UNIVERSITÉ PARIS. DIDEROT (Paris 7)

THÈSE

pour obtenir le grade de

Docteur de l'université Paris. Diderot

Spécialité: Mathématiques Appliquées

présentée par

François BACHOC

Estimation paramétrique de la fonction de covariance dans le modèle de Krigeage par processus Gaussiens. Application à la quantification des incertitudes en simulation numérique

---

Soutenue le 3 octobre 2013 devant le jury composé de:

Présidente	Mme Dominique PICARD	Université Paris. Diderot
Directeur de Thèse	M. Josselin GARNIER	Université Paris. Diderot
Encadrant CEA	M. Jean-Marc MARTINEZ	CEA Saclay
Rapporteur	M. Luc PRONZATO	Université de Nice-Sophia Antipolis
Examinateurs	M. Hannes LEEB	Université de Vienne
	M. Nicolas VAYATIS	ENS Cachan
	M. Emmanuel VAZQUEZ	École supérieure d'électricité

D'après les rapports de:

M. Luc PRONZATO	Université de Nice-Sophia Antipolis
M. Michael STEIN	Université de Chicago

Thèse réalisée dans le cadre d'un Contrat de Formation par la Recherche au Commissariat à l'Energie Atomique et aux énergies alternatives de Saclay

Remerciements

Mes remerciements vont tout d'abord au duo d'encadrants qui m'a accompagné pendant ces trois années de thèse. Josselin Garnier, en tant que directeur de thèse, a gardé un intérêt et une attention constante sur notre travail malgré ses nombreuses occupations. Il m'a aussi conseillé avec efficacité et franchise et a constitué une caution scientifique sur de nombreux produits de cette thèse. Je retiendrai également sa rapidité et sa rigueur dans la relecture de mes productions écrites. Jean-Marc Martinez, encadrant CEA, est à l'initiative de cette thèse, en décembre 2009. Depuis, il m'a initié à la recherche appliquée, a toujours été la pour moi, et a manifesté un souci constant de valoriser notre travail et de communiquer à son sujet. Nos échanges étaient chaleureux et stimulants, et m'ont permis de découvrir une grande variété de problèmes statistiques issus de la R&D.

Je suis honoré que Luc Pronzato et Michael Stein soient rapporteurs de ce travail. Les nombreux commentaires et suggestions de Luc Pronzato, ainsi que sa disponibilité pour m'aider à les prendre en compte, me permettront certainement d'avancer. *I also thank Michael Stein for his detailed report. I hope that working on his comments will prove fruitful in the future.*

Hannes Leeb has accepted to be a member of the jury, and I thank him for that. I also seize the opportunity to acknowledge the kind welcome I received in Vienna from him, Benedikt Poetscher, Birgit Ewald and the praedocs David, Ivana, Lukas and Nina. I hope that we will work well in the near future. J'adresse mes remerciements à Dominique Picard qui accepte de faire partie de ce jury. J'attends notre rencontre avec intérêt. En 2009/2010, j'ai eu le plaisir de suivre deux cours de Nicolas Vayatis, et il a bien voulu m'accorder de son temps pour me conseiller. Je le remercie pour cela, ainsi que pour être maintenant membre du jury de cette thèse. Emmanuel Vazquez a également accepté de me conseiller en tout début de thèse. Son mémoire a aussi occupé une place importante dans mes documents de travail. Je souhaite le remercier de compléter ce jury.

Durant cette thèse, j'ai essentiellement travaillé au centre CEA de Saclay. Je salue donc d'abord mes collègues de bureau successifs: Francis, lors de mon stage, Thu Huyen, maintenant docteure et Kieu et Takoua, à qui je souhaite bonne chance pour la suite de leurs travaux. Je remercie Vincent Bergeaud et Jacques Segré pour leur soutien et leur confiance. Je suis reconnaissant à mes deux chefs de service successifs, Danier Caruge et Bernard Faydide, de m'avoir accueilli et d'avoir témoigné de l'intérêt pour cette thèse. Il m'est agréable de signaler que ma vie administrative a été facilitée par Evelyne Macanda et Elsa Rodrigues. Enfin, j'adresse mes salutations aux membres du Laboratoire de Génie Logiciel et Simulation que j'ai particulièrement côtoyés, ainsi qu'aux membres de notre service, avec qui j'ai pu échanger, notamment les stagiaires et thésards.

A Saclay, j'ai trouvé un accueil bienveillant et une ouverture d'esprit de la part des membres de notre groupe incertitude: Gilles Arnaud, Agnes de Crecy, Fabrice Gaudier, Nicolas Gilardi et Jean-Marc Martinez. Le travail réalisé avec Guillaume Bois, sur les données thermohydrauliques, a été intéressant et fructueux. Enfin, je souhaite remercier ici Karim Ammar, doctorant, pour notre collaboration, son intérêt pour le Krigeage et pour m'avoir initié à son domaine de travail.

Le Groupement De Recherche MASCOT-NUM et le consortium ReDICE ont rendu mon travail de thèse plus stimulant. Je salue donc les membres de ces deux communautés avec qui

j'ai pu échanger. Certains m'ont en particulier fait profiter de leur expérience, sur les plans scientifique ou, plus largement, professionnel. Je pense ici à Fabrice Gamboa, David Ginsbourger, Bertrand Iooss et Olivier Roustant. Enfin, je conclurai ce passage sur la communauté incertitude en saluant les doctorants, ou jeunes docteurs, que j'ai rencontrés: Miguel, Pierre et Nabil, jeunes docteurs, Loïc, Gaëlle, Clément, Alexandre et Paul, dont les soutenances sont proches, et Vincent, Guillaume, Simon, Michael et Olivier, à qui je souhaite la meilleure réussite pour la suite de leurs thèses.

J'envoie un salut à mes camarades du projet REDVAR du CEMRACS 2013, Achref et Lionel.

Je suis reconnaissant à Erick Herbin pour ses conseils avant cette thèse, et pour avoir participé à l'élaboration de celle-ci.

Je saisis ici l'occasion de remercier deux de mes anciens professeurs de Mathématiques, dont je garde un souvenir agréable des cours: Sylvain Courjaud, au lycée, et Guillaume Roussel, en classe de MPSI 3.

Et pour conclure cette séquence de remerciements, je remercie pour leur soutien mes amis, mes parents, ma sœur, *amaxi* et le reste de ma famille.

François Bachoc, le 19 septembre 2013

Résumé

L'estimation paramétrique de la fonction de covariance d'un processus Gaussien est étudiée, dans le cadre du modèle de Krigeage. Les estimateurs par Maximum de Vraisemblance et Validation Croisée sont considérés. Le cas correctement spécifié, dans lequel la fonction de covariance du processus Gaussien appartient à l'ensemble paramétrique de fonctions de covariance, est d'abord traité dans un cadre asymptotique par expansion. Le plan d'expériences considéré est une grille régulière multidimensionnelle perturbée aléatoirement. Un résultat de consistance et de normalité asymptotique est montré pour les deux estimateurs. Il est ensuite mis en évidence que des amplitudes de perturbation importantes sont toujours préférables pour l'estimation par Maximum de Vraisemblance. Le cas incorrectement spécifié, dans lequel l'ensemble paramétrique utilisé pour l'estimation ne contient pas la fonction de covariance du processus Gaussien, est ensuite étudié. Il est montré que la Validation Croisée est alors plus robuste que le Maximum de Vraisemblance. Enfin, deux applications du modèle de Krigeage par processus Gaussiens sont effectuées sur des données industrielles. Pour un problème de validation du modèle de frottement pariétal du code de thermohydraulique FLICA 4, en présence de résultats expérimentaux, il est montré que la modélisation par processus Gaussiens de l'erreur de modèle du code FLICA 4 permet d'améliorer considérablement ses prédictions. Enfin, pour un problème de métamodélisation du code de thermomécanique GERMINAL, l'intérêt du modèle de Krigeage par processus Gaussiens, par rapport à des méthodes par réseaux de neurones, est montré.

Abstract

The parametric estimation of the covariance function of a Gaussian process is studied, in the framework of the Kriging model. Maximum Likelihood and Cross Validation estimators are considered. The correctly specified case, in which the covariance function of the Gaussian process does belong to the parametric set used for estimation, is first studied in an increasing-domain asymptotic framework. The sampling considered is a randomly perturbed multidimensional regular grid. Consistency and asymptotic normality are proved for the two estimators. It is then put into evidence that strong perturbations of the regular grid are always beneficial to Maximum Likelihood estimation. The incorrectly specified case, in which the covariance function of the Gaussian process does not belong to the parametric set used for estimation, is then studied. It is shown that Cross Validation is more robust than Maximum Likelihood in this case. Finally, two applications of the Kriging model with Gaussian processes are carried out on industrial data. For a validation problem of the friction model of the thermal-hydraulic code FLICA 4, where experimental results are available, it is shown that Gaussian process modeling of the FLICA 4 code model error enables to considerably improve its predictions. Finally, for a metamodeling problem of the GERMINAL thermal-mechanical code, the interest of the Kriging model with Gaussian processes, compared to neural network methods, is shown.

Contents

Chapter 1

Introduction

The analysis of computer experiments

In the past decades, a new field of statistics, the design and analysis of computer experiments ([SWMW89], [SWN03]), has gradually gained a lot of interest from the statistical community, from both a theoretical and applied point of view. In this thesis, we define a computer experiment by the collection of an input point $\boldsymbol{x}$, a computer model function f_{mod}, and the output of the simulation $y = f_{mod}(\boldsymbol{x})$. The term computer experiment means that the use of the computer model f_{mod}, for obtaining the simulated value of a given phenomenon of interest at $\boldsymbol{x}$, shares several characteristics with the classical notion of physical experiment.

The first point is that a computer simulation is potentially costly (in terms of time and of hardware necessary for the simulation), so that it is already an issue in itself to select the simulation input $\boldsymbol{x}$ that would give the maximal amount of information, for the minimal cost. Shall this input point be specified, recent computer simulators come with a large number of optional parameters, that also have to be fixed. These parameters enable the computer model to be a versatile and accurate representation of reality, but add more complexity for the simulator user. Finally, since computer models address more and more complex physical phenomena (in particular, multi-physics or multi-scale phenomena), it is not yet certain that their accuracies are sufficient, for the considered applications.

We have hence, informally, listed three main problems related to computer experiments. First, for computational cost reasons, the computer model function $f_{mod}(\boldsymbol{x})$ can not be calculated for arbitrarily many inputs $\boldsymbol{x}$. This makes it computationally prohibitive to directly solve the calibration and validation problems, that we present below, as well as to address other analyses involving many calls to the f_{mod} function (such as sensitivity analysis or global optimization). The set of techniques for building a cheap and reasonably accurate approximation of this function constitutes the field of metamodeling. In this thesis, we study meta-modeling methods in which the computer model function is treated as a black box, known only from its inputs and outputs. Thus, a meta-model $\hat{f}_{mod}(\boldsymbol{x})$ of $f_{mod}(\boldsymbol{x})$ is built from a set $(\boldsymbol{x}^{(1)}, y_1), ..., (\boldsymbol{x}^{(n)}, y_n)$ of input and output points.

Calibration corresponds, when an input point $\boldsymbol{x}$ is fixed for the computer model f_{mod}, to fix the optional parameters necessary to carry out the simulation. The computer model function

is thus denoted $f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the calibration parameter and has to be fixed prior to carrying out a computer experiment for the input $\boldsymbol{x}$. It may also be expected that an uncertainty quantification be associated to the selected value for $\boldsymbol{\beta}$.

Validation is the problem of quantifying the discrepancy between the computer model, ideally calibrated, and the true underlying physical system. More precisely, denoting $f_{real}(\boldsymbol{x})$ the variable of interest of the physical system at $\boldsymbol{x}$, we are interested in quantifying the residual error $f_{real}(\boldsymbol{x}) - f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$, when $\boldsymbol{\beta}$ is correctly calibrated.

Gaussian process models

The central probabilistic notion that we consider in this thesis, for addressing the calibration, validation and meta-modeling problems presented above, is the notion of Gaussian process. In this context, a Gaussian process is considered as a Bayesian *a priori* distribution on a deterministic function [RW06]. More precisely, the deterministic function is either the computer model, for the meta-modeling problem, or possibly the deterministic error function between the physical system and the correctly calibrated code, for the validation problem.

Gaussian processes have become popular for representing random functions, because of their tractability (all linear functionals of a Gaussian process remain Gaussian variables), their conceptual simplicity (they are defined only by a mean and a covariance function) and the fact that they constitute a reasonable representation for a large class of deterministic continuous functions. Furthermore, Gaussian process models yield confidence intervals, for the value of a random function at particular input points, that are easily computable. Note here the parallelism between the utilization of Gaussian processes for representing, say, a deterministic computer model, and, since a longer time, the utilization of random processes to represent a deterministic geostatistical function. This last paradigm is known as Kriging [Mat70].

The treatment of a Gaussian process model is most classically carried out in a two-step approach. First, the mean and covariance functions of the Gaussian process are estimated from a data set, that is a set of input and output points stemming from the same trajectory of the Gaussian process. The estimation of the covariance function is generally carried out within a fixed parametric family, so that it boils down to estimating a finite number of hyper-parameters, characterizing it. The mean function is most of the times selected in a linearly-parameterized set. Second, the covariance function is fixed to the obtained estimate, so that, using simple matrix-vector formulas the conditional distribution of any linear functional of the Gaussian process, given the observed values, remains Gaussian and can be computed. The uncertainty resulting from the estimation of the linearly-parameterized mean function is also taken into account with explicit formula.

Concerning the covariance function estimation, the most widely used estimation method is the Maximum Likelihood method [MM84]. The popularity of this method is notably justified by the attractive general properties of asymptotic consistency and normality for Maximum Likelihood estimators. These general properties can indeed be verified in the Gaussian process framework [MM84].

Another estimation method is the Cross Validation method [Dub83], [ZW10]. This method

consists in defining an empirical, cross-validation based, prediction criterion on the data set of the Gaussian process. Hence, an attractive feature is that Cross Validation selects a covariance hyper-parameter directly according to its empirical prediction results on the data set, and that the prediction criterion can be chosen. This shall make it possible, for the Cross Validation method, to yield particularly robust estimations, according to the selected criterion. Particularly, in this thesis, we show in chapter 6, that, for approximating the conditional mean function and the pointwise conditional variances of the Gaussian process, a Cross Validation procedure is more robust than Maximum Likelihood, to misspecifications of the parametric family of covariance functions.

Both the Maximum Likelihood and Cross Validation estimators are defined as minimizing criteria of the observed values, that have explicit matrix-vector expressions.

Notice that, while the criterion to minimize is well-defined for Maximum Likelihood, Cross Validation is, as we have said, a general method yielding several possible criteria [ZW10], [RW06]. These criteria could be more or less appropriate, according to the objective of the Gaussian process model (e.g. simply approximating the conditional mean function of the Gaussian process, according to the observations, or approximating its full conditional distribution). We consider the choice of the Cross Validation procedure as an open problem. In this thesis, we have chosen to address the natural Cross Validation criterion that consists in the Leave-One-Out mean square prediction error, which is associated to the objective of approximating the conditional mean function of the Gaussian process.

Despite the conceptual simplicity of Gaussian process models, the problem of the covariance function estimation (for example of which covariance function estimator to use), and of the quantification of the influence of estimation on prediction, is not fully understood yet. The main obstacle for a complete mathematical treatment is the dependence between all the observations that are made of a Gaussian process. Notice also that the estimators of covariance hyper-parameters are only known as being statistical M-estimators, with an explicit criterion function. Furthermore, for a fixed covariance function, even if the prediction formulas are explicit, it is not easy to derive general conclusions from them, notably because they incorporate an inverse matrix term.

As a consequence, most general theoretical results in the direction of covariance function estimation and of its influence on prediction are asymptotic (in the sense that the number of observation points goes to infinity). The reader may refer to [MM84] and e.g. [Yin91], [Zha04] for an asymptotic analysis of the Maximum Likelihood estimator. Concerning asymptotic results for the prediction problem, we refer to [Ste99].

Contributions of the thesis

The thesis makes several significant contributions to the field of Gaussian process modeling for the analysis of computer experiments. First, the covariance function estimation problem is investigated, from both a finite sample and an asymptotic point of view. In particular, the Cross Validation estimator is analyzed, and compared with Maximum Likelihood, and the impact of the design of experiments on the quality of the covariance function estimation is studied.

Second, the utilization of Gaussian process models for calibration, validation and meta-modeling of computer models is treated, from the methodological point of view, and in two real-case studies on two industrial computer models.

Organization of the manuscript

The thesis is organized into three parts. Part I constitutes a review of the state of the art regarding Gaussian process modeling.

In chapter 2, we review the finite-sample treatment of Gaussian process models, when the covariance function is fixed. We describe the influence of the covariance function on the nature of the trajectories of the Gaussian process, and we review the classical covariance function families, that we use in the thesis. We also review a variety of explicit formulas and methods for prediction, conditional simulation and Cross Validation. For Cross Validation, we propose practical and simple matrix-vector formulas, obtained form the virtual Cross Validation formulas of [Dub83].

In chapter 3, we address covariance function estimation for Gaussian processes. The chapter starts with an introduction to statistical parametric estimation. The most classical asymptotic consistency and efficiency results, for the Maximum Likelihood estimator with independent and identically distributed observations, are presented. Then, the different Maximum Likelihood and Cross Validation methods for covariance function estimation are introduced. A large variety of explicit formulas, including the gradients of the criteria, are gathered, which can be useful from a practical point of view.

Chapter 4 constitutes an introduction to the existing asymptotic results for Gaussian process models. First, the two classical fixed-domain and increasing-domain asymptotic frameworks are presented. Then, fixed-domain asymptotic results are discussed for prediction with fixed covariance function. These results concern the asymptotic consistency of Kriging predictions and a quantification of the asymptotic influence of the covariance function choice. Finally, the existing asymptotic results on covariance function estimation by Maximum Likelihood are presented.

Part II is dedicated to our contributions to the covariance function estimation problem for Gaussian processes.

In chapter 5, we address an increasing-domain asymptotic framework, which yields three main conclusions. First, we prove that, in this favorable context for estimation where Maximum Likelihood is known to be consistent and asymptotically normal, Cross Validation is also consistent and has the same convergence rate as Maximum Likelihood. This is a desirable theoretical result, for Cross Validation to be considered in practice. Second, we confirm that Maximum Likelihood yields a smaller asymptotic variance than Cross Validation. Indeed, chapter 5 addresses what we call the well-specified framework, where the true covariance function of the Gaussian process does belong to the parametric family used for estimation. Maximum Likelihood estimators are classically preferable in this context. The third conclusion of chapter 5 concerns the impact of the spatial sampling on the covariance function estimation, for the Maximum Likelihood and Cross Validation estimators. An asymptotic confirmation is given to the

commonly admitted fact that using groups of observation points with small spacing is beneficial to covariance function estimation. Finally, the prediction error, using a consistent estimator of the covariance parameters, is analyzed in details.

In chapter 6, we carry out a finite-sample comparison of Maximum Likelihood and Cross Validation in what we call the misspecified framework. This means that the true covariance function of the Gaussian process does not belong to the parametric family of covariance functions used for estimation. In this context, we show that Cross Validation is more robust than Maximum Likelihood. We follow a two-step approach. In a first step, we address theoretically the case of the estimation of a single variance hyper-parameter, where the correlation function is fixed and misspecified. Then, we numerically confirm the results of the first step, in the case where variance and correlation hyper-parameters are estimated from data.

In part III, we address the application of Gaussian process models to the calibration, validation and metamodeling of computer models.

In chapter 7, we review the existing methodologies for addressing calibration and validation. These two problems can equivalently be considered as the problem of modeling the discrepancies between a set of experimental results, and the associated set of computer-model results.

We distinguish two classes of methods. First, we review the methods considering the underlying physical system as intrinsically random. This randomness is governed by a randomness in the calibration parameters of the computer model, so that the goal is to estimate their distribution. We hence review the existing methods, relying or not on a linear approximation of the computer model with respect to its calibration parameters.

The second class of methods for calibration and validation treats the physical system as deterministic, and introduce the notion of model error. The model error is the bias between the physical system and the perfectly calibrated computer model. It is represented by a trajectory of a Gaussian process. We introduce the different objectives associated to this statistical model, and we review the methodology, in the case where no linear approximation of the computer model is done. The most important feature of this methodology is that, eventually, the prediction of the physical system for an input point $\boldsymbol{x}$, where no experiments have been done, is composed of the calibrated code, completed by an inference of the model error function at $\boldsymbol{x}$.

We then present the simplifications of the method above when a linear approximation of the computer model is done, with respect to its model parameters. This is the case we focus on in this thesis.

In chapter 8, we apply the Gaussian process modeling of the model error, with the linear approximation of the computer model with respect to its model parameters, to the thermal-hydraulic code FLICA 4, for which a set of experimental results is available. We show that taking the model error into account (that is to say, predicting an experimental result by the calibrated FLICA 4 code completed by the model error inference) yields a significantly smaller prediction error than when using only the calibrated FLICA 4 code.

In chapter 9, we address the meta-modeling of the GERMINAL thermal-mechanical code. The meta-modeling method consists in a classical Gaussian process model, in which the GERMINAL computer model is represented as a trajectory of a Gaussian process. We show that the Gaussian process model yields good prediction results, compared to an alternative meta-

modeling method based on an artificial neural network, that has also been applied to the GERMINAL thermal-mechanical code. Furthermore, the probabilistic model, underlying the Gaussian process predictions, enables to select automatically outlier points in a base of results of the GERMINAL thermal-mechanical code. This is an attractive practical feature of Gaussian process models, since it is prohibitive to check manually the validity of all the GERMINAL results. It is confirmed that the output points that are selected by the Gaussian process model do correspond to computation failures of the GERMINAL code.

Part I

Kriging models

Chapter 2

Kriging models with known covariance function

In this chapter we present classical results on Kriging models, in the case when the covariance function of the Gaussian process is assumed to be known. The mean function is either known, or assumed in a linear, finite-dimensional family of functions.

In section 2.1, we present the basic properties for Gaussian processes. These properties are for instance stationarity and regularity. Then, we show how these properties are linked with the covariance function of the Gaussian process. We conclude the section by presenting the covariance function families we study in the thesis, and the associated properties of regularity.

In section 2.2, we address prediction and conditional simulation, when a set of observed values is available for the Gaussian process. We present the simple, ordinary and universal Kriging frameworks, and we review the most classical formulas of the literature for prediction and conditional simulation. Then, we present the Cross Validation concepts, and we give the associated virtual Cross Validation formulas. Finally, we give a few words on the parallelism between Gaussian process prediction and ridge regression in Reproducing Kernel Hilbert Space.

2.1 Gaussian processes

2.1.1 Definition and properties of Gaussian processes

Random processes

We give a short introduction to random processes. For further reference on random processes (including the mathematical construction), we refer, e.g, to the chapter seven of the monograph [Bil12].

In all the manuscript, we consider a domain of interest $\mathcal{D} \subset \mathbb{R}^d$. The main probabilistic notion we use for Kriging models is the notion of random process, presented in definition 2.1.

Definition 2.1. *A real-valued random process (or random function) on $\mathcal{D}$ is an application Y, that associates a random variable $Y(\boldsymbol{x})$ to each $\boldsymbol{x} \in \mathcal{D}$. All the random variables $Y(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{D}$, are defined respectively to a common probability space $(\Omega, \mathcal{F}, P)$.*

Remark 2.2. *In the manuscript, mention to the probability space $(\Omega, \mathcal{F}, P)$ is generally omitted, for concision. Nevertheless, the probability space is sometimes explicitly used, particularly in chapter 5.*

The fact that the probability space is common for the random variables $Y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$, is very important. Indeed, it enables to talk about the trajectories of a random process, as presented in definition 2.3.

Definition 2.3. *For each fixed $\omega \in \Omega$, the real-valued function $\boldsymbol{x} \to Y(\omega, \boldsymbol{x})$ is called a trajectory (or a realization or a sample function) of the random process Y.*

Let us consider an example for definitions 2.1 and 2.3. With $(\Omega, \mathcal{F}, P)$ a probability space and U a real random variable on $(\Omega, \mathcal{F}, P)$, following the uniform distribution on $[-\pi, \pi]$, we consider the random process $Y(\omega, x) = \cos(U(\omega) + x)$. The random process Y is a sinusoid with deterministic period and random phase. Its trajectories are sinusoid with period 2π, and the phase varies among the trajectories.

The notion of trajectory of a random function of definition 2.3 is at least as important as the formal definition 2.1 from an interpretation point of view. Indeed, in the same way as a random number is a number that can change according to a random phenomenon, a random function is a function that can change according to a random phenomenon. In the manuscript, we will essentially postulate that a deterministic function is actually a trajectory of a random function. The interpretation is that a deterministic function can be seen as the result of a "past" random phenomenon (which is unknown and now over). Hence it is conceivable that the random phenomenon could have had different results, which would have yielded different deterministic functions.

The fact that the probability space is common for the random variables $Y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$ is also important to define the finite-dimensional distributions of a random function, in definition 2.4.

Definition 2.4. *For any n points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$, the multidimensional probability distribution of the random vector $Y(\boldsymbol{x}^{(1)})$,...,$Y(\boldsymbol{x}^{(n)})$ is called a finite-dimensional distribution of the random function Y.*

The notion of finite-dimensional distribution is the basis of the predictions and conditional simulations of section 2.2. Roughly speaking, the fact that there is a probability distribution for the random vector $\big(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}), Y(\boldsymbol{x})\big)$ enables us to predict the value of $Y(\boldsymbol{x})$, after observing the values of $Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)})$.

Introduction to the multidimensional Gaussian distribution

We give a short introduction to the Gaussian multidimensional distribution. For other introductions to the multidimensional Gaussian distribution, we refer e.g to appendix B.1 of [SWN03] or appendix A.2 of [RW06].

Consider $n > 1$ and a real random vector $\boldsymbol{y} = (y_1, ..., y_n)$. This random vector is said to be a Gaussian vector if the two following equivalent conditions are verified.

- For any $a_1, ..., a_n \in \mathbb{R}$, the random variable $\sum_{i=1}^{n} a_i y_i$ follows a Gaussian distribution.

- There exists a vector $\boldsymbol{m}$ of size n and a $n \times n$ non-negative matrix $\mathbf{K}$ so that the random vector $\boldsymbol{y}$ has characteristic function $\boldsymbol{u} \to \exp\left(\mathrm{i}\boldsymbol{u}^t\boldsymbol{m} - \frac{1}{2}\boldsymbol{u}^t\mathbf{K}\boldsymbol{u}\right)$

If the two conditions are verified, we will write it $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}, \mathbf{K})$ and furthermore we have $\boldsymbol{m} = \mathbb{E}(\boldsymbol{y})$ and $\mathbf{K} = Cov(\boldsymbol{y})$.

When $\mathbf{K}$ is non-singular, the probability density function of $\boldsymbol{y}$ at $\boldsymbol{x} \in \mathbb{R}^n$ is, with $\boldsymbol{m} = \mathbb{E}(\boldsymbol{y})$ and $\mathbf{K} = Cov(\boldsymbol{y})$

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\mathbf{K}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^t\mathbf{K}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right), \tag{2.1}$$

where $|\mathbf{K}|$ stands for the determinant of $\mathbf{K}$.

When $\mathbf{K}$ is singular, there exists a hyperplane of $\mathbb{R}^n$ which is the support of $\boldsymbol{y}$ (meaning that $\boldsymbol{y}$ almost surely belongs to this hyperplane) and so that, restricted on this hyperplane, $\boldsymbol{y}$ has a probability density function of the form (2.1) (with respect to the Lebesgue measure over the hyperplane).

We conclude the introduction to the multi-dimensional Gaussian distribution by stating the Gaussian conditioning theorem.

Theorem 2.5. *Consider a Gaussian vector of size $n = n_1 + n_2$ of the form*

$$\begin{pmatrix} \boldsymbol{y}^{(1)} \\ \boldsymbol{y}^{(2)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{m}^{(1)} \\ \boldsymbol{m}^{(2)} \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{1,1} & \mathbf{K}_{1,2} \\ \mathbf{K}_{2,1} & \mathbf{K}_{2,2} \end{pmatrix}\right)$$

Then, when $\mathbf{K}_{1,1}$ is non-singular, conditionally to $\boldsymbol{y}^{(1)} = \boldsymbol{v}^{(1)}$, $\boldsymbol{y}^{(2)}$ follows a

$$\mathcal{N}(\boldsymbol{m}^{(2)} + \mathbf{K}_{2,1}\mathbf{K}_{1,1}^{-1}(\boldsymbol{v}^{(1)} - \boldsymbol{m}^{(1)}), \mathbf{K}_{2,2} - \mathbf{K}_{2,1}\mathbf{K}_{1,1}^{-1}\mathbf{K}_{1,2})$$

distribution.

Roughly speaking, theorem 2.5 gives, from two Gaussian vectors, the distribution of the second one conditionally to the first one. The fact that this conditional distribution remains Gaussian is remarkable, and is one of the reasons for the popularity of Gaussian-based probabilistic models.

Remark 2.6. *In theorem 2.5, at first sight, it seems necessary that the $n_1 \times n_1$ matrix $\mathbf{K}_{1,1}$ be non-singular. We now give a short discussion on how to proceed when $\mathbf{K}_{1,1}$ is singular. The important point is that the mathematical definition of conditional distributions is still valid when the conditioning random vector is redundant. For instance, for two random variables X, Z, the conditional distribution of X according to the degenerate random vector (Z, Z) is well-defined and is simply the conditional distribution of X according to Z.*

In the case when $\mathbf{K}_{1,1}$ is singular, there exists $n_1' < n_1$, a $n_1' \times n_1$ matrix $\mathbf{P}_1$ and a $(n_1 - n_1') \times n_1$ matrix $\mathbf{P}_2$ so that $\mathbf{P}_1\boldsymbol{y}^{(1)}$ has a non-singular covariance matrix and $\mathbf{P}_2\boldsymbol{y}^{(1)} = 0$ almost surely. Then the support of $\boldsymbol{y}$ is a hyperplane of dimension n_1' (meaning that $\boldsymbol{y}$ almost surely belongs to this hyperplane), so that $\mathcal{L}(\boldsymbol{y}^{(2)}|\boldsymbol{y}^{(1)}) = \mathcal{L}(\boldsymbol{y}^{(2)}|\mathbf{P}_1\boldsymbol{y}^{(1)})$.

Hence, with $Cov(\boldsymbol{y}^{(2)}, \mathbf{P}_1\boldsymbol{y}^{(1)}) = \mathbf{K}_{2,1}\mathbf{P}_1^t$ and $Cov(\mathbf{P}_1\boldsymbol{y}^{(1)}) = \mathbf{P}_1\mathbf{K}_{1,1}\mathbf{P}_1^t$, we get from theorem 2.5

$$\mathcal{L}\left(\boldsymbol{y}^{(2)}|\boldsymbol{y}^{(1)} = \boldsymbol{v}^{(1)}\right) =$$
$$\mathcal{N}\left[\boldsymbol{m}^{(2)} + \mathbf{K}_{2,1}\mathbf{P}_1^t(\mathbf{P}_1\mathbf{K}_{1,1}\mathbf{P}_1^t)^{-1}\mathbf{P}_1(\boldsymbol{v}^{(1)} - \boldsymbol{m}^{(1)}), \mathbf{K}_{2,2} - \mathbf{K}_{2,1}\mathbf{P}_1^t(\mathbf{P}_1\mathbf{K}_{1,1}\mathbf{P}_1^t)^{-1}\mathbf{P}_1\mathbf{K}_{1,2}\right],$$

so that there is also a matrix-vector formula in the case where $\mathbf{K}_{1,1}$ *is singular.*

Consider now a symmetric SVD (see e.g. [GL96]) $\mathbf{K}_{1,1} = \mathbf{U}\mathbf{D}\mathbf{U}^t$ *with orthogonal* $n_1 \times n_1$ *matrix* $\mathbf{U}$ *and diagonal matrix* $\mathbf{D}$ *with diagonal values* $\lambda_1 > ... > \lambda_{n_1'} > \lambda_{n_1'+1} = ... = \lambda_{n_1} = 0$. *Let, with* $n_1 \times n_1'$ *matrix* $\mathbf{U}_{n_1'}$ *and* $n_1 \times (n_1 - n_1')$ *matrix* $\mathbf{U}_{n_1-n_1'}$,

$$\mathbf{U} = \left(\begin{array}{cc} \mathbf{U}_{n_1'}, & \mathbf{U}_{n_1-n_1'} \end{array} \right)$$

Then $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *can be* $\mathbf{U}_{n_1'}^t$ *and* $\mathbf{U}_{n_1-n_1'}^t$. *This yields, with* $\mathbf{K}_{1,1} = \mathbf{U}_{n_1'}\mathbf{D}_{n_1'}\mathbf{U}_{n_1'}^t$, *with* $\mathbf{D}_{n_1'}$ *the* $n_1' \times n_1'$ *diagonal matrix with diagonal values* $\lambda_1, ..., \lambda_{n_1'}$,

$$\begin{aligned} \mathbb{E}(\boldsymbol{y}^{(2)}|\boldsymbol{y}^{(1)} = \boldsymbol{v}^{(1)}) &= \boldsymbol{m}^{(2)} + \mathbf{K}_{2,1}\mathbf{U}_{n_1'}(\mathbf{U}_{n_1'}^t\mathbf{U}_{n_1'}\mathbf{D}_{n_1'}\mathbf{U}_{n_1'}^t\mathbf{U}_{n_1'})^{-1}\mathbf{U}_{n_1'}^t(\boldsymbol{v}^{(1)} - \boldsymbol{m}^{(1)}) \\ &= \boldsymbol{m}^{(2)} + \mathbf{K}_{2,1}\mathbf{U}_{n_1'}\mathbf{D}_{n_1'}^{-1}\mathbf{U}_{n_1'}^t(\boldsymbol{v}^{(1)} - \boldsymbol{m}^{(1)}) \end{aligned}$$

and

$$\begin{aligned} Cov(\boldsymbol{y}^{(2)}|\boldsymbol{y}^{(1)} = \boldsymbol{v}^{(1)}) &= \mathbf{K}_{2,2} - \mathbf{K}_{2,1}\mathbf{U}_{n_1'}(\mathbf{U}_{n_1'}^t\mathbf{U}_{n_1'}\mathbf{D}_{n_1'}\mathbf{U}_{n_1'}^t\mathbf{U}_{n_1'})^{-1}\mathbf{U}_{n_1'}^t\mathbf{K}_{1,2} \\ &= \mathbf{K}_{2,2} - \mathbf{K}_{2,1}\mathbf{U}_{n_1'}\mathbf{D}_{n_1'}^{-1}\mathbf{U}_{n_1'}^t\mathbf{K}_{1,2}. \end{aligned}$$

Hence, we see that we can compute $\mathcal{L}(\boldsymbol{y}^{(2)}|\boldsymbol{y}^{(1)} = \mathbf{V}^{(1)})$ *by using a SVD of the singular matrix* $\mathbf{K}_{1,1}$. *Therefore, the computational cost is of the same order as in the non-singular case. Note also that the matrix* $\mathbf{K}_{1,1}^- := \mathbf{U}_{n_1'}\mathbf{D}_{n_1'}^{-1}\mathbf{U}_{n_1'}^t$ *is a pseudo inverse of* $\mathbf{K}_{1,1}$ *that is to say, it verifies* $\mathbf{K}_{1,1}\mathbf{K}_{1,1}^-\mathbf{K}_{1,1} = \mathbf{K}_{1,1}$ *and* $\mathbf{K}_{1,1}^-\mathbf{K}_{1,1}\mathbf{K}_{1,1}^- = \mathbf{K}_{1,1}^-$. *Hence, remark 2.6 can be summarized by the easy to remember rule: if* $\mathbf{K}_{1,1}$ *is singular, replace its inverse by the pseudo-inverse above in the formulas for the Gaussian conditioning theorem.*

Finally, when the matrix $\mathbf{K}_{1,1}$ *is theoretically non-singular but ill-conditioned, it may be advised to approximate its lowest eigenvalues by zero and to use the formulas of remark 2.6.*

Gaussian processes

In the manuscript, we especially study a particular class of random processes: the Gaussian processes. These processes are based on the multi-variable Gaussian distribution presented above.

Definition 2.7. *A random process is a Gaussian process if its finite-dimensional distributions are multidimensional Gaussian distributions.*

Assuming that a random process at hand is Gaussian is classical for several reasons. First, the Gaussian distribution is generally an acceptable choice to model the statistical distribution of a random variable which has a priori reasons to be symmetric, unimodal, and with probability density function decreasing when one goes away from the mean value.

Second, as we see in section 2.2, using a Gaussian process considerably simplifies the treatment of a given problem at hand, both conceptually and in practice. Conceptually, it ensures that the overall random process is easy to define, and that is stays Gaussian after conditioning to a set of observation points (theorem 2.5). In practice, linear finite-dimensional treatments boils down to classical vector-matrix formulas that have a relatively low computational cost.

Finally, let us also mention that, among the different classes of random processes, the Gaussian processes are the ones for which the most theory has been done. For example, there exists several monographs giving detailed results for Gaussian processes, for instance on the properties of the trajectory functions ([Adl90]) and on the prediction problem ([Ste99]).

In the rest of the manuscript, we always consider Gaussian processes. Nevertheless, many notions or results that are presented hold for general random processes.

Mean and covariance functions

A multidimensional Gaussian distribution is characterized by its mean vector and its covariance matrix. In the same way, a Gaussian process is characterized by its mean and covariance functions, that are defined below.

Definition 2.8. *The mean function of a Gaussian process Y is the application m: $\mathcal{D} \to \mathbb{R}$, defined by $m(\boldsymbol{x}) = \mathbb{E}(Y(\boldsymbol{x}))$.*

Definition 2.9. *The covariance function of a Gaussian process Y is the application K: $\mathcal{D} \times \mathcal{D} \to \mathbb{R}$, defined by $K(\boldsymbol{x}, \boldsymbol{y}) = Cov(Y(\boldsymbol{x}), Y(\boldsymbol{y}))$.*

In definition 2.8 and 2.9, the mean function can be any function m: $\mathcal{D} \to \mathbb{R}$. However there is an important constraint on the covariance function K. Indeed, for any $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$, the $n \times n$ covariance matrix $\mathbf{K}$, defined by $\mathbf{K}_{i,j} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$, must be non-negative. Hence, the covariance function K must be positive-definite, as defined in definition 2.10.

Definition 2.10. *A function K: $\mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is positive definite if, for any $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$, the $n \times n$ covariance matrix $\mathbf{K}$, defined by $\mathbf{K}_{i,j} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$, is non-negative.*

A positive-definite function is also called a kernel, and its application to the general field of machine learning has yielded the denomination of kernel methods. There is a fair amount of literature on studying the positive-definiteness of bivariate functions K: $\mathcal{D} \times \mathcal{D} \to \mathbb{R}$ (e.g [SS02], ch.13).

In subsection 2.1.2, we give a review of the covariance functions we consider in the manuscript.

Stationarity

The notion of stationarity corresponds to a random process which has the same behavior, regardless of the location on the domain $\mathcal{D}$. The precise definition is given below.

Definition 2.11. *A random process Y is stationary if, for all $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$ and $\boldsymbol{h} \in \mathbb{R}^d$, so that $\boldsymbol{x}^{(1)} + \boldsymbol{h}, ..., \boldsymbol{x}^{(n)} + \boldsymbol{h}$ remain in $\mathcal{D}$, the finite-dimensional distribution of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ is the same as the finite-dimensional distribution at $\boldsymbol{x}^{(1)} + \boldsymbol{h}, ..., \boldsymbol{x}^{(n)} + \boldsymbol{h}$.*

When modeling a deterministic function as the trajectory of a Gaussian process, assuming stationarity corresponds to considering that the deterministic function has the same nature (in terms of regularity and variation scale) in all the domain. This is the most classical case, that we consider in all the manuscript. Concerning non-stationary Gaussian processes, let us mention that they start to be proposed in operational Kriging packages, like DiceKriging ([RGD12]).

For a Gaussian process, stationarity is characterized in terms of conditions on the mean and covariance functions, presented in definition 2.12 and proposition 2.13.

Definition 2.12. *Let $\mathcal{D}_d = \{\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \in \mathcal{D}\}$. A covariance function K is stationary if it can be written, for any $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \in \mathcal{D}$, $K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = K(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})$, where, for convenience of notation, we use the same notation K for both a bivariable function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and a monovariable function $K : \mathcal{D}_d \to \mathbb{R}$.*

Proposition 2.13. *A Gaussian process is stationary if and only if its mean function is constant and its covariance function is stationary.*

Proof. If the mean function is constant and the covariance function is stationary, the Gaussian vectors $(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))$ and $(Y(\boldsymbol{x}^{(1)} + \boldsymbol{h}), ..., Y(\boldsymbol{x}^{(n)} + \boldsymbol{h}))$ have the same mean vector and covariance matrix. Hence, they have the same distribution.

Let m be the mean function. If there exists $\boldsymbol{h}$ so that $m(\boldsymbol{x}) \neq m(\boldsymbol{x} + \boldsymbol{h})$, then the random variables $Y(\boldsymbol{x})$ and $Y(\boldsymbol{x} + \boldsymbol{h})$ do not have the same mean. Let K be the covariance function. If there exists $\boldsymbol{h}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ so that $K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \neq K(\boldsymbol{x}^{(1)} + \boldsymbol{h}, \boldsymbol{x}^{(2)} + \boldsymbol{h})$, then the Gaussian vectors $(Y(\boldsymbol{x}^{(1)}), Y(\boldsymbol{x}^{(2)}))$ and $(Y(\boldsymbol{x}^{(1)} + \boldsymbol{h}), Y(\boldsymbol{x}^{(2)} + \boldsymbol{h}))$ do not have the same covariance matrix. □

We conclude the discussion on stationarity by presenting the Bochner's theorem, which states that any continuous stationary covariance function is the Fourier transform of a non-negative measure.

Theorem 2.14. *A continuous function $K : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it can be written as $K(\boldsymbol{x}) = \int_{\mathbb{R}^d} \mu(d\boldsymbol{\omega}) e^{\mathrm{i}\boldsymbol{\omega}.\boldsymbol{x}}$, where μ is a finite non-negative measure.*

A proof of theorem 2.14 is given in [GS74], p.208. Let us just mention that this result is intuitive since, for any $\alpha_1, ..., \alpha_n \in \mathbb{R}$, when K is positive definite

$$0 \leq \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \int_{\mathbb{R}^d} \mu(d\boldsymbol{\omega}) \left| \sum_{i=1}^{n} \alpha_i e^{\left(\mathrm{i}\boldsymbol{\omega}.\boldsymbol{x}^{(i)}\right)} \right|^2,$$

so that it makes sense that the measure μ is non-negative.

Regularity

Since continuity and differentiability are important features of deterministic functions, and since we aim at modeling deterministic functions as Gaussian process trajectories, the following notions of regularity for a Gaussian process are important.

The two most used notions are mean square regularity, and almost sure regularity of the trajectories. Roughly speaking, the former notion is the most convenient to handle mathematically, and the latter notion makes the most sense from an applied point of view. Indeed, as we see in definition 2.18, almost sure regularity is an information related to the only trajectory of a Gaussian process that the practitioner has at hand.

We now give the definitions of mean square continuity and mean square derivability.

Definition 2.15. *A Gaussian process Y is mean square continuous on $\mathcal{D}$ if, for any $\boldsymbol{x}^{(0)} \in \mathcal{D}$, $Y(\boldsymbol{x})$ goes to $Y(\boldsymbol{x}^{(0)})$ in the mean square sense when $\boldsymbol{x} \to \boldsymbol{x}^{(0)}$.*

Definition 2.16. *A Gaussian process Y is mean square differentiable on $\mathcal{D}$ if there exist d Gaussian processes $\frac{\partial}{\partial x_1} Y, ..., \frac{\partial}{\partial x_d} Y$ so that, for any $k \in \{1, ..., d\}, \boldsymbol{x}^{(0)} \in \mathcal{D}$, with $\boldsymbol{e}^{(k)}$ the k-th*

base vector, $\frac{Y(\boldsymbol{x}^{(0)}+h\boldsymbol{e}^{(k)})-Y(\boldsymbol{x}^{(0)})}{h}$ *goes to* $\frac{\partial}{\partial x_k}Y(\boldsymbol{x}^{(0)})$*, in the mean square sense, when the scalar* h *goes to zero.*

By induction, we then define the notion of multiple differentiability.

Definition 2.17. *A Gaussian process* Y *is* k *times mean square differentiable on* $\mathcal{D}$ *if it is* $k-1$ *times mean square differentiable, and for any* $i_1, ..., i_{k-1} \in \{1, ..., d\}$*, the Gaussian process* $\frac{\partial}{\partial x_{i_{k-1}}} ... \frac{\partial}{\partial x_{i_1}} Y$ *is mean square differentiable, with mean square derivative processes* $\frac{\partial}{\partial x_1}\frac{\partial}{\partial x_{i_{k-1}}} ... \frac{\partial}{\partial x_{i_1}} Y, ..., \frac{\partial}{\partial x_d}\frac{\partial}{\partial x_{i_{k-1}}} ... \frac{\partial}{\partial x_{i_1}} Y$*.*

In subsection 2.1.2, we will see that there is a simple relationship between the mean square regularity of the Gaussian Process and the regularity of its covariance function.

We now define the notions of almost sure regularity.

Definition 2.18. *A Gaussian process* Y *is almost surely continuous (*k *times differentiable) if, almost surely on the probability space* $(\Omega, \mathcal{F}, P)$*, the function* $\boldsymbol{x} \to Y(\omega, \boldsymbol{x})$ *is continuous (*k *times differentiable).*

Remark 2.19. *In definition 2.18, unless stated otherwise, differentiability is defined in the Frechet sense.*

2.1.2 The relationship between the covariance function and the trajectories of a Gaussian process

Relation between the regularity of the covariance function and the mean square regularity

The two following propositions give simple relationships between the mean square regularity and the regularity of the covariance function.

Proposition 2.20. *Let* $\mathcal{D} \subset \mathbb{R}^d$*. A centered Gaussian process* Y *is mean square continuous if and only if its covariance function is continuous at each pair* $(\boldsymbol{x}, \boldsymbol{x})$*,* $\boldsymbol{x} \in \mathcal{D}$*. Furthermore, if a covariance function is continuous at each pair* $(\boldsymbol{x}, \boldsymbol{x})$*,* $\boldsymbol{x} \in \mathcal{D}$*, then it is continuous on* $\mathcal{D} \times \mathcal{D}$*.*

Proof. The equivalence in proposition 2.20 is proved by writing $\mathbb{E}((Y(\boldsymbol{x}+\boldsymbol{h})-Y(\boldsymbol{x}))^2)$ in terms of the covariance function K. The second part is proved in [Adl81]. □

Proposition 2.21. *Let* $\mathcal{D} \subset \mathbb{R}^d$*. For a centered Gaussian process* Y*, for any* $k \in \mathbb{N}$*,* $i_1, ..., i_k \in \{1, ..., d\}$ *if the derivative function* $\frac{\partial^2}{\partial x_{i_1}\partial y_{i_1}} ... \frac{\partial^2}{\partial x_{i_k}\partial y_{i_k}} K$ *exists and is finite then* $\frac{\partial}{\partial x_{i_1}} ... \frac{\partial}{\partial x_{i_k}} Y$ *exists in the mean square sense and is a Gaussian process.*

Proof. The proof for $k = 1$ can be found in [CL67] for instance. The proof for $k > 1$ is done by induction on k by using

$$Cov\left(\frac{\partial Y}{\partial x_i}(\boldsymbol{x}^{(1)}), \frac{\partial Y}{\partial x_i}(\boldsymbol{x}^{(2)})\right) = \frac{\partial^2 K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})}{\partial x_i \partial y_i}. \tag{2.2}$$

The relation (2.2) is shown by writing

$$Cov\left(\frac{Y(\boldsymbol{x}^{(1)}+h\boldsymbol{e}^{(i)})-Y(\boldsymbol{x}^{(1)})}{h},\frac{Y(\boldsymbol{x}^{(2)}+h\boldsymbol{e}^{(i)})-Y(\boldsymbol{x}^{(2)})}{h}\right)=$$
$$\frac{1}{h^2}K(\boldsymbol{x}^{(1)}+h\boldsymbol{e}^{(i)},\boldsymbol{x}^{(2)}+h\boldsymbol{e}^{(i)})+\frac{1}{h^2}K(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)})$$
$$-\frac{1}{h^2}K(\boldsymbol{x}^{(1)}+h\boldsymbol{e}^{(i)},\boldsymbol{x}^{(2)})-\frac{1}{h^2}K(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}+h\boldsymbol{e}^{(i)}).$$

Now, if two random variables X_1 and X_2 converge in the mean square sense, their covariance converges to the covariance of their limits. Hence

$$Cov\left(\frac{\partial}{\partial x_i}Y(\boldsymbol{x}^{(1)}),\frac{\partial}{\partial x_i}Y(\boldsymbol{x}^{(2)})\right)=\frac{\partial}{\partial x_i}\frac{\partial}{\partial y_i}K(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}).$$

□

Concerning almost-sure regularity, we use the notion of a modification of a stochastic process.

Definition 2.22. *Let Y_1 and Y_2 be two stochastic processes on $\mathcal{D}$, with common probability space $(\Omega,\mathcal{F},P)$. Y_2 is a modification of Y_1 if, for all $\boldsymbol{x}\in\mathcal{D}$, $Y_1(\boldsymbol{x})=Y_2(\boldsymbol{x})$ almost surely.*

Concerning almost sure regularity, the two following propositions give sufficient conditions on the covariance function for a Gaussian process to be almost-surely continuous and almost-surely k times continuously differentiable.

Proposition 2.23, addressing almost sure continuity is proved in [Adl81].

Proposition 2.23. *Let Y be a Gaussian process on $\mathcal{D}\subset\mathbb{R}^d$, with covariance function K so that there exists $C<+\infty$ and $\epsilon>0$ so that, for $|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)}|$ small enough,*

$$K(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(1)})+K(\boldsymbol{x}^{(2)},\boldsymbol{x}^{(2)})-2K(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)})\leq\frac{C}{|\ln|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)}||^{1+\epsilon}}.$$

Then, there exists a Gaussian process $\tilde{Y}$, that is a modification of Y and that is almost surely continuous.

Remark 2.24. *In proposition 2.23, note that the Gaussian process at hand Y is not necessarily almost surely continuous. Only a second Gaussian process $\tilde{Y}$, that is a modification of Y, is so. This fact is illustrated in an elementary example in [Doo53]. The example is also presented in [Vaz05], chapter 2.1.*

Now in practice, for a Gaussian process verifying proposition 2.23, the almost surely continuous modification $\tilde{Y}$ always makes more sense than Y when Y is not almost surely continuous. Hence, we will always consider that we work with a modification of the Gaussian process Y having the most almost sure regularity. We will no longer mention this.

Because the condition in proposition 2.23 is expressed in terms of $\frac{1}{|\ln|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)}||^{1+\epsilon}}$, which vanishes very slowly when $|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)}|$ goes to 0, it is argued in [Abr97] that all continuous covariance functions can, in practice, be considered as yielding continuous trajectories almost surely.

Proposition 2.25 addresses sufficient conditions for a Gaussian process to be almost-surely k times differentiable. The interpretation is that it is sufficient that the covariance function be "a bit more" than $2k$ times continuously differentiable.

Proposition 2.25. *Let $k \in \mathbb{N}$. Let Y be a Gaussian process on $D_x \subset \mathbb{R}^d$, with covariance function K $2k$ times differentiable. Let, for any $i_1, ..., i_k \in \{1, ..., d\}$,*

$$\partial^2_{i_1,...,i_k} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \frac{\partial^2}{\partial x_{i_1} \partial y_{i_1}} ... \frac{\partial^2}{\partial x_{i_k} \partial y_{i_k}} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}).$$

Assume that, for $\alpha > 3$, and for $|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}|$ small enough,

$$\partial^2_{i_1,...,i_k} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}) + \partial^2_{i_1,...,i_k} K(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}) - 2\partial^2_{i_1,...,i_k} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \leq \frac{C}{|\ln |\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}||^\alpha}. \quad (2.3)$$

Then the Gaussian process Y is almost-surely k times continuously differentiable.

Proof. The proposition, for $d = 1$ corresponds to [CL67], p185. We are not aware of a corresponding multi-dimensional formulation in the literature. Hence, we give a short proof for consistency.

We show the proposition for $k = 1$, the case $k > 1$ is proved by induction using the same technique. Let $\boldsymbol{x}, \boldsymbol{h} \in \mathcal{D}$. The covariance function $\tilde{K}$ of the one-dimensional Gaussian process $t \to Y(\boldsymbol{x} + t\boldsymbol{h})$, defined on $[-a, a]$ for positive a small enough, verifies

$$\frac{\partial^2}{\partial x \partial y} \tilde{K}(t + dt, t + dt) + \frac{\partial^2}{\partial x \partial y} \tilde{K}(t, t) - 2\frac{\partial^2}{\partial x \partial y} \tilde{K}(t + dt, t) \leq \frac{C}{|\ln |dt||^\alpha}.$$

Hence, from [CL67], p185, the one-dimensional Gaussian process $t \to Y(\boldsymbol{x} + t\boldsymbol{h})$ is almost surely C^1 on $[-t, t]$.

Repeating the argument over $\boldsymbol{h}$ and $\boldsymbol{x}$, we show that Y is almost surely Gateaux differentiable. We have also shown that the Gateaux derivatives are almost-surely continuous. The Gateaux derivatives in the almost-sure sense and in the mean square sense are equal (they both correspond to the same limits in probability). The Gateaux derivatives in the mean square sense are linear, because K is two times Frechet differentiable. Hence, almost surely, Y is Gateaux differentiable, with linear and continuous Gateaux derivatives. Hence, Y is almost surely Frechet differentiable with continuous gradient. □

For a stationary Gaussian process on $\mathbb{R}$, the following proposition gives the simplest relation for mean square regularity, that can also be characterized in terms of the Fourier transform of the covariance function.

Proposition 2.26. *For a stationary Gaussian process on $\mathbb{R}$, the following assertions verify $i) \Rightarrow ii) \Rightarrow iii)$.*

i) The Fourier transform $\hat{K}$ of K (so that $K(x) = \int_{\mathbb{R}} \hat{K}(\omega) e^{\mathrm{i}\omega x} d\omega$) verifies

$$\int_{\mathbb{R}} \omega^{2k} \hat{K}(\omega) d\omega < +\infty.$$

ii) The covariance function K of Y is $2k$ times differentiable.

iii) Y is k times mean square differentiable.

Proof. From proposition 2.21 and the relation $Cov(\frac{\partial Y}{\partial x}(x_1), \frac{\partial Y}{\partial x}(x_2)) = -\frac{\partial^2 K(x_1 - x_2)}{\partial x^2}$. □

Hence we have the rule of thumb "$\omega^{2k}\hat{K}(\omega)$ is summable" implies "$K(x)$ is $2k$ times differentiable" implies "Y is k times mean square differentiable".

Matérn model on $\mathbb{R}$

Proposition 2.26 gives motivation for a covariance model where the regularity at zero is tunable, or equivalently where the vanishing rate at $+\infty$ of the Fourier transform of the covariance function is tunable. The Matérn model satisfies this, and its systematical use to model stationary Gaussian processes is hence recommended ([Ste99]). The Matérn model is parameterized by the hyper-parameters $\sigma^2 > 0$, $\ell > 0$ and $\nu > 0$ and defined by

$$\hat{K}(\omega) = \sigma^2 \frac{\Gamma\left(\nu + \frac{1}{2}\right)(2\sqrt{\nu})^{2\nu}}{\ell^{2\nu}\sqrt{\pi}\Gamma(\nu)} \frac{1}{(4\frac{\nu}{\ell^2} + \omega^2)^{\frac{1}{2}+\nu}} \tag{2.4}$$

We see that $\omega^{2k}\hat{K}(\omega)$ is summable whenever $\nu > k$. Therefore, in view of propositions 2.25 and 2.26, ν is called the smoothness hyper-parameter and Y is k times mean square differentiable and k times almost-surely differentiable whenever $\nu > k$. The two other hyper-parameters σ^2 and ℓ have, we find, a clearer interpretation after giving the temporal equivalent of (2.4),

$$K(x) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}x}{\ell}\right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}x}{\ell}\right), \tag{2.5}$$

where K_ν is a modified Bessel function ([AS65] p.374-379). In (2.5), the three hyper-parameters are as follow.

- σ^2 is the variance parameter. The parameterization is so that $K(0) = \sigma^2$. The larger σ^2 is, the larger the scale of the trajectories is, as illustrated in figure 2.1.
- ℓ is the correlation length hyper-parameter. The larger ℓ is, the more Y is correlated between two fixed points x_1 and x_2 and hence, the more the trajectories of Y vary slowly with respect to x. In figure 2.2, we illustrate this by plotting trajectories of centered Gaussian processes with varying ℓ for the covariance function.
- ν is the smoothness hyper-parameter. Y is k times mean square and almost surely differentiable whenever $\nu > k$. In figure 2.3, we plot trajectories of centered Gaussian processes with varying ν for the covariance function. It is clear that, the larger ν is, the smoother the trajectories are.

We conclude the presentation of the Matérn model in $\mathbb{R}$ by mentioning that the covariance has a simpler expression than in (2.5) when $\nu = k + \frac{1}{2}$, with integer k ([Ste99], p31). The limit $\nu \to +\infty$ also gives a simpler Gaussian form for the covariance. The Matérn covariance functions for $\nu = \frac{1}{2}$, $\nu = \frac{3}{2}$, $\nu = \frac{5}{2}$ and $\nu = +\infty$ are classical submodels, parameterized by σ^2 and ℓ and are called the exponential, Matérn $\frac{3}{2}$, Matérn $\frac{5}{2}$ and Gaussian correlation function. In table 2.1, we give the expressions of these submodels.

Remark 2.27. *$\nu = \frac{1}{2}$ actually corresponds to $K(x) = \sigma^2 e^{-\sqrt{2}\frac{|x|}{\ell}}$. Nevertheless, we define the exponential model by $K(x) = \sigma^2 e^{-\frac{|x|}{\ell}}$ for convenience.*

Remark 2.28. *The fact that, when $\nu \to +\infty$, the Matérn model with hyper-parameters (σ^2, ℓ, ν) converges to the Gaussian model with hyper-parameters (σ^2, ℓ) is worth insisting on. Indeed, it means that a given value of the hyper-parameter ℓ has the same impact (in terms of scale of variation for the Gaussian process) regardless of the value of ν. Thus, in the Matérn model of*

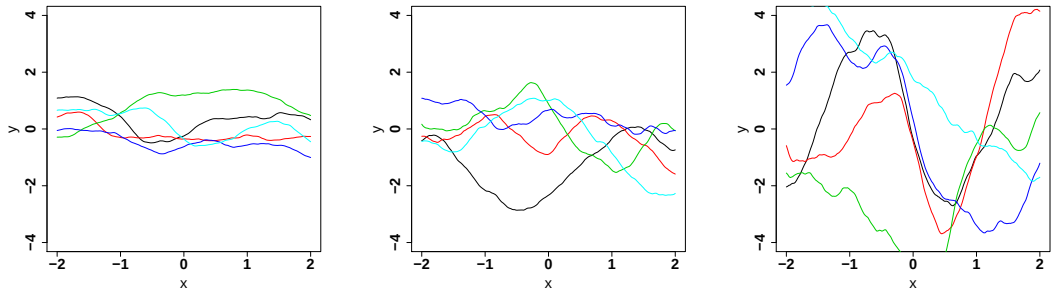

Figure 2.1: Influence of the variance hyper-parameter for the Matérn model of (2.5). Plot of trajectories of Gaussian processes with the Matérn covariance function with correlation length $\ell = 1$, smoothness parameter $\nu = \frac{3}{2}$ and variance $\sigma^2 = \frac{1}{2}, 1, 2$ from left to right.

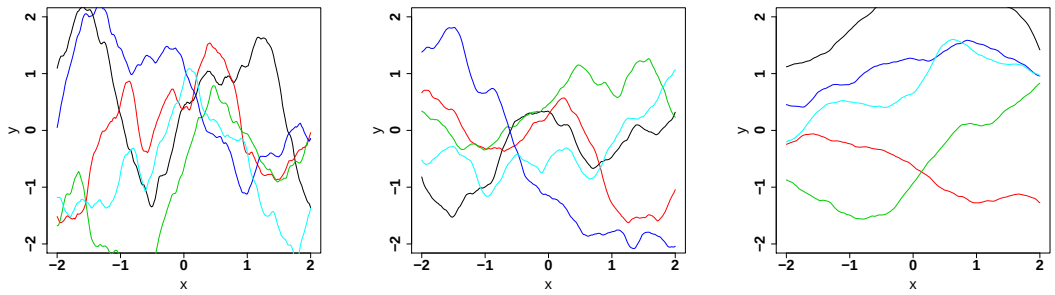

Figure 2.2: Influence of the correlation length for the Matérn model of (2.5). Plot of trajectories of Gaussian processes with the Matérn covariance function with variance $\sigma^2 = 1$, smoothness parameter $\nu = \frac{3}{2}$ and correlation length $\ell = \frac{1}{2}, 1, 2$ from left to right.

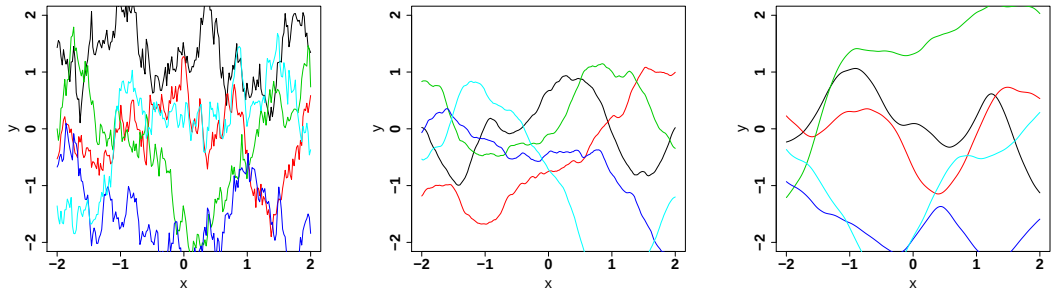

Figure 2.3: Influence of the smoothness parameter for the Matérn model of (2.5). Plot of trajectories of Gaussian processes with the Matérn covariance function with variance $\sigma^2 = 1$, correlation length $\ell = 1$ and smoothness parameter $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$ from left to right.

submodel name	corresponding ν	expression
exponential	$\frac{1}{2}$	$\sigma^2 e^{-\frac{\lvert x\rvert}{\ell}}$
Matérn $\frac{3}{2}$	$\frac{3}{2}$	$\sigma^2(1+\sqrt{6}\frac{\lvert x\rvert}{\ell})e^{-\sqrt{6}\frac{\lvert x\rvert}{\ell}}$
Matérn $\frac{5}{2}$	$\frac{5}{2}$	$\sigma^2(1+\sqrt{10}\frac{\lvert x\rvert}{\ell}+\frac{10}{3}\frac{\lvert x\rvert^2}{\ell^2})e^{-\sqrt{10}\frac{\lvert x\rvert}{\ell}}$
Gaussian	$+\infty$	$\sigma^2 e^{-\frac{x^2}{\ell^2}}$

Table 2.1: Expressions of the exponential, Matérn $\frac{3}{2}$, Matérn $\frac{5}{2}$ and Gaussian covariance functions on $\mathbb{R}$ and corresponding smoothness parameter ν of the Matérn model in (2.5). $\nu=\frac{1}{2}$ actually corresponds to $K(x)=\sigma^2 e^{-\sqrt{2}\frac{\lvert x\rvert}{\ell}}$ but we define the exponential model by $K(x)=\sigma^2 e^{-\frac{\lvert x\rvert}{\ell}}$ for convenience.

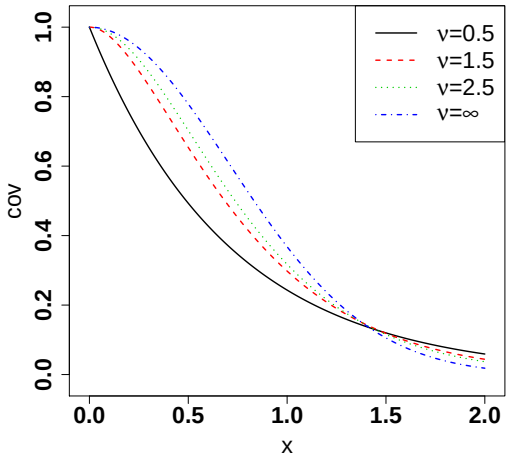

Figure 2.4: Plot of the Matérn covariance function with $\sigma^2=1$, $\ell=1$ and $\nu=\frac{1}{2}$, $\nu=\frac{3}{2}$, $\nu=\frac{5}{2}$ and $\nu=\infty$. Remark: for $\nu=\frac{1}{2}$, we plot $t\to\sigma^2 e^{\left(-\sqrt{2}\frac{\lvert t\rvert}{\ell}\right)}$.

(2.5), *the three hyper-parameters* σ^2, ℓ *and* ν *impact respectively on the variance, the scale of variation and the regularity of the Gaussian process and the three effects are rather independent. We illustrate this in figure 2.4, where we plot the Matérn covariance function with* $\sigma^2=1$, $\ell=1$ *and* $\nu=\frac{1}{2}$, $\nu=\frac{3}{2}$, $\nu=\frac{5}{2}$ *and* $\nu=\infty$. *We see that only the regularity at zero of the covariance function varies. Notably, the hyper-parameter* $\ell=1$ *has the same impact on the global decreasing rate of the covariance function for all the values of* ν.

Thus, although (2.5) *seems rather complicated, the three hyper-parameters have simple interpretations. Notice finally that if in* (2.5) *we tried to "simplify" the covariance function by using, say,* $\left(\frac{x}{\ell}\right)^{\nu} K_{\nu}\left(\frac{x}{\ell}\right)$, *a given value of the hyper-parameter* ℓ *would have a strongly different impact when* ν *is small and when* ν *is large. Thus, the interpretation of the hyper-parameters* ℓ *and* ν *would not be independent and would, therefore, be significantly less intuitive.*

Multi-dimensional Matérn model

We now generalize the Matérn model for dimension $d > 1$. There are two methods for doing so, defining two different Matérn models in dimension $d > 1$. Both are parameterized by $\sigma^2 > 0$, $\ell_1, ..., \ell_d > 0$ and $\nu > 0$.

The first model is the isotropic Matérn model. It is defined by

$$K(\boldsymbol{x}) = \sigma^2 K_{m,\nu}\left(\sqrt{\left[\sum_{i=1}^{d} \frac{x_i^2}{\ell_i^2}\right]}\right), \tag{2.6}$$

where $K_{m,\nu}$ is the one-dimensional Matérn covariance function of (2.5) with variance $\sigma^2 = 1$, correlation length $\ell = 1$ and smoothness parameter ν. This model is called isotropic because the Gaussian process parameterized by $(\frac{x_1}{\ell_1}, ..., \frac{x_d}{\ell_d})$ is isotropic, as defined in definition 2.29 ([Ste99], p17).

Definition 2.29. *A stationary Gaussian process Y is isotropic if, for any orthogonal matrix $\mathbf{M}$, the distribution of the Gaussian process $\boldsymbol{x} \to Y(\mathbf{M}\boldsymbol{x})$ is the same as the distribution of the Gaussian process Y.*

To see that, for a centered Gaussian process Y, with isotropic Matérn covariance function with hyper-parameters $\ell_1, ..., \ell_d, \nu$, the Gaussian process $(\frac{x_1}{\ell_1}, ..., \frac{x_d}{\ell_d}) \to Y(x_1, ..., x_d)$ is isotropic, note that its covariance function is $K(\boldsymbol{h}) = \tilde{K}(|\boldsymbol{h}|)^2$, which is a sufficient condition for isotropy because $|\boldsymbol{h}|^2 = |\mathbf{M}\boldsymbol{h}|^2$ for any orthogonal matrix $\mathbf{M}$.

The second model is the tensorized Matérn model. It is defined by

$$K(\boldsymbol{x}) = \sigma^2 \prod_{i=1}^{d} K_{m,\nu}\left(\frac{x_i}{\ell_i}\right), \tag{2.7}$$

where $K_{m,\nu}$ is the one-dimensional Matérn covariance function of (2.5) with variance $\sigma^2 = 1$, correlation length $\ell = 1$ and smoothness parameter ν.

For both versions of the multi-dimensional Matérn model, the hyper-parameters $\sigma^2 > 0$, $\ell_1, ..., \ell_d > 0$ and $\nu > 0$ have the same interpretation.

- σ^2 is the variance hyper-parameter. Its interpretation is the same as for the one-dimensional case of (2.5).
- ν is the smoothness hyper-parameter. For $\nu > k$, the covariance functions of (2.6) and (2.7) are $2k$ times differentiable, so that because of propositions 2.21 and 2.25 the Gaussian process Y is k times mean square and almost surely differentiable.
- $\ell_1, ..., \ell_d$ are the correlation length hyper-parameters corresponding to the d components. ℓ_k corresponds, similarly to the one-dimensional case of (2.5), to the scale of variation of the Gaussian process Y relatively to the component x_k. Consider that $\mathcal{D} = [0, 1]^d$, so that the scale of the correlation lengths are comparable. If one of the correlation lengths ℓ_k is significantly larger than at least one of the others, then it boils down to considering that the trajectories of the Gaussian process Y almost do not depend on x_k. If all the correlation lengths are significantly larger than one, then this corresponds to a Gaussian process that has the distribution of a constant function whose constant value follows a Gaussian distribution.

The two versions of the multi-dimensional Matérn model seem to appear rather equally in the literature. In his monograph [Ste99] p55, Stein criticizes the use of the tensorized version because of its strong dependence with respect to the system of axes. The choice of axes is indeed rather arbitrary for natural data (whose case is the context of Stein's remark). Nevertheless, they may have more meaning for the analysis of computer experiments, where they correspond to quantities of different nature. On the other hand, in the packages PErK ([SWN03], appendix C) or DICE Kriging ([RGD12]), tensorized Matérn covariance functions are proposed (where they are called separable). In our work, we have used both versions of the multi-dimensional Matérn model.

Other covariance models

The other covariance model we have used is the power-exponential model, parameterized by $\sigma^2 > 0$, $\ell_1, ..., \ell_d > 0$ and $0 < p \leq 2$. It is defined by

$$K(\boldsymbol{x}) = \sigma^2 \prod_{i=1}^{d} e^{\left(-\left|\frac{x_i}{\ell_i}\right|^p\right)}, \tag{2.8}$$

The hyper-parameters σ^2 and $\ell_1, ..., \ell_d$ have the same interpretation as for the multidimensional Matérn model. However, the power parameter p gives less flexibility concerning the smoothness of the Gaussian process Y. Indeed Y is infinitely mean square and almost surely differentiable for $p = 2$ and only mean square and almost surely continuous for $0 < p < 2$.

Other classical covariance functions exist in the literature, that we have not used in this work. We refer e.g. to [Abr97] or section 4.2 of [RW06].

2.2 Prediction and conditional simulation for Gaussian processes

2.2.1 Ordinary, simple and universal Kriging models

A Kriging model [Mat70] consists in inferring the values of a random field Y at unobserved points given observations of Y at other points. Hence, in the manuscript, we work in a Kriging framework, with the additional assumption that the random field Y is Gaussian.

Until now, we have focused our attention on the covariance function of the Gaussian process Y. The assumptions made on the mean function can be important as well, although they are generally less important than for the covariance function ([Ste99], p138).

There are three subcases of Kriging model, depending on the assumption made on the mean function of Y.

In **Simple Kriging**, the mean function is assumed to be known. Equivalently, when working in the simple Kriging framework, we will consider a centered Gaussian process Y.

In **Ordinary Kriging**, the mean function is assumed to be constant and unknown.

In **Universal Kriging**, the mean function at $\boldsymbol{x} \in \mathcal{D}$ is assumed to be of the form $\sum_{i=1}^{m} \beta_i g_i(\boldsymbol{x})$, with known functions g_i and unknown scalar coefficients β_i.

As we see in subsection 2.2.2, $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^t$ can be estimated by Maximum Likelihood in the case of simple Kriging, and by Maximum Likelihood or a Bayesian method with Gaussian prior distribution in the cases of ordinary and universal Kriging.

2.2.2 Point-wise prediction

In this subsection, the Gaussian process Y is observed at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, with observed values $\boldsymbol{y} = (y_1, ..., y_n) = (Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))$. We want to predict the value of Y at a fixed point $\boldsymbol{x}^{(new)}$. We hence denote $\mathbf{K}$ as the $n \times n$ covariance matrix of Y at $(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)})$ and $\boldsymbol{k}(\boldsymbol{x}^{(new)})$ as the $n \times 1$ covariance vector of Y between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}^{(new)}$. All the formulas presented in this subsection 2.2.2 can be found, for instance in [SWN03].

Case of simple Kriging

In the case of simple Kriging, we call prediction, or predictive mean, the conditional mean of $Y(\boldsymbol{x}^{(new)})$ according to $\boldsymbol{y}$, given by theorem 2.5,

$$\hat{y}(\boldsymbol{x}^{(new)}) := \mathbb{E}(Y(\boldsymbol{x}^{(new)})|\boldsymbol{y}) = \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1} \boldsymbol{y}. \tag{2.9}$$

We call predictive variance the conditional variance (theorem 2.5),

$$\hat{\sigma}^2(\boldsymbol{x}^{(new)}) := Var(Y(\boldsymbol{x}^{(new)})|\boldsymbol{y}) = Var(Y(\boldsymbol{x}^{(new)})) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1} \boldsymbol{k}(\boldsymbol{x}^{(new)}). \tag{2.10}$$

Remark 2.30. *In the case where* $\mathbf{K}$ *is singular, we refer to remark 2.6 for the definition and computation of* (2.9) *and* (2.10). *Roughly speaking it is sufficient to replace* $\mathbf{K}^{-1}$ *by a pseudo-inverse of* $\mathbf{K}$*, which is equivalent to obtaining, from the redundant Gaussian vector* $\boldsymbol{y}$*, a lower-dimensional non-degenerate Gaussian vector incorporating all the randomness of* $\boldsymbol{y}$. *This remark holds for* (2.11)-(2.18). *In the sequel, we do not make the remark anymore, and we assume that* $\mathbf{K}$ *is non-singular.*

We make the following remarks for (2.9) and (2.10).

- The prediction of (2.9) is a linear function of the Gaussian vector $\boldsymbol{y}$. For Gaussian vectors, the conditional mean indeed coincides with a linear prediction.
- The observations being fixed, the prediction of (2.9) can be written as

  $$\sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(new)}),$$

  with K the covariance function of Y. As the classical covariance functions are decreasing functions of the distance between $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(new)}$, the prediction function $\boldsymbol{x}^{(new)} \to \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(new)})$ vanishes when $\boldsymbol{x}^{(new)}$ is far from the observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. Hence the prediction of (2.9) is essentially meant for interpolation.
- When $\boldsymbol{x}^{(new)} = \boldsymbol{x}^{(i)}$ for a particular i, we can show that in (2.9) and (2.10), $\hat{y}(\boldsymbol{x}^{(i)}) = y_i$ and $\hat{\sigma}^2(\boldsymbol{x}^{(i)}) = 0$. This is expected since, when predicting a value that we know, the prediction is the value itself and the associated uncertainty (the predictive variance) is zero.

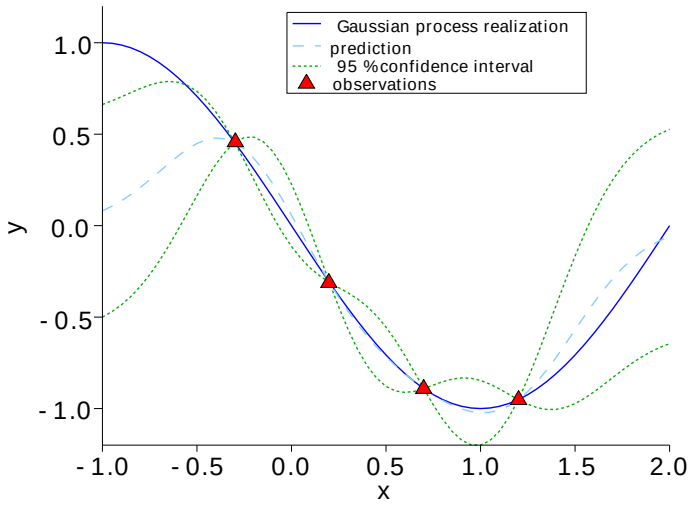

Figure 2.5: Illustration of the simple Kriging prediction of (2.9) and (2.10). The function $sin(\frac{\pi x}{2})$ is assumed to be a trajectory of a Gaussian process with Gaussian covariance function with $\sigma^2 = 0.3^2$ and $\ell = \frac{1}{2}$. 95% confidence intervals are of the form $[\hat{y}(\boldsymbol{x}^{(new)}) - 1.96\hat{\sigma}(\boldsymbol{x}^{(new)}), \hat{y}(\boldsymbol{x}^{(new)}) + 1.96\hat{\sigma}(\boldsymbol{x}^{(new)})]$.

In fact, we have the stronger result $\mathcal{L}\left(Y(\boldsymbol{x}^{(new)})|\boldsymbol{y}\right) = \mathcal{N}(\hat{y}(\boldsymbol{x}^{(new)}), \hat{\sigma}^2(\boldsymbol{x}^{(new)}))$, where $\mathcal{N}(m, \sigma^2)$ is the Gaussian distribution with mean m and variance σ^2. Thus, we can build confidence intervals for $Y(\boldsymbol{x}^{(new)})$, for instance 95% confidence intervals of the form $[\hat{y}(\boldsymbol{x}^{(new)}) - 1.96\hat{\sigma}(\boldsymbol{x}^{(new)}), \hat{y}(\boldsymbol{x}^{(new)}) + 1.96\hat{\sigma}(\boldsymbol{x}^{(new)})]$.

In figure 2.5, we give a one-dimensional illustration of (2.9) and (2.10). We observe that, as discussed, the prediction interpolates the known values exactly, the confidence intervals have length zero at the known value points, their widths increase when going away from known value points, and the prediction function goes to zero as we go in the extrapolation domain (away from all the known value points).

Case of ordinary or universal Kriging

For the case of ordinary or universal Kriging, we denote by $\mathbf{H}$ the $n \times m$ matrix so that $H_{i,j} = g_j(\boldsymbol{x}^{(i)})$, where the mean function is assumed to be of the form $\sum_{i=1}^{m} \beta_i g_i$, with known functions g_i and unknown coefficients β_i. We also denote by $\boldsymbol{h}$ the $m \times 1$ vector so that $h_j = g_j(\boldsymbol{x}^{(new)})$.

We distinguish two cases for the coefficient vector $\boldsymbol{\beta}$. We call the frequentist case, or no prior information case, the case where $\boldsymbol{\beta}$ is an unknown constant. We call Bayesian case, or prior information case, the case where $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_{prior}, \mathbf{Q}_{prior})$, with known a priori mean vector $\boldsymbol{\beta}_{prior}$ and covariance matrix $\mathbf{Q}_{prior}$. We refer, e.g. to [Rob01] for an introduction to Bayesian statistics.

In the frequentist case, the Maximum Likelihood estimator of $\boldsymbol{\beta}$ is, after writing a zero-gradient condition,

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{y}. \tag{2.11}$$

The estimator in (2.11) is unbiased and, after a direct calculation, has covariance matrix

$$Cov(\hat{\boldsymbol{\beta}}) = (\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}. \tag{2.12}$$

Remark 2.31. *In* (2.11) *and* (2.12)*, the matrix* $(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})$ *is assumed to be well-defined and non-singular.*

We will discuss here the case where $\mathbf{K}$ *is non-singular so that the matrix* $(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})$ *is well-defined. Hence,* $(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})$ *is singular if and only if* $\mathbf{H}$ *does not have a full rank (we consider, in the manuscript* $m < n$*, which corresponds to all our application cases).*

If $\mathbf{H}$ *is not of full-rank, let* m' *be its rank. Then we can write* $\mathbb{R}^m$ *as* $E_1 \bigoplus E_2$ *(meaning that each element of* $\mathbb{R}^m$ *is the sum of two unique elements of* E_1 *and* E_2*), where* E_1 *has dimension* m'*,* E_2 *has dimension* $m - m'$ *and* E_1, E_2 *satisfy* $\mathbf{H}E_1$ *has dimension* m' *and* $\mathbf{H}E_2 = \{0\}$*. Roughly speaking* E_1 *corresponds to the part of* $\boldsymbol{\beta}$ *that has an impact for the regression function on the points* $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ *and* E_2 *corresponds to the part of* $\boldsymbol{\beta}$ *that has no impact on the regression function on the points* $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$*. Hence the* E_1 *part can be estimated while the* E_2 *part can not. From this fact, two frameworks are possible.*

First, if, for a prediction point $\boldsymbol{x}^{(new)}$*, the* $m \times 1$ *vector* $\boldsymbol{h}^{(new)}$ *defined by* $h_i^{(new)} = g_i(\boldsymbol{x}^{(new)})$ *verifies* $\left(\boldsymbol{h}^{(new)}\right)^t E_2 \neq \{0\}$*, then the* E_2 *component of* $\boldsymbol{\beta}$ *has a non-zero and totally non-quantifiable impact on the value of* $Y(\boldsymbol{x}^{(new)})$*. In this case, it is impossible to predict* $Y(\boldsymbol{x}^{(new)})$ *because the design of experiments totally ignores some aspects of the regression function that impact the prediction points. The only way to solve this first issue is to add well-chosen points to the design of experiments. We will assume that this has been done in all the sequel. A typical example for this first issue is when the regression model is of the form* $g(x, y) = \beta_1 x + \beta_2 y$*, when all the observation points are of the form* $(x_1, 0), ..., (x_n, 0)$*, and when we want to predict at the* (x_{new}, y_{new}) *with non-zero* y_{new}*.*

The second framework is when, for all prediction points $\boldsymbol{x}^{(new)}$*,* $\boldsymbol{h}^{(new)}$ *does verify* ${\boldsymbol{h}^{(new)}}^t E_2 = \{0\}$*. In this case, the* E_2 *component of* $\boldsymbol{\beta}$ *is both inestimable and has no impact on prediction. Hence, it shall simply be ignored. To do so, let* E_1*, of dimension* m'*, be parameterized bijectively by the linear application* $\mathbf{P} : \mathbb{R}^{m'} \to E_1$*. Then we can set* $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{m'}$ *and* $\tilde{H}$ *as the linear application from* $\mathbb{R}^{m'}$ *to* $\mathbb{R}^n$ *defined as* $\tilde{H} = \mathbf{H} \circ \mathbf{P}$*. For a new prediction point, we also use* $(\tilde{\boldsymbol{h}}^{(new)})^t = (\boldsymbol{h}^{new})^t \circ \mathbf{P}$*. This second case corresponds to an unidentifiability, or to an over-parameterization for* $\boldsymbol{\beta}$*. A typical example for this second case is when the regression model is of the form* $g(x) = \beta_1 x + \beta_2 x$*.*

We now explain, in practice, how to compute E_1 *and* E_2 *for* $\boldsymbol{\beta}$*, and how to proceed with the reparameterization described above. Consider a SVD,* $\mathbf{H} = \mathbf{USV}^t$ *with* $n \times m$ *matrix* $\mathbf{U}$ *so that* $\mathbf{U}^t\mathbf{U} = \mathbf{I}_m$*,* $m \times m$ *matrix* $\mathbf{V}$ *so that* $\mathbf{V}^t\mathbf{V} = \mathbf{I}_m$ *and diagonal matrix* $\mathbf{S}$ *with diagonal values* $\lambda_1 > ... > \lambda_{m'} > \lambda_{m'+1} = ... = \lambda_m = 0$*. Then, let, with* $m \times m'$ *matrix* $\mathbf{V}_{m'}$*,*

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{m'} & \mathbf{V}_{m-m'} \end{pmatrix}.$$

With $\mathbf{0}_{a,b}$ *the* $a \times b$ *zero matrix, the spaces* $E_1 := (\mathbf{V}_{m'}, \mathbf{0}_{m,m-m'})\mathbb{R}^m$ *and* $E_2 := (\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'})\mathbb{R}^m$ *verify* $\mathbf{H}E_2 = \{0\}$ *and* $\mathbf{H}E_1$ *has dimension* m'*. Now, if for a prediction point* $\boldsymbol{x}^{(new)}$*,* $\boldsymbol{h}^{(new)}$ *verifies* ${\boldsymbol{h}^{(new)}}^t(\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'}) \neq 0$*, the design of experiments needs to be completed as explained above. If, for all prediction points* $\boldsymbol{x}^{(new)}$*,* $\boldsymbol{h}^{(new)}$ *verifies* ${\boldsymbol{h}^{(new)}}^t(\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'}) = 0$*,*

the solution is to use a m'-dimensional regression parameter $\tilde{\boldsymbol{\beta}}$ that is non-redundant. For this, note that for $\boldsymbol{w} \in E_1$ of the form $\boldsymbol{w} = (\mathbf{V}_{m'}, \mathbf{0}_{m,m-m'})\boldsymbol{v}$, $\boldsymbol{v} \in \mathbb{R}^m$,

$$\mathbf{H}\boldsymbol{w} = \mathbf{H}(\mathbf{V}_{m'}, \mathbf{0}_{m,m-m'})\boldsymbol{v} = \mathbf{U}_{m'}\mathbf{S}_{m'}\boldsymbol{v}_{m'},$$

with $\mathbf{U}_{m'}$ the matrix of the m' first columns of $\mathbf{U}$, $\mathbf{S}_{m'}$ the diagonal matrix of the m' non-zero diagonal elements of $\mathbf{S}$ and $\boldsymbol{v}_{m'}$ the vector of the m' first components of $\boldsymbol{v}$. Said differently, with

$$\tilde{H} = \mathbf{U}_{m'}\mathbf{S}_{m'}$$

on the design of experiments, the regression function family $(\boldsymbol{\beta} \to \mathbf{H}\boldsymbol{\beta})_{\boldsymbol{\beta}\in\mathbb{R}^m}$ is the same as the regression function family $(\tilde{\boldsymbol{\beta}} \to \tilde{\mathbf{H}}\tilde{\boldsymbol{\beta}})_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^{m'}}$. For a new prediction point, the corresponding reparameterization is

$$(\tilde{\boldsymbol{h}}^{(new)})^t = (\boldsymbol{h}^{(new)})^t(\mathbf{V}_{m'}, \mathbf{0}_{m,m-m'}).$$

Hence, we have shown, from a SVD of $\mathbf{H}$, how to solve the case when $\mathbf{H}$ is singular. In the case where $\mathbf{H}$ is non-singular but ill-conditioned, we can adopt the same techniques. Assume that the $m - m'$ last eigenvalues of $\mathbf{S}$ are very small compared to the other ones. Then, if $(\boldsymbol{h}^{(new)})^t(\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'})$ is much larger than the lines of $\mathbf{H}(\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'})$, the design of experiments is numerically incomplete, and the prediction task is strongly compromised. If $(\boldsymbol{h}^{(new)})^t(\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'})$ is not larger than the lines of $\mathbf{H}(\mathbf{0}_{m,m'}, \mathbf{V}_{m-m'})$, then we can set the $m - m'$ last eigenvalues of $\mathbf{S}$ to zero and proceed as described above.

In the sequel, we will not discuss the singularity issues again. We will assume that $\mathbf{K}$ is non-singular and that $\mathbf{H}$ has a full rank.

We see in (2.11) that if there is a $\boldsymbol{\beta}$ so that $\mathbf{H}\boldsymbol{\beta} = \boldsymbol{y}$, then we have $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. This means that, if we are in the favorable case when the mean function model can perfectly reproduce the known values, then the estimation of the mean function will achieve this perfect reproduction, as should be expected. Finally, as the random vector $\hat{\boldsymbol{\beta}}$ has Gaussian distribution, its covariance matrix (2.12) is sufficient to yield confidence ellipsoids for $\boldsymbol{\beta}$.

In the Bayesian case, the posterior distribution of $\boldsymbol{\beta}$ given the known values $\boldsymbol{y}$ is Gaussian with mean vector

$$\boldsymbol{\beta}_{post} = \boldsymbol{\beta}_{prior} + (\mathbf{Q}_{prior}^{-1} + \mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{H}\boldsymbol{\beta}_{prior}), \tag{2.13}$$

and covariance matrix

$$\mathbf{Q}_{post} = (\mathbf{Q}_{prior}^{-1} + \mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}. \tag{2.14}$$

We refer to [SWN03], section 4 for the proof of (2.13) and (2.14). We can note that, when $\mathbf{Q}_{prior}^{-1} \to 0$, then the Bayesian estimation of $\boldsymbol{\beta}$ tends to the frequentist one. This is an intuitive fact, because $\mathbf{Q}_{prior}^{-1}$ small corresponds to a small a priori knowledge of $\boldsymbol{\beta}$ and hence should, in the limit case, correspond to an absence of knowledge.

We now present the formulas for the prediction at a new point $\boldsymbol{x}^{(new)}$. We denote $\boldsymbol{h}(\boldsymbol{x}^{(new)})$ the vector of the regression functions at $\boldsymbol{x}^{(new)}$ defined by $\left(\boldsymbol{h}(\boldsymbol{x}^{(new)})\right)_i = h_i(\boldsymbol{x}^{(new)})$. We denote $\boldsymbol{k}(\boldsymbol{x}^{(new)})$ the $n \times 1$ covariance vector of Y between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}^{(new)}$.

In the frequentist case, the Best Linear Unbiased Predictor (BLUP) of $Y(\boldsymbol{x}^{(new)})$ with respect to the vector of observations $\boldsymbol{y}$ (that we also call prediction) is

$$\hat{y}(\boldsymbol{x}^{(new)}) = (\boldsymbol{h}(\boldsymbol{x}^{(new)}))^t\hat{\boldsymbol{\beta}} + (\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{\beta}}), \tag{2.15}$$

with $\hat{\boldsymbol{\beta}}$ given by (2.11).

We refer to [SWMW89] for a detailed definition of the BLUP and the proof of (2.15). The prediction can be interpreted as the conditional mean of (2.9), in which the true and unknown value $\boldsymbol{\beta}$ is replaced by its estimation $\hat{\boldsymbol{\beta}}$. Otherwise, we can make the same remarks for (2.15) as for (2.9).

The mean square error of the BLUP is (see section 4 of [SWN03] for a proof)

$$\begin{aligned}\hat{\sigma}^2(\boldsymbol{x}^{(new)}) &:= \mathbb{E}\left[\left(Y(\boldsymbol{x}^{(new)}) - \hat{y}(\boldsymbol{x}^{(new)})\right)^2\right] \\ &= Var(Y(\boldsymbol{x}^{(new)})) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)}) \\ +(\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t&(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}(\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)})).\end{aligned} \tag{2.16}$$

Since only linear combinations have been used, the BLUP has Gaussian distribution and the mean square error allows to build confidence intervals. The predictive variance (2.16) can be interpreted as the conditional variance in (2.10), plus a non-negative term due to the uncertainty on the estimation of $\boldsymbol{\beta}$.

In the Bayesian case, the posterior distribution of $Y(\boldsymbol{x}^{(new)})$ given the observations $\boldsymbol{y}$ is Gaussian with mean

$$\hat{y}(\boldsymbol{x}^{(new)}) = (\boldsymbol{h}(\boldsymbol{x}^{(new)}))^t\boldsymbol{\beta}_{post} + (\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t\mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{H}\boldsymbol{\beta}_{post}), \tag{2.17}$$

and variance

$$\begin{aligned}&\hat{\sigma}^2(\boldsymbol{x}^{(new)}) = Var(Y(\boldsymbol{x}^{(new)})) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)}) \\ +\quad &(\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H} + \mathbf{Q}_{prior}^{-1})^{-1}(\boldsymbol{h} - \mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)})).\end{aligned} \tag{2.18}$$

Equations (2.17) and (2.18) are also proved in section 4 of [SWN03]. We can make the same remarks for these equations as for (2.15) and (2.16). Similarly to the estimation of $\boldsymbol{\beta}$, the limit when $\mathbf{Q}_{prior}^{-1} \to 0$ of the prediction in the Bayesian case is the prediction in the frequentist case.

In figures 2.6 and 2.7, we give a one-dimensional illustrative example of the estimation of $\boldsymbol{\beta}$ and the prediction and predictive variance. The function $x \to x^2$ on $[0, 1]$ is assumed to be the trajectory of a Gaussian process, with mean function of the form $\beta_0 + \beta_1 x$ and Gaussian covariance function with $\sigma = 0.3$ and $\ell = 0.5$. In the Bayesian case, the a priori distribution of $\boldsymbol{\beta}$ is Gaussian with mean vector $(0.2, 0.1)^t$ and diagonal covariance matrix, with diagonal vector $(0.09, 0.09)^t$. The values of Y are known at the points 0.2, 0.5 and 0.8.

In figure 2.6, we consider the frequentist case. We first see that there is a negative correlation in the estimation of $\boldsymbol{\beta}$. This correlation can be interpreted. Indeed if β_0, the value at 0 of the line $x \to \beta_0 + \beta_1 x$ is increased, then, for the line to remain close to the parabola $x \to x^2$, the slope of the line (β_1) must be decreased. Furthermore, an important remark is that the estimated line is above and does not go through the three known value points. This is surprising at first sight, all the more so since a least square estimator of $\boldsymbol{\beta}$ would go through the three points. This is because, as it is shown in (2.15), the estimated line is not intended to constitute a predictive model of the parabola. Indeed it is completed by the inferred deviation from the mean function,

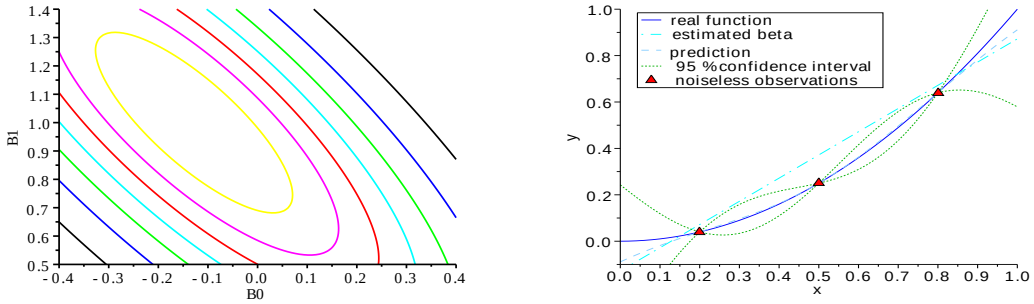

Figure 2.6: Estimation of $\boldsymbol{\beta}$ and prediction in the frequentist case. The function $x \to x^2$ on $[0,1]$ is assumed to be the trajectory of a Gaussian process, with mean function of the form $\beta_0 + \beta_1 x$ and Gaussian covariance function with $\sigma = 0.3$ and $\ell = 0.5$. Left: Iso-density curves of the probability density function for the estimation of $\boldsymbol{\beta}$, given by (2.11) and (2.12). Right: Estimated line (2.11), real parabola, prediction (2.17) and 95% confidence intervals (2.18) of the form $[\hat{y}(\boldsymbol{x}^{(new)}) - 1.96\hat{\sigma}(\boldsymbol{x}^{(new)}), \hat{y}(\boldsymbol{x}^{(new)}) + 1.96\hat{\sigma}(\boldsymbol{x}^{(new)})]$.

from the three known value points. We see in figure 2.6 that the prediction curve approximates almost perfectly the parabola. Let us also note that in the extrapolation region ($0 \leq x \leq 0.2$ and $0.8 \leq x \leq 1$), the estimated line approximates better the parabola than a line which would go between the three observation points.

In figure 2.7, we consider the Bayesian case. By looking at the right plot, we can see that, from the prior $\boldsymbol{\beta}$ to the posterior $\boldsymbol{\beta}$, the line goes substantially closer to the three observation points. Nevertheless, it is not as close as in the frequentist case. This is a classical case in the Bayesian case (as well as in Bayesian statistics in general), when the known value points and the a priori distribution are in disagreement, the posterior mean of $\boldsymbol{\beta}$ is a compromise between the frequentist estimate and the prior mean. Looking on the left plot, we see that a negative correlation between the two components of $\boldsymbol{\beta}$ appears in the posterior distribution of $\boldsymbol{\beta}$.

Finally, the conclusion concerning the prediction and the predictive variances are the same as for figure 2.5.

Case of noisy observations

Assume that the observations of the Gaussian process are noisy. Formally, this corresponds to considering that the observation at $\boldsymbol{x}^{(i)}$ is $y_i = Y(\boldsymbol{x}^{(i)}) + \epsilon_i$, where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^t$ is a centered Gaussian vector with covariance matrix $\mathbf{K}_{mes}$. In this case, the covariance matrix of the observation vector $\boldsymbol{y}$ becomes $\mathbf{K}_{obs} := Cov(\boldsymbol{y}) = \mathbf{K} + \mathbf{K}_{mes}$, with $\mathbf{K}$ the covariance matrix of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$.

For estimation of $\boldsymbol{\beta}$, one has the same formulas as (2.11), (2.12), (2.13) and (2.14), by replacing $\mathbf{K}$ by $\mathbf{K}_{obs}$. Similarly, for prediction of $Y(\boldsymbol{x}^{(new)})$, we have the same formulas as (2.9), (2.10), (2.15), (2.16), (2.17) and (2.18), by replacing $\mathbf{K}$ by $\mathbf{K}_{obs}$. This is shown by noting that the proofs of (2.9)-(2.18) only use linear algebra so that they remain the same when $Cov(\boldsymbol{y}) = \mathbf{K} + \mathbf{K}_{mes}$ and $Cov(\boldsymbol{y}, Y(\boldsymbol{x}^{(new)})) = \boldsymbol{r}(\boldsymbol{x}^{(new)})$ (because the measurement errors are

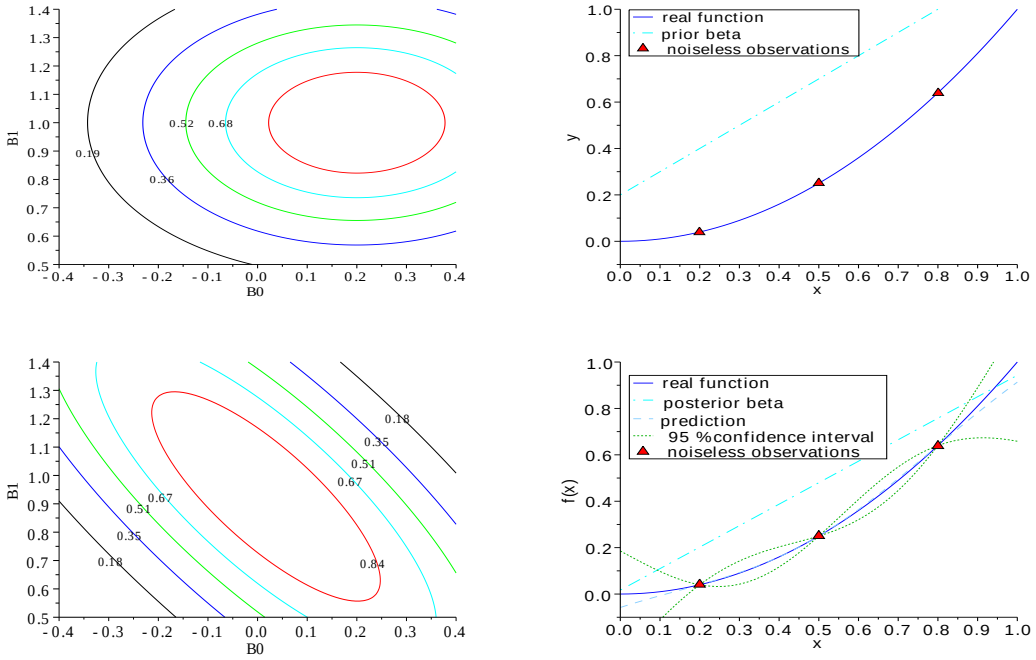

Figure 2.7: Estimation of $\boldsymbol{\beta}$ and prediction in the Bayesian case. Same settings as in figure 2.6, where the a priori distribution of $\boldsymbol{\beta}$ is Gaussian with mean vector $(0.2, 0.1)^t$ and diagonal covariance matrix, with diagonal vector $(0.09, 0.09)^t$. Top left: iso-density curves of the prior probability density function of $\boldsymbol{\beta}$. Bottom left: iso-density curves of the posterior probability density function of $\boldsymbol{\beta}$ given by (2.13) and (2.14). Right: Estimated line with the prior (top) and posterior (bottom, (2.13)) mean values for $\boldsymbol{\beta}$, real parabola, prediction (2.17) and 95% confidence intervals (2.18) of the form $[\hat{y}(\boldsymbol{x}^{(new)}) - 1.96\hat{\sigma}(\boldsymbol{x}^{(new)}), \hat{y}(\boldsymbol{x}^{(new)}) + 1.96\hat{\sigma}(\boldsymbol{x}^{(new)})]$.

uncorrelated with the Gaussian process Y).

Note that, in the case of noisy observations, we can see that $\hat{y}(\boldsymbol{x}^{(i)})$ is not necessarily equal to y_i. Similarly, the predictive variance $\hat{\sigma}^2(\boldsymbol{x}^{(i)})$ is not zero at an observation point $\boldsymbol{x}^{(i)}$. Indeed, the exact value of $Y(\boldsymbol{x}^{(i)})$ remains unknown after the observations, because of the measurement errors.

Even when there is no measurement error, it is common practice to use a matrix $\mathbf{K}_{mes}$ of the form $\lambda \mathbf{I}_n$, with small λ, because the matrix $\mathbf{K} + \lambda \mathbf{I}_n$ is better-conditioned than the matrix $\mathbf{K}$. Although with this practice the Kriging prediction does not interpolate the observations exactly, we call this practice a "numerical nugget effect" in this thesis. This denomination is inspired by [AC12], where the term nugget parameter is used for the parameter λ above. We give more details about the numerical nugget effect in chapter 6 when discussing the practical estimation of the covariance function.

2.2.3 Conditional simulation of Gaussian processes

The formulas of subsection 2.2.2 allow to simulate one-dimensional trajectories $Y(\boldsymbol{x}^{(new)})$ conditionally to the vector $\boldsymbol{y}$ of observed values of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. This one-dimensional simulation might not be sufficient for certain tasks. For example, when modeling monotonic functions ([VM12]), one can not use the point-wise prediction formulas of subsection 2.2.2, because the conditional mean function of (2.9) is not necessarily monotonic, even if the observations are. Hence, we have to simulate trajectories of the Gaussian process, conditionally to the observation vector $\boldsymbol{y}$, and to average only the ones that are monotonic. This is an example of a use of trajectories of a Gaussian process, conditionally to the observation vector $\boldsymbol{y}$. Other classical uses, either conceptual or practical, of these conditional simulations are the multipoint Efficient Global Optimization algorithm (see e.g [CG13a]) and the Stepwise Uncertainty Reduction methods ([BGL+12]).

The simulation of conditional trajectories of Gaussian processes can be obtained from the following proposition, giving the conditional mean and covariance functions of a Gaussian process, according to a vector of observations.

Proposition 2.32. *Let Y be a Gaussian process, observed at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, with observation vector $\boldsymbol{y}$. Let $\mathbf{K}$ be the covariance matrix of $\boldsymbol{y}$ and $\boldsymbol{k}(\boldsymbol{x})$ the covariance vector of Y between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$. Then, according to $\boldsymbol{y}$, the random process Y is Gaussian, with mean function $\boldsymbol{x} \to m(\boldsymbol{x}|\boldsymbol{y})$ and covariance function $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \to K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}|\boldsymbol{y})$.*

In the simple Kriging case m and K are given by

$$m(\boldsymbol{x}|\boldsymbol{y}) = \boldsymbol{k}(\boldsymbol{x})^t \mathbf{K}^{-1} \boldsymbol{y},$$

and

$$K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}|\boldsymbol{y}) = K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) - \boldsymbol{k}(\boldsymbol{x}^{(1)})^t \mathbf{K} \boldsymbol{k}(\boldsymbol{x}^{(2)}).$$

In the ordinary or universal Kriging case, in the frequentist framework, m and K are given by, with $\hat{\boldsymbol{\beta}}$ and $\mathbf{H}$ as in (2.11),

$$m(\boldsymbol{x}|\boldsymbol{y}) = (\boldsymbol{h}(\boldsymbol{x}))^t \hat{\boldsymbol{\beta}} + (\boldsymbol{r}(\boldsymbol{x}))^t \mathbf{K}^{-1} (\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{\beta}}),$$

and

$$\begin{aligned} & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}|\boldsymbol{y}) \\ = \; & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) - \boldsymbol{r}(\boldsymbol{x}^{(1)})^t \mathbf{K}^{-1} \boldsymbol{r}(\boldsymbol{x}^{(2)}) \\ & + (\boldsymbol{h}(\boldsymbol{x}^{(1)}) - \mathbf{H}^t \mathbf{K}^{-1} \boldsymbol{r}(\boldsymbol{x}^{(1)}))^t (\mathbf{H}^t \mathbf{K}^{-1} \mathbf{H})^{-1} (\boldsymbol{h}(\boldsymbol{x}^{(2)}) - \mathbf{H}^t \mathbf{K}^{-1} \boldsymbol{r}(\boldsymbol{x}^{(2)})). \end{aligned}$$

In the ordinary or universal Kriging case, in the Bayesian framework, m and K are given by, with $\boldsymbol{\beta}_{post}$ and $\mathbf{H}$ as in (2.13),

$$m(\boldsymbol{x}|\boldsymbol{y}) = (\boldsymbol{h}(\boldsymbol{x}))^t \boldsymbol{\beta}_{post} + (\boldsymbol{r}(\boldsymbol{x}))^t \mathbf{K}^{-1} (\boldsymbol{y} - \mathbf{H}\boldsymbol{\beta}_{post}),$$

and

$$\begin{aligned} & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}|\boldsymbol{y}) \\ = \; & K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) - \boldsymbol{r}(\boldsymbol{x}^{(1)})^t \mathbf{K}^{-1} \boldsymbol{r}(\boldsymbol{x}^{(2)}) \\ & + (\boldsymbol{h}(\boldsymbol{x}^{(1)}) - \mathbf{H}^t \mathbf{K}^{-1} \boldsymbol{r}(\boldsymbol{x}^{(1)}))^t (\mathbf{H}^t \mathbf{K}^{-1} \mathbf{H} + \mathbf{Q}_{prior}^{-1})^{-1} (\boldsymbol{h}(\boldsymbol{x}^{(2)}) - \mathbf{H}^t \mathbf{K}^{-1} \boldsymbol{r}(\boldsymbol{x}^{(2)})). \end{aligned}$$

Remark 2.33. *In the universal Kriging case in the frequentist framework, $m(\boldsymbol{x}|\boldsymbol{y})$ and $K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}|\boldsymbol{y})$ actually do not constitute the conditional distribution of $Y(\boldsymbol{x})$ given $\boldsymbol{y}$ (notice that this would be the case if we allowed $\boldsymbol{\beta}$ to have an improper and non-informative prior distribution [SWN03], but we do not treat improper prior distributions in this thesis). This conditional distribution depends on the true and unknown regression parameter $\boldsymbol{\beta}$ and is hence uncomputable. The "conditional mean function" we give in proposition 2.32 is actually the Best Linear Unbiased Predictor (BLUP) function and the "conditional covariance function" is the unconditional covariance function of the error process of this BLUP, $\boldsymbol{x} \to m(\boldsymbol{x}|\boldsymbol{y}) - Y(\boldsymbol{x})$. Since these mean and covariance functions are the most reasonable that we can compute, we make the classical slight approximation of naming the distribution they yield the conditional distribution of Y.*

From the mean and covariance functions of proposition 2.32, we are able to simulate conditional trajectories. In figure 2.8, we plot a one-dimensional example of conditional simulations with five exact observation points. We see that all the conditional trajectories pass through the five exact observation points. Furthermore, the conditional simulations have all the most variability when we are away from the observation points.

There are different kinds of methods in the literature for simulating trajectories of Gaussian processes, like the ones in figures 2.3 and 2.8. We refer to [CD99] for an introduction to the subject, and we now present some classical methods. In the rest of subsection 2.2.3, we aim at simulating trajectories of a Gaussian process Y on $\mathcal{D}$. Y has an arbitrary covariance function K (hence including the conditional covariance function of proposition 2.32).

Cholesky decomposition

In our work, we have always used the Cholesky decomposition method. This method aims at simulating Y at n points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$. Consider a Cholesky decomposition $\mathbf{K} = \mathbf{C}\mathbf{C}^t$ of the covariance matrix $\mathbf{K}$ at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, with a vector $\boldsymbol{z}$ following a $\mathcal{N}(0, \mathbf{I}_n)$ distribution (easy to simulate). Let $\boldsymbol{m}$ be the mean vector of $\boldsymbol{y}$. Then the vector $\boldsymbol{y} := \boldsymbol{m} + \mathbf{C}\boldsymbol{z}$ follows a $\mathcal{N}(\boldsymbol{m}, \mathbf{K})$ distribution. The advantage of the Cholesky decomposition is its simplicity, because there are no

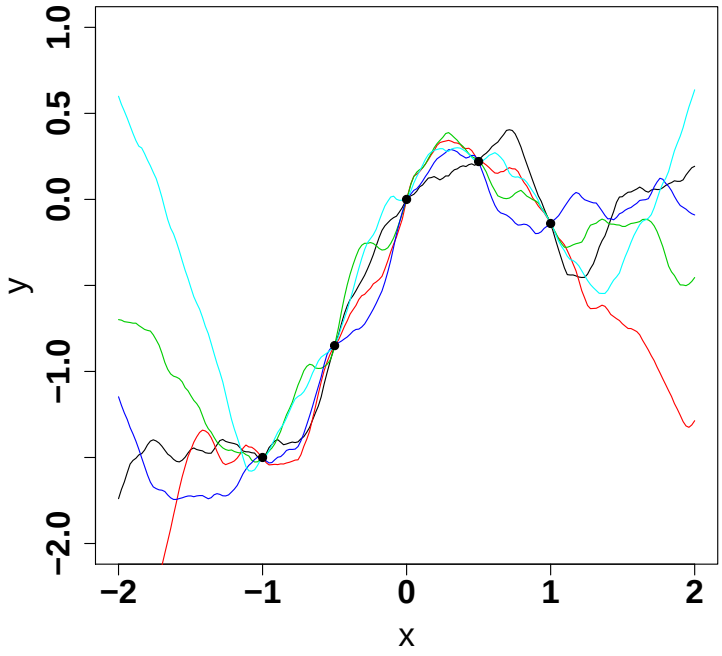

Figure 2.8: Illustration of the conditional simulations of proposition 2.32. A centered Gaussian process, with Matérn $\frac{3}{2}$ covariance function with $\sigma^2 = 1$ and $\ell = 1$, is observed at five observation points (black circles). Conditional simulations are plotted. All the conditional trajectories pass through the five observation points.

conditions on the covariance function K and on the sample points where the Gaussian process is simulated. The drawback of the Cholesky decomposition method is its computational cost: $0(n^3)$ in time for computing the Cholesky decomposition, and $O(n^2)$ for storing the Cholesky decomposition. Note however that, once the Cholesky decomposition is computed and stored, the marginal cost of a simulation is reduced to $O(n^2)$.

Karhunen Loève expansion

Consider a Gaussian process Y on $\mathcal{D} = [0,1]^d$ with stationary covariance function K. The Karhunen Loève expansion method is based on the following Mercer theorem.

Theorem 2.34. *Consider a stationary covariance function K, continuous on $\mathcal{D} \times \mathcal{D}$, with $\mathcal{D} = [0,1]^d$. Then there exists a sequence $(\lambda_i^2)_{i\in\mathbb{N}^*}$ of non-negative scalars, and a sequence $(e_i)_{i\in\mathbb{N}^*}$ of continuous functions $e_i : \mathcal{D} \to \mathbb{R}$ so that $\int_{\mathcal{D}} e_i e_j = \delta_{i,j}$, the $(e_i)_{i\in\mathbb{N}^*}$ form a basis of $L^2(\mathcal{D})$ (the Hilbert space of the square-summable functions on $\mathcal{D}$), and*

$$\int_{\mathcal{D}} K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) e_i(\boldsymbol{x}^{(2)}) d\boldsymbol{x}^{(2)} = \lambda_i^2 e_i(\boldsymbol{x}^{(1)}).$$

Furthermore

$$K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \sum_{i=1}^{+\infty} \lambda_i^2 e_i(\boldsymbol{x}^{(1)}) e_i(\boldsymbol{x}^{(2)}).$$

Proof. See for instance [Aub00]. □

Assume that K is continuous. From theorem 2.34 on K, consider the Gaussian process defined (in the mean square limit sense) by

$$Y(\boldsymbol{x}) = \sum_{i=1}^{+\infty} Z_i \lambda_i e_i(\boldsymbol{x}), \tag{2.19}$$

where the Z_i are *iid* standard Gaussian variables. We calculate

$$\begin{aligned} Cov(Y(\boldsymbol{x}^{(1)}), Y(\boldsymbol{x}^{(2)})) &= Cov(\sum_{i=1}^{+\infty} Z_i \lambda_i e_i(\boldsymbol{x}^{(1)}), \sum_{i=1}^{+\infty} Z_i \lambda_i e_i(\boldsymbol{x}^{(2)})) \\ &= \sum_{i=1}^{+\infty} \lambda_i^2 e_i(\boldsymbol{x}^{(1)}) e_i(\boldsymbol{x}^{(2)}). \\ &= K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}), \end{aligned}$$

so that the Gaussian process in (2.19) does have covariance function K. Hence, (2.19) is called the Karhunen Loève expansion of the Gaussian process Y. (2.19) can be used to simulate Y.

Note firstly that, if the eigenfunctions e_i and the eigenvalues λ_i in (2.19) have explicit expressions, then Y can be simulated efficiently. Indeed, on can truncate (2.19) and if the number of remaining terms is small compared to the number of points n where the Gaussian process is simulated, the computational cost is $O(n)$ operations.

If the eigenfunctions are not explicit, (2.19) is first approximated for N large by

$$Y(\boldsymbol{x}^{(N_1,\ldots,N_d)}) = \sum_{(n_1,\ldots,n_d)\in\{1,\ldots,N\}^d} Z_{n_1,\ldots,n_d} \lambda_{n_1,\ldots,n_d} e_{n_1,\ldots,n_d}(\boldsymbol{x}^{(N_1,\ldots,N_d)}), \tag{2.20}$$

where $\boldsymbol{x}^{(N_1,...,N_d)} = (\frac{N_1}{N}, ..., \frac{N_d}{N})$, the $Z_{n_1,...,n_d}$ are *iid* standard Gaussian variables, the

$$(e_{n_1,...,n_d})_{(n_1,...,n_d)\in\{1,...,N\}^d}$$

and the

$$(\lambda^2_{n_1,...,n_d})_{(n_1,...,n_d)\in\{1,...,N\}^d}$$

are the N^d first eigenfunctions and eigenvalues of the operator K.

Then, let $\mathbf{K}$ be the $N^d \times N^d$ covariance matrix of K at the $\boldsymbol{x}^{(N_1,...,N_d)}$, for $(N_1, ..., N_d) \in \{1, ..., N\}^d$. Then, the vector of the $e_{n_1,...,n_d}(\boldsymbol{x}^{(N_1,...,N_d)})$, for varying $N_1, ..., N_d$, is approximated by the eigenvector of $\mathbf{K}$ corresponding to the index $n_1, ..., n_d$. Similarly $\lambda_{n_1,...,n_d}$ is approximated by the eigenvalue of $\mathbf{K}$ corresponding to the index $n_1, ..., n_d$. These approximations amount to first diagonalizing $\mathbf{K}$ as $\mathbf{K} = \mathbf{PDP}^t$, with $\mathbf{PP}^t = \mathbf{I}_n$ and $\mathbf{D}$ diagonal with diagonal values the $(\lambda^2_{n_1,...,n_d})_{(n_1,...,n_d)\in\{1,...,N\}^d}$. Then we can compute $\boldsymbol{v}_Y$, the vector of Y at the $(\boldsymbol{x}^{(N_1,...,N_d)})_{(N_1,...,N_d)\in\{1,...,N\}^d}$, as

$$\boldsymbol{v}_Y = \mathbf{PD}^{\frac{1}{2}}\boldsymbol{v}_z, \tag{2.21}$$

where $\mathbf{D}^{\frac{1}{2}}$ is the diagonal matrix with diagonal elements the square roots of the diagonal elements of $\mathbf{D}$ and $\boldsymbol{v}_z$ is a vector of *iid* standard Gaussian variables. Because of (2.21), the computational simulation of Y using the Karhunen Loève representation is generally similar to the Cholesky method, because both consist in computing a square-root matrix of the covariance matrix $\mathbf{K}$.

Note finally that, when K can be written as a tensor product $K(\boldsymbol{x}) = K_1(x_1)...K_d(x_d)$, the $N^d \times N^d$ covariance matrix $\mathbf{K}$ can be diagonalized by successively diagonalizing the $N \times N$ matrices $\mathbf{K}_1, ..., \mathbf{K}_d$, where $\mathbf{K}_i$ is the covariance matrix of K_i at the points $\frac{1}{N}, ..., \frac{N}{N}$. Indeed, let $\lambda^i_1, ..., \lambda^i_N$ be the eigenvalues of $\mathbf{K}_i$, with eigenvectors $\boldsymbol{v}^{(i,1)}, ..., \boldsymbol{v}^{(i,N)}$. Then the eigenvalue of $\mathbf{K}$ for the index $n_1, ..., n_d$ is $\lambda^1_{n_1}...\lambda^d_{n_d}$. The component $N_1, ..., N_d$ of the corresponding eigenvector is $v^{(1,n_1)}_{N_1}, ..., v^{(1,n_d)}_{N_d}$. Thus, when K is a tensor product, the computational cost of the Karhunen Loève expansion method goes down from $O((N^d)^3)$ to $O(dN^3 + (N^d)^2)$.

Spectral method

The spectral method aims at simulating Y when the covariance function K is stationary and when $\mathcal{D}$ is a hyper-rectangle of $\mathbb{R}^d$. For this specialized problem, the spectral method, as we will see, is computationally efficient. Indeed, for computing a simulated process at n points, the computational cost is $O(n \ln(n))$.

The method is based on the following spectral representation of the stationary Gaussian process Y (similarly to [Ste99], p23)

$$Y(\boldsymbol{x}) = \int_{\mathbb{R}^d} \cos(\boldsymbol{\omega}^t\boldsymbol{x})M_1(d\boldsymbol{\omega}) + \int_{\mathbb{R}^d} \sin(\boldsymbol{\omega}^t\boldsymbol{x})M_2(d\boldsymbol{\omega}), \tag{2.22}$$

where M_1 and M_2 are random measures verifying, for $k = 1, 2$, for any disjoint Borel sets Δ_1 and Δ_2, $M_k(\Delta_1 \cup \Delta_2) = M_k(\Delta_1) + M_k(\Delta_2)$ and $M_k(\Delta_1)$ is Gaussian with mean 0 and variance $\int_{\Delta_1} \hat{K}(\boldsymbol{\omega})d\boldsymbol{\omega}$. Furthermore, for any Borel sets Δ_3 and Δ_4, $M_1(\Delta_3)$ and $M_2(\Delta_4)$ are independent.

Informally, the computation of the covariance function of $Y(\boldsymbol{x})$ in (2.22) is as follows (see [Ste99], p23).

$$\begin{aligned} Cov\left(Y(\boldsymbol{x}),Y(\boldsymbol{x}')\right) &= Cov\left(\int_{\mathbb{R}^d}\cos\left(\boldsymbol{\omega}^t\boldsymbol{x}\right)M_1(d\boldsymbol{\omega}),\int_{\mathbb{R}^d}\cos\left(\boldsymbol{\omega}^t\boldsymbol{x}'\right)M_1(d\boldsymbol{\omega})\right) \\ &\quad +Cov\left(\int_{\mathbb{R}^d}\sin\left(\boldsymbol{\omega}^t\boldsymbol{x}\right)M_2(d\boldsymbol{\omega}),\int_{\mathbb{R}^d}\sin\left(\boldsymbol{\omega}^t\boldsymbol{x}'\right)M_2(d\boldsymbol{\omega})\right) \\ &= \int_{\mathbb{R}^d}\cos\left(\boldsymbol{\omega}^t\boldsymbol{x}\right)\cos\left(\boldsymbol{\omega}^t\boldsymbol{x}'\right)\hat{K}(\boldsymbol{\omega})d\boldsymbol{\omega} \\ &\quad +\int_{\mathbb{R}^d}\sin\left(\boldsymbol{\omega}^t\boldsymbol{x}\right)\sin\left(\boldsymbol{\omega}^t\boldsymbol{x}'\right)\hat{K}(\boldsymbol{\omega})d\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^d}\cos\left(\boldsymbol{\omega}^t(\boldsymbol{x}-\boldsymbol{x}')\right)\hat{K}(\boldsymbol{\omega})d\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^d}e^{\mathrm{i}\boldsymbol{\omega}^t(\boldsymbol{x}-\boldsymbol{x}')}\hat{K}(\boldsymbol{\omega})d\boldsymbol{\omega} \\ &= K(\boldsymbol{x}-\boldsymbol{x}'), \end{aligned}$$

where the second to last line holds because $\int_{\mathbb{R}^d}\sin\left(\boldsymbol{\omega}^t(\boldsymbol{x}-\boldsymbol{x}')\right)\hat{K}(\boldsymbol{\omega})d\boldsymbol{\omega}$ is the imaginary part of $K(\boldsymbol{x}-\boldsymbol{x}')$ and is 0.

The spectral representation (2.22) is the limit, in distribution, of the discrete representation

$$\begin{aligned} Y_N(\boldsymbol{x}) = & \qquad (2.23) \\ &\sum_{(n_1,\ldots,n_d)\in\{1,\ldots,N\}^d}\left(\frac{2\sqrt{N}}{N}\right)^d\cos\left(\mathrm{i}\left(\boldsymbol{\omega}^{(n_1,\ldots,n_d)}\right)^t\boldsymbol{x}\right)\sqrt{\left(\hat{K}(\boldsymbol{\omega}^{(n_1,\ldots,n_d)})\right)}Z_{n_1,\ldots,n_d} \\ &+\sum_{(n_1,\ldots,n_d)\in\{1,\ldots,N\}^d}\left(\frac{2\sqrt{N}}{N}\right)^d\sin\left(\mathrm{i}\left(\boldsymbol{\omega}^{(n_1,\ldots,n_d)}\right)^t\boldsymbol{x}\right)\sqrt{\left(\hat{K}(\boldsymbol{\omega}^{(n_1,\ldots,n_d)})\right)}Z'_{n_1,\ldots,n_d}, \end{aligned}$$

where $\boldsymbol{\omega}^{(n_1,\ldots,n_d)}=(-\sqrt{N}+2n_1\frac{\sqrt{N}}{N},\ldots,-\sqrt{N}+2n_d\frac{\sqrt{N}}{N})$, and the $Z_{n_1,\ldots,n_d}$, $Z'_{n_1,\ldots,n_d}$ are *iid* standard Gaussian variables. (2.23) is just a random Riemann sum representation of the random integral (2.22). More general forms are possible, but we will consider (2.23) for concision. Spectral methods consist in computing (2.23) for N large. Because the points $\boldsymbol{\omega}^{(n_1,\ldots,n_d)}$ constitute a tensorized grid, (2.23) can be computed for $\boldsymbol{x}$ in a tensorized grid of the same dimension in $O(N^d\ln(N^d))$ using Fast Fourier Transform (FFT) techniques.

Circulant embedding

The circulant embedding method aims at simulating a centered and stationary Gaussian process Y at n points $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)}$ forming a regular grid of $\mathcal{D}=[0,1]^d$. Compared to the spectral method, which aims at simulating stationary processes on dense regular grids, the circulant embedding method is not always feasible (depending on the covariance function K). However, when this method is feasible, it yields simulated trajectories with the exact target distribution, while the distribution of the trajectories obtained from the spectral method is an approximation of the target distribution.

We now present the circulant embedding method in dimension $d=1$, in a way inspired by [Die97]. Let, for $i=1,\ldots,n$, $x_i=\frac{i}{n}$ be the simulation points. The covariance matrix $\mathbf{K}$ of Y

at these points has hence general term $(\mathbf{K})_{i,j} = K(\frac{i-j}{n})$ and is hence a Toeplitz matrix ($(\mathbf{K})_{i,j}$ only depends on $i-j$, see e.g. [Gra01]).

Let $k_i = K(\frac{i}{n})$. Consider now the matrix $\tilde{\mathbf{K}}$, of size $2n-2$ defined so that $\tilde{\mathbf{K}}$ is Toeplitz and symmetric, with first row the vector $\tilde{\boldsymbol{k}} = (k_1, ..., k_{n-1}, k_n, k_{n-1}, ..., k_2)$. For instance, with $n = 4$, $\tilde{\mathbf{K}}$ is as follows

$$\tilde{\mathbf{K}} = \begin{pmatrix} k_1 & k_2 & k_3 & k_4 & k_3 & k_2 \\ k_2 & k_1 & k_2 & k_3 & k_4 & k_3 \\ k_3 & k_2 & k_1 & k_2 & k_3 & k_4 \\ k_4 & k_3 & k_2 & k_1 & k_2 & k_3 \\ k_3 & k_4 & k_3 & k_2 & k_1 & k_2 \\ k_2 & k_3 & k_4 & k_3 & k_2 & k_1 \end{pmatrix}.$$

Notice that the matrix obtained from the n first rows and columns of $\tilde{\mathbf{K}}$ is $\mathbf{K}$.

Now, the applicability of the circulant embedding method depends on whether $\tilde{\mathbf{K}}$ is a non-negative matrix. If it is the case, then, as we will see, it is computationally efficient to simulate $\mathcal{N}(0, \tilde{\mathbf{K}})$ random vectors. By selecting their n first components, we obtain $\mathcal{N}(0, \mathbf{K})$ random vectors.

We thus now show how to simulate $\mathcal{N}(0, \tilde{\mathbf{K}})$ random vectors. The matrix $\tilde{\mathbf{K}}$ is circulant ([Bar90]), so it can be written as ([Bar90]) $\tilde{\mathbf{K}} = \tilde{\mathbf{P}}\tilde{\mathbf{D}}\tilde{\mathbf{P}}^h$, where $\tilde{\mathbf{P}}_{k,l} = e^{\mathrm{i}2\pi\frac{kl}{2n-2}}$, $\tilde{\mathbf{P}}^h_{k,l} = e^{-\mathrm{i}2\pi\frac{kl}{2n-2}}$, where $\mathrm{i}^2 = -1$, and $\tilde{\mathbf{D}}$ is diagonal with diagonal term k equal to

$$\frac{1}{2n-2}\sum_{l=1}^{2n-2} e^{\mathrm{i}2\pi\frac{kl}{2n-2}}\tilde{k}_l. \tag{2.24}$$

Then, $\tilde{\mathbf{K}}$ is non-negative if and only if all the terms (2.24) are non-negative. This means that the applicability of the circulant embedding method can be checked at the cost of only one FFT (to compute the terms (2.24)).

Once $\tilde{\mathbf{D}}$ is computed, the simulation method consists in generating pairs of independent random vectors $\boldsymbol{\epsilon}^{(1)}, \boldsymbol{\epsilon}^{(2)}$, with $\mathcal{N}(0, \mathbf{I}_{2n-2})$ distribution, and in setting

$$\tilde{\boldsymbol{y}} = \tilde{\mathbf{P}}\tilde{\mathbf{D}}^{\frac{1}{2}}\left(\boldsymbol{\epsilon}^{(1)} + \mathrm{i}\boldsymbol{\epsilon}^{(2)}\right).$$

The computation of $\tilde{\boldsymbol{y}}$ is carried out by FFT, with a $O(n\ln(n))$ computation cost. Then, the real and imaginary parts of $\tilde{\boldsymbol{y}}$ have a $\mathcal{N}(0, \tilde{\mathbf{K}})$ distribution, so by extracting their n first components, we obtain vectors with a $\mathcal{N}(0, \mathbf{K})$ distribution.

As a summary, in dimension $d = 1$, the applicability of the circulant embedding method is checked with a $0(n\ln(n))$ computation cost, and, in case of applicability, one simulation is performed with a $0(n\ln(n))$ computation cost. The obtained simulation has exactly the target distribution, contrary to the case of spectral methods.

This principle generalizes in dimension $d > 1$, for which we refer to [Die97]. The computational cost remains $0(n\ln(n))$ for checking the validity and $0(n\ln(n))$ per simulation.

Finally, we also refer to [Die97] for theoretical results ensuring that the terms in (2.24) are non-negative, for particular covariance functions.

Conditioning methods

Consider Y at any n points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. One can always write the joint distribution of $y_1, ..., y_n$ with pdf

$$p(y_1, ..., y_n) = p(y_1)p(y_2|y_1)...p(y_n|y_1, ..., y_{n-1}) \tag{2.25}$$

From (2.25), the random vector $y_1, ..., y_n$ can be sampled by sampling y_1, then sampling y_2 conditionally to y_1, and so on until sampling y_n conditionally to $y_1, ..., y_{n-1}$. These iterative samplings are based on the Kriging equations. It is worth mentioning that the computation of $p(y_{k+1}|y_1, ..., y_k)$ can be efficiently deduced from $p(y_k|y_1, ..., y_{k-1})$ using

$$p(y_{k+1}|y_1, ..., y_k) = p_{|y_1,...,y_{k-1}}(y_{k+1}|y_k),$$

where $p_{|y_1,...,y_{k-1}}(y_{k+1}|y_k)$ is the conditional pdf of y_{k+1} given y_k, when their joint pdf is $p(y_k, y_{k+1}|y_1, ..., y_{k-1})$. This efficient computation is used for Kriging, e.g in [CG13b].

Despite these computationally efficient updates, the conditioning method above is not more efficient than, e.g, a Cholesky decomposition. It can become more efficient if y_k is conditioned only by its nearest neighbors instead of all the previously simulated variables $y_1, ..., y_{k-1}$. Nevertheless, this simplification yields an error in the distribution of the simulated $y_1, ..., y_n$, which needs to be quantified.

A decomposition between unconditional simulation and conditional prediction

We conclude subsection 2.2.3 about conditional simulation by presenting how conditional simulations of a Gaussian process can be obtained from unconditional simulations and conditional predictions. Consider the Gaussian process Y, observed at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ with observation vector $\boldsymbol{y}^0$. Let us define the Gaussian process Z by,

$$Z(\boldsymbol{x}) = \mathbb{E}(Y(\boldsymbol{x})|\boldsymbol{y}^0) + Y(\boldsymbol{x}) - \mathbb{E}(Y(\boldsymbol{x})|\boldsymbol{y}), \tag{2.26}$$

where in (2.26), $\boldsymbol{y} = (Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))^t)$ and Y follows its unconditional distribution. Then, it can be verified ([CD99]) that Z follows the distribution of Y conditionally to $\boldsymbol{y}^0$.

The relation (2.26) is of particular interest when simulating a stationary Gaussian process on a grid, conditionally to a set of observations. Indeed, the unconditional simulation can be carried out efficiently using the spectral method, or the circulant embedding method when it is valid, while the conditional prediction is computed in $O(N)$, where N is the number of points where the conditional prediction is computed.

This eventually gives a $O(N \ln(N))$ method for simulating a Gaussian process at N points, when its distribution is the one of a stationary Gaussian process conditioned by $n \ll N$ observed values.

2.2.4 Cross Validation formulas

In this subsection we start by discussing the Cross Validation principles, and then we give the virtual Cross Validation formulas of [Dub83].

Cross Validation principles

In the general framework of statistical prediction, the quality of a predictor should not be evaluated on the data that helped to build it ([HTF08] chapter 7). This is particularly true for the Gaussian process prediction of (2.9), (2.15) and (2.17), since in the noiseless case it yields an interpolation of the observations. When a rather limited number of observations is available, Cross Validation is a very natural method to assess the predictive capability of a prediction model.

In Kriging models, we are particularly focused on the Leave-One-Out (LOO) technique. For observed vales $\boldsymbol{y} = (y_1, ..., y_n)^t$ of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, LOO is based on the LOO predictions and predictive variances of y_i according to $y_1, ..., y_{i-1}, y_{i+1}, ..., y_n$, for $1 \leq i \leq n$. LOO predictions $\hat{y}_i$ are defined by (2.9), (2.15) and (2.17) for the simple Kriging, frequentist universal Kriging and Bayesian universal Kriging cases. LOO predictive variances $\hat{\sigma}_i^2$ are defined by (2.10), (2.16) and (2.18) for the simple Kriging, frequentist universal Kriging and Bayesian universal Kriging cases.

There are two main uses of the LOO prediction and predictive variance vectors $(\hat{y}_i)_{i=1...n}$ and $(\hat{\sigma}_i^2)_{i=1...n}$. First, we can make a verification of a covariance function at hand, by checking that it gives acceptable predictions and predictive variances. For prediction, the simplest criterion is the LOO Mean Square Error (MSE)

$$MSE_{LOO} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2.27}$$

This criterion should be as small as possible. For predictive variance, a classical criterion is

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}_i^2}. \tag{2.28}$$

It is noted in [Cre93] p.102, that, if the covariance function is correctly specified, then we should expect (2.28) to be close to 1.

In the case when we have a new set of observation points $\boldsymbol{x}^{(n+1)}, ..., \boldsymbol{x}^{(n+p)}$, with observed values $y_{n+1}, ..., y_{n+p}$, other criteria are proposed for the validation of the covariance function ([BO08]). Roughly speaking, these criteria are based on decorrelating the prediction errors $y_{n+i} - \hat{y}_{n+i}$, for $1 \leq i \leq p$. Let us note that we can also decorrelate the LOO errors $y_i - \hat{y}_i$, for $1 \leq i \leq n$. Nevertheless, our opinion is that doing so (for instance by doing a Normality test on the decorrelated LOO errors), is closer in spirit to classical statistical tests on a correlated Gaussian vector than to LOO. Therefore, we rather use the criteria (2.27) and (2.28) when validating a Gaussian process model explicitly by LOO.

The second use of the LOO predictions and predictive variances is for selecting a covariance function, which is the subject of chapter 3.

Virtual Cross Validation formula

The following proposition gives explicit formulas that allow to calculate the $\hat{y}_i, 1 \leq i \leq n$ and the $\hat{\sigma}_i^2, 1 \leq i \leq n$, without solving n different linear systems of size $n - 1$.

Proposition 2.35. *Let Y be a Gaussian process on $\mathcal{D}$, with mean function of the form $\boldsymbol{x} \to \sum_{i=1}^{m} \beta_i g_i(\boldsymbol{x})$, with known functions g_i and unknown coefficients β_i, and with covariance function K. Let $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ be observation points with observation vector $\boldsymbol{y} = (y_1, ..., y_n)$. Let $\mathbf{K}$ be defined by $K_{i,j} = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ and $\mathbf{H}$ be defined by $H_{i,j} = g_j(\boldsymbol{x}^{(i)})$. Then we have*

$$y_i - \hat{y}_i = \frac{1}{(\tilde{\mathbf{K}}^-)_{i,i}} (\tilde{\mathbf{K}}^- \tilde{\boldsymbol{y}})_i$$

and

$$\hat{\sigma}_i^2 = \frac{1}{(\tilde{\mathbf{K}}^-)_{i,i}},$$

with $\tilde{\mathbf{K}}^-$ being $\mathbf{K}^{-1}$ in the simple Kriging case, $\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}^{-1}$ in the frequentist universal Kriging case and $(\mathbf{H}\mathbf{Q}_{prior}\mathbf{H}^t + \mathbf{K})^{-1}$ in the Bayesian universal Kriging case. In the simple Kriging case or the universal Kriging case in the frequentist framework for $\boldsymbol{\beta}$, denote $\tilde{\boldsymbol{y}} = \boldsymbol{y}$, while in the Bayesian framework for $\boldsymbol{\beta}$, denote $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{\beta}_{prior}$.

The formulas leading to proposition 2.35 are classical in the simple Kriging case (see e.g [Rip81], ch.5.2), and were proved by [Dub83] in the universal Kriging case.

The formulas of proposition 2.35 show that we can compute the LOO errors and predictive variances by inverting a unique $n \times n$ matrix. Therefore, computing these errors and predictive variances has the same computational complexity, $O(n^3)$, than calculating the likelihood of the observation vector $\boldsymbol{y}$. This is the basis of the covariance function estimation by LOO, as an alternative to Maximum Likelihood (ML), presented in chapter 3.

Finally, let us note that it is shown in [Dub83] how to generalize the LOO formulas to the case of k-fold cross validation. k-fold Cross Validation is when a given observation is predicted after having removed k observations instead of one. Since we did not study k-fold Cross Validation, we do not give more detail about the corresponding virtual Cross Validation formulas. We refer to [Dub83] on this subject.

2.2.5 Alternative RKHS formulation

There is a parallelism between the simple Kriging conditional mean of (2.9) and Kernel ridge regression ([SS02]). We refer to the PhD thesis [Vaz05] for a detailed analysis of this parallelism.

We will just give a basic interpretation of the link between the conditional mean in case of noisy observations and the expression of the prediction in the Kernel ridge regression framework.

Let us first recall the prediction for Kriging with noisy observations. Assume that the covariance matrix of the measurement error is $\mathbf{K}_{mes} = \sigma_{mes}^2 \mathbf{I}$. Then, the prediction of $Y(\boldsymbol{x}^{(new)})$ given the observation vector $\boldsymbol{y}$ is

$$\hat{y}(\boldsymbol{x}^{(new)}) = \boldsymbol{k}^t (\mathbf{K} + \sigma_{mes}^2 \mathbf{I})^{-1} \boldsymbol{y}, \tag{2.29}$$

with $\mathbf{K}$ the covariance matrix of $\boldsymbol{y}$ and $\boldsymbol{k}$ the covariance vector between $\boldsymbol{y}$ and $Y(\boldsymbol{x}^{(new)})$.

Let us now introduce the Kernel ridge regression framework. This framework is based on considering the Hilbert space $\mathcal{F} \subset \mathbb{R}^{\mathcal{D}}$ defined as the closure of the linear span of the functions $\boldsymbol{x}^{(1)} \to K(\boldsymbol{x}, \boldsymbol{x}^{(1)})$, for $\boldsymbol{x} \in \mathcal{D}$. The closure is defined w.r.t the norm associated to the dot product defined by $\langle K(\boldsymbol{x}^{(1)}, .) | K(\boldsymbol{x}^{(2)}, .) \rangle_{RKHS} = K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$. This dot product is extended

to $\mathcal{F}$ by continuity. This was just a sketch of the mathematical construction of $\mathcal{F}$, and we refer to [SS02] for the detailed mathematical construction.

The Hilbert space $\mathcal{F}$ is defined as a Reproducing Kernel Hilbert Space (RKHS) with reproducing Kernel K because it verifies, for any $f \in \mathcal{F}$ and $\boldsymbol{x} \in \mathcal{D}$

$$\langle f | K(\boldsymbol{x}, .) \rangle_{RKHS} = f(\boldsymbol{x}).$$

Given the RKHS $\mathcal{F}$, we can define the mapping $\mathcal{M} : \mathcal{D} \to \mathcal{F}$, so that $\mathcal{M}(\boldsymbol{x})$ is the function $\boldsymbol{x}^{(1)} \to K(\boldsymbol{x}, \boldsymbol{x}^{(1)})$. This allows to map data from an arbitrary space $\mathcal{D}$ to a Hilbert space $\mathcal{F}$.

Remark 2.36. *Note that it is not necessary that $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be continuous for the Hilbert space $\mathcal{F}$ to be constructed. RKHS methods are indeed classically used with inputs $\boldsymbol{x}$ without continuous structure, such as mathematical representations of character string or of DNA sequences ([STV04]).*

Given the RKHS mathematical framework, kernel ridge regression consists in, from a set of observations $y_1, ..., y_n$ of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, finding $f \in \mathcal{F}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}^{(i)}) - y_i)^2 + \frac{\sigma_{mes}^2}{n} ||f||_{RKHS}^2, \tag{2.30}$$

where $||.||_{RKHS}$ is the norm corresponding to the dot product $\langle .|. \rangle_{RKHS}$, $||f||_{RKHS}$ is interpreted as a measure of the complexity of the function f and $\frac{\sigma_{mes}^2}{n}$ is the regularization parameter. It is shown by the representer theorem ([SS02]) that the function f minimizing (2.30) is of the form $\hat{y}_{\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\hat{y}_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(new)}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i K(\boldsymbol{x}^{(new)}, \boldsymbol{x}^{(i)})$.

As in (2.30) $||\hat{y}_{\boldsymbol{\alpha}}||_{RKHS}^2 = \boldsymbol{\alpha}^t \mathbf{K} \boldsymbol{\alpha}$, one can show, from a straightforward zero-gradient condition, that the solution of the minimization problem (2.30) is given by $\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \sigma_{mes}^2 \mathbf{I})^{-1} \boldsymbol{y}$, thus giving the same prediction as in (2.29). This equivalence makes sense, because when the variance σ_{mes}^2 of the measurement error is large, the observations are unreliable, so a twisted form should not be imposed on the prediction function to reproduce them. Similarly, the weight of the observation reproduction term in (2.30) should be small compared to the complexity penalization term.

Note finally that the virtual Cross Validation formulas are also known in the context of Kernel methods ([Wah90]), and that they are recommended for selecting the regularization parameter σ_{mes}^2 in (2.30).

Chapter 3

Covariance function estimation for Kriging models

In section 3.1, we give an introduction to parametric estimation. We first present the classical properties for an estimator. Then we present the classical asymptotic results for Maximum Likelihood with independent and identically distributed observations. In section 3.2 we present the parametric estimation problem for the covariance function of a Gaussian process. We present and discuss the Maximum Likelihood and Cross Validation estimators. We give the explicit gradients of all the criteria, derived from Maximum Likelihood and Cross Validation, that have to be optimized numerically. We conclude by considering the relatively open problem of taking into account the uncertainty on the covariance function in the Kriging predictions.

3.1 Introduction to parametric estimation

In the whole section 3.1, we consider a vector $\boldsymbol{y}$ of n scalar random observations. In the two following subsections, we first give the basic definitions and properties for the estimation of the parameter characterizing the unknown distribution of $\boldsymbol{y}$. Then we give some classical asymptotic results when the number of observations n goes to $+\infty$.

3.1.1 Definition and properties for parametric estimation

The first notion is the notion of parametric family of distributions for $\boldsymbol{y}$, presented in the following definition.

Definition 3.1. *A parametric family of distributions for $\boldsymbol{y}$ is a parametric family of distributions on $\mathbb{R}^n$, defined by*

$$\mathcal{P} = \{P_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\},$$

where $P_{\boldsymbol{\psi}}$ is a distribution on $\mathbb{R}^n$ and Ψ is a subset of $\mathbb{R}^p$. Unless explicitly stated otherwise, there exists $\boldsymbol{\psi}^{(0)} \in \Psi$ so that the distribution of $\boldsymbol{y}$ is $P_{\boldsymbol{\psi}^{(0)}}$. We will sometimes emphasize this by saying that the model $\mathcal{P}$ is well-specified.

The basic idea of definition 3.1 is that considering directly all the possible distributions as candidates for the generation of $\boldsymbol{y}$ makes no sense. Indeed, we can always consider that $\boldsymbol{y} = (y_1, ..., y_n)$ is generated by a tensor product of Dirac distribution at the y_i, $1 \leq i \leq n$. This distribution will always be the best fit for $\boldsymbol{y}$, but it will make no sense in any applied context. Therefore, we first identify a reasonable set of distributions $\mathcal{P}$, generally using knowledge of the nature of the observations $y_1, ..., y_n$. After that, the corresponding parameter $\boldsymbol{\psi}$ is estimated, generally by using methods that are more automatic, such as the Maximum Likelihood method of definition 3.8.

In the theoretical analysis of parametric estimation, the assumption that the model $\mathcal{P}$ in definition 3.1 is well-specified is very classical. Nevertheless, this assumption is not always done. The term misspecified model has appeared in the literature, in the case where the true distribution of $\boldsymbol{y}$ does not belong to $\mathcal{P}$. In this case, we refer to [Whi82] for asymptotic results for the Maximum Likelihood estimator.

We now give the definition of an estimator, which corresponds to selecting a distribution in $\mathcal{P}$.

Definition 3.2. *An estimator $\hat{\boldsymbol{\psi}}$ is a deterministic function from $\mathbb{R}^n$ to Ψ. $\hat{\boldsymbol{\psi}}(\boldsymbol{y})$ is the estimation of $\boldsymbol{\psi}$, according to the vector of observations $\boldsymbol{y}$.*

Remark 3.3. *We shall write $\hat{\boldsymbol{\psi}}$ for $\hat{\boldsymbol{\psi}}(\boldsymbol{y})$ for concision.*

We see in definition 3.2, that $P_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}$ is the distribution that is concluded to have generated the observation vector $\boldsymbol{y}$. Since the objective is that $P_{\hat{\boldsymbol{\psi}}(\boldsymbol{y})}$ is as close as possible to $P_{\boldsymbol{\psi}^{(0)}}$, the estimated parameter $\hat{\boldsymbol{\psi}}$ should be as close as possible to $\boldsymbol{\psi}^{(0)}$ (in the well-specified case). Since $\hat{\boldsymbol{\psi}}$ is a random vector, its distribution should be concentrated around $\boldsymbol{\psi}^{(0)}$. The following notions of bias and Mean Square Error (MSE) quantify this concentration.

Definition 3.4. *The bias of an estimator $\hat{\boldsymbol{\psi}}$ is the $p \times 1$ vector with ith component $\mathbb{E}(\hat{\boldsymbol{\psi}}_i) - \boldsymbol{\psi}_i^{(0)}$. An estimator is said to be unbiased when $\mathbb{E}(\hat{\boldsymbol{\psi}}_i) = \boldsymbol{\psi}_i^{(0)}$ for $1 \leq i \leq p$.*

Definition 3.5. *The Mean Square Error vector of an estimator $\hat{\boldsymbol{\psi}}$ is the $p \times 1$ vector with ith component $\mathbb{E}\left\{\left(\hat{\boldsymbol{\psi}}_i - \boldsymbol{\psi}_i^{(0)}\right)^2\right\}$.*

The Mean Square Error and the bias of the estimator $\hat{\boldsymbol{\psi}}$ are linked to its variance by the classical identity, for $1 \leq i \leq p$

$$\mathbb{E}\left\{\left(\hat{\boldsymbol{\psi}}_i - \boldsymbol{\psi}_i^{(0)}\right)^2\right\} = \left(\mathbb{E}(\hat{\boldsymbol{\psi}}_i) - \boldsymbol{\psi}_i^{(0)}\right)^2 + Var(\hat{\boldsymbol{\psi}}_i). \tag{3.1}$$

In (3.1), the MSE on the left term is the objective function to minimize for the estimator $\hat{\boldsymbol{\psi}}$. Naturally, for the MSE to be small, both the bias and the variance of the estimator have to be small. Nevertheless, reducing the bias and the variance can be antagonistic, so that a trade-off may have to be found between them, known as the bias-variance trade-off. An example of this trade-off is the introduction of penalization in the Maximum Likelihood estimator for Kriging, that reduces the variance, but at the cost of a small bias ([LS05]).

Finally, when the estimator is unbiased, there is a classical lower bound for its variance, the Cramér Rao bound given in proposition 3.16.

We conclude the subsection by presenting the most classical estimators. The first general family of estimators are the M-estimators, which correspond to minimizing a criterion depending on the observations (which is generally interpreted as a data reproduction criterion).

Definition 3.6. *A M-estimator is an estimator $\hat{\boldsymbol{\psi}}$ so that there exists a deterministic function $c : \Psi \times \mathbb{R}^n \to \mathbb{R}$ so that*

$$\hat{\boldsymbol{\psi}} \in \underset{\boldsymbol{\psi}\in\Psi}{\operatorname{argmin}}\, c(\boldsymbol{\psi}, \boldsymbol{y}).$$

The second family of estimators are the Z estimators, which correspond to verifying a set of equations. An example of Z-estimator is a M-estimator for which it is shown that the minimization of the criterion implies the nullity of its gradient.

Definition 3.7. *A Z-estimator is an estimator $\hat{\boldsymbol{\psi}}$ so that there exists a deterministic function $g : \Psi \times \mathbb{R}^n \to \mathbb{R}^p$ so that*

$$g(\hat{\boldsymbol{\psi}}, \boldsymbol{y}) = 0.$$

Finally, we define the Maximum Likelihood (ML) estimator.

Definition 3.8. *Assume all the distributions $P_{\boldsymbol{\psi}}$ have a probability density function $l_{\boldsymbol{\psi}}$ on $\mathbb{R}^n$. Then the Maximum Likelihood (ML) estimator $\hat{\boldsymbol{\psi}}_{ML}$ of $\boldsymbol{\psi}$ is defined by*

$$\hat{\boldsymbol{\psi}}_{ML} \in \underset{\boldsymbol{\psi}\in\Psi}{\operatorname{argmax}}\, l_{\boldsymbol{\psi}}(\boldsymbol{y}).$$

The ML estimator is both a M-estimator and a Z-estimator, under smoothness condition on the family of probability density functions $\{l_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\}$.

The ML estimator is perhaps the most studied theoretically, and the most used in practice. The main reason is that, as presented in subsection 3.1.2, there is an intrinsic relation between the Cramer-Rao bound and the likelihood function, allowing the ML estimator to reach asymptotically the Cramer-Rao bound.

Before considering, in subsection 3.1.2, the asymptotic framework $n \to +\infty$, let us give a simple example for the Maximum Likelihood estimator.

Example 3.9. *Consider in definition 3.8, $\psi \in \mathbb{R}$ and $l_{\boldsymbol{\psi}}$ is the joint pdf of n iid Gaussian variables with mean ψ and known variance 1. In this case, we can write*

$$l_\psi(y_1, ..., y_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \psi)^2\right). \tag{3.2}$$

Maximizing (3.2) with respect to ψ yields the ML estimator

$$\hat{\psi}_{ML} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right).$$

Hence, in the iid Gaussian case with known variance, the ML estimator estimates the mean parameter ψ by the empirical mean of the observation sample $y_1, ..., y_n$.

One can also calculate $\mathbb{E}(\hat{\psi}_{ML}) = \mathbb{E}(y_1) = \psi^{(0)}$, so that ML estimator is unbiased here. Finally, its variance is $Var(\hat{\psi}_{ML}) = \frac{1}{n}$.

3.1.2 Classical asymptotic results for parametric estimation

To define an asymptotic framework, it is first necessary to let the size n of the observation vector vary. However, we shall keep in mind that the parametric family of distributions of definition 3.1 depends on the number of observations n. It is hence necessary to parameterize all these different distributions on $\mathbb{R}^n$, for n varying, by a parameter $\boldsymbol{\psi}$ independent of n.

In subsection 3.2.1, we will see that the distribution of a random vector $\boldsymbol{y}$, of size n, coming from a centered Gaussian process Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$ can be parameterized independently of n, by a parameter $\boldsymbol{\psi}$ characterizing the covariance function of Y.

In this subsection 3.1.2, we will consider the case where $\boldsymbol{y}$ is composed of n *iid* random variables, so that $\boldsymbol{\psi}$ is a parameter for their common distribution and is hence independent of n.

In view of the discussion above, we will consider the framework of definition 3.10 in this subsection 3.1.2.

Definition 3.10. *Let $\boldsymbol{y}$ be a random vector of size n, with $n \in \mathbb{N}^*$ varying. Assume that the components $y_1, ..., y_n$ of $\boldsymbol{y}$ are iid. A parametric family of iid distributions for $\boldsymbol{y}$, is a parametric family of distributions on $\mathbb{R}$, defined by*

$$\mathcal{P} = \{P_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\},$$

where $P_{\boldsymbol{\psi}}$ is a probability distribution on $\mathbb{R}$ and Ψ is a subset of $\mathbb{R}^p$.

Unless explicitly stated otherwise, there exists $\boldsymbol{\psi}^{(0)} \in \Psi$ so that the common distribution of $y_1, ..., y_n$ is $P_{\boldsymbol{\psi}^{(0)}}$.

An example of parametric family of *iid* distributions is the example 3.9 of the *iid* Gaussian variables, with unknown mean and known variance.

Once the *iid* case is settled, so that the parameter $\boldsymbol{\psi}$ to be estimated is independent of n, the asymptotic framework $n \to +\infty$ has a double objective. The first objective is to answer the question: If there is a very large number of *iid* observations, can we know for sure from which common distribution they stem? The second objective of asymptotic theory is to give an approximation of the distribution of an estimator $\hat{\boldsymbol{\psi}}$, in a given finite-sample situation where n is large.

Consistency

Let us address the question: If there is a very large number of *iid* observations, can we know for sure from which common distribution they stem? This question corresponds to whether the random vector $\hat{\boldsymbol{\psi}}$ goes to $\boldsymbol{\psi}^{(0)}$ when $n \to +\infty$. This corresponds to the notion of consistency of definition 3.11.

Definition 3.11. *Consider the iid framework of definition 3.10. An estimator $\hat{\boldsymbol{\psi}}$ is consistent if $\hat{\boldsymbol{\psi}}$ goes to $\boldsymbol{\psi}^{(0)}$ in probability when $n \to +\infty$. An estimator $\hat{\boldsymbol{\psi}}$ is strongly consistent if $\hat{\boldsymbol{\psi}}$ goes almost surely to $\boldsymbol{\psi}^{(0)}$ when $n \to +\infty$.*

For instance, we have seen, in the example 3.9 of the *iid* Gaussian variables with unknown mean and known variance, that $\mathbb{E}(\hat{\psi}_{ML}) = \psi^{(0)}$ and $Var(\hat{\psi}_{ML}) = \frac{1}{n}$. Hence, this ML estimator

is consistent in the sense of definition 3.11, and even strongly consistent by the strong law of large numbers.

The consistency in the simple example above can be generalized to the case of the ML estimator, in the *iid* framework addressed in this subsection 3.1.2. Indeed, roughly speaking, the *iid* framework is favorable for the estimation of $\boldsymbol{\psi}$, because the information brought by the observations $y_1, ..., y_n$ on $P_{\boldsymbol{\psi}^{(0)}}$ will not be redundant. This is confirmed by the following proposition, showing that, under mild conditions the ML estimator is consistent in the *iid* framework.

Proposition 3.12. *Consider the iid framework of definition 3.10, where $P_{\boldsymbol{\psi}}$ is a distribution on $\mathbb{R}$ having a probability density function $l_{\boldsymbol{\psi}}$ with respect to the Lebesgue measure. Assume that all parameters $\boldsymbol{\psi}$ give distinct distributions $P_{\boldsymbol{\psi}}$. Assume that Ψ is compact. Assume that $\boldsymbol{\psi} \to \ln(l_{\boldsymbol{\psi}}(y))$ is continuously differentiable for any $y \in \mathbb{R}$ and that $\sup_{\boldsymbol{\psi}\in\Psi} |\ln(l_{\boldsymbol{\psi}}(y))|$ and $\sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial}{\partial \boldsymbol{\psi}} \ln(l_{\boldsymbol{\psi}}(y))|$ are summable (with respect to the distribution of y on $\mathbb{R}$ given by $P_{\boldsymbol{\psi}^{(0)}}$). Then the ML estimator $\hat{\boldsymbol{\psi}}_{ML}$ is consistent.*

Proof. For any n,

$$\hat{\boldsymbol{\psi}}_{ML} \in \operatorname*{argmin}_{\boldsymbol{\psi}\in\Psi} \frac{1}{n} \sum_{i=1}^{n} \ln(l_{\boldsymbol{\psi}}(y_i))$$

Denoting $M_n(\boldsymbol{\psi}) = \frac{1}{n}\sum_{i=1}^{n} \ln(l_{\boldsymbol{\psi}}(y_i))$ and using the strong law of large numbers, $M_n(\boldsymbol{\psi})$ goes in probability, for each $\boldsymbol{\psi}$ to

$$M(\boldsymbol{\psi}) := \int_{\mathbb{R}} \ln(l_{\boldsymbol{\psi}}(\boldsymbol{z})) l_{\boldsymbol{\psi}^{(0)}}(z) dz.$$

Furthermore, let $t > 0$ and, for $\epsilon > 0$, let $\boldsymbol{\psi}^{(1)}, ..., \boldsymbol{\psi}^{(N)}$ so that $\sup_{\boldsymbol{\psi}\in\Psi} \inf_{1\le i\le N} |\boldsymbol{\psi} - \boldsymbol{\psi}^{(i)}| \le \epsilon$. Then, with a constant $a < +\infty$, for all $\boldsymbol{\psi} \in \Psi$

$$\begin{aligned}
& |M_n(\boldsymbol{\psi}) - M(\boldsymbol{\psi})| \\
\le\ & \min_{1\le i\le N} \left(|M_n(\boldsymbol{\psi}) - M_n(\boldsymbol{\psi}^{(i)})| + |M_n(\boldsymbol{\psi}^{(i)}) - M(\boldsymbol{\psi}^{(i)})| + |M(\boldsymbol{\psi}^{(i)}) - M(\boldsymbol{\psi})| \right) \\
\le\ & \max_{1\le i\le N} |M_n(\boldsymbol{\psi}^{(i)}) - M(\boldsymbol{\psi}^{(i)})| + a\epsilon \frac{1}{n} \sup_{\boldsymbol{\psi}\in\Psi} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\psi}} \ln(l_{\boldsymbol{\psi}}(y_i)) \right| \\
& + a\epsilon \sup_{\boldsymbol{\psi}\in\Psi} \left| \int_{\mathbb{R}} \frac{\partial}{\partial \boldsymbol{\psi}} \ln(l_{\boldsymbol{\psi}}(z)) l_{\boldsymbol{\psi}^{(0)}}(z) dz \right|.
\end{aligned}$$

Hence

$$\begin{aligned}
& \sup_{\boldsymbol{\psi}\in\Psi} |M_n(\boldsymbol{\psi}) - M(\boldsymbol{\psi})| \\
\le\ & \max_{1\le i\le N} |M_n(\boldsymbol{\psi}^{(i)}) - M(\boldsymbol{\psi}^{(i)})| \\
& + a\epsilon \frac{1}{n} \sum_{i=1}^{n} \sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial}{\partial \boldsymbol{\psi}} \ln(l_{\boldsymbol{\psi}}(y_i))| + a\epsilon \int_{\mathbb{R}} \sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial}{\partial \boldsymbol{\psi}} \ln(l_{\boldsymbol{\psi}}(z))| l_{\boldsymbol{\psi}^{(0)}}(z) dz \\
=\ & o_p(1) + a\epsilon \left(2 \int_{\mathbb{R}} \sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial}{\partial \boldsymbol{\psi}} \ln(l_{\boldsymbol{\psi}}(z))| l_{\boldsymbol{\psi}^{(0)}}(z) dz + o_p(1) \right) \\
=\ & o_p(1) + \epsilon K,
\end{aligned}$$

for a constant $K < +\infty$. Hence, for any n larger than N large enough,

$$P\left(\sup_{\boldsymbol{\psi}\in\Psi} |M_n(\boldsymbol{\psi}) - M(\boldsymbol{\psi})| \geq t\right) \leq \epsilon + \mathbf{1}_{\epsilon K \geq \frac{t}{2}}.$$

Hence, $\sup_{\boldsymbol{\psi}\in\Psi} |M_n(\boldsymbol{\psi}) - M(\boldsymbol{\psi})|$ goes to zero in probability.

Now the function $M(\boldsymbol{\psi})$ is continuous by the dominated convergence theorem. It is proved in theorem 5.35 of [Van98] that $M(\boldsymbol{\psi}^{(0)}) > M(\boldsymbol{\psi})$ for any $\boldsymbol{\psi} \neq \boldsymbol{\psi}^{(0)}$. Because Ψ is compact, we have then, for any $\alpha > 0$

$$\sup_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq\alpha} M(\boldsymbol{\psi}) < M(\boldsymbol{\psi}^{(0)}).$$

Hence, because of theorem 5.7 of [Van98], $\hat{\boldsymbol{\psi}}_{ML}$ is consistent. □

Let us consider again the example 3.9 of the *iid* Gaussian observations. By taking two times the opposite of the logarithm of their likelihood, we can write the ML estimator of the mean ψ as

$$\hat{\psi}_{ML} \in \operatorname*{argmin}_{\psi\in\mathbb{R}} L(\psi) \quad \text{with} \quad L(\psi) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \psi)^2. \tag{3.3}$$

Although the framework here is so that $L(\psi)$ can be minimized explicitly, it is worth noting that it is composed of a sum of *iid* terms. Each term, in the mean sense, is minimized only by the true ψ. Hence, it is intuitive that, using the law of large numbers, when the number of observations is large, the modified likelihood function $\psi \to L(\psi)$ is close to the mean likelihood function $\psi \to \mathbb{E}((Y-\psi)^2) = 1 + (\psi_0 - \psi)^2$, where Y follows the $\mathcal{N}(\psi_0, 1^2)$ distribution. Since this mean likelihood function is minimized only by the true ψ, the ML estimator $\hat{\psi}_{ML}$ is close to the true ψ when the number of observation is large. This discussion hence explains why proposition 3.12 is intuitive.

We now illustrate the example in figure 3.1. We set $\psi_0 = 1$ as the true mean. We plot realizations of $\psi \to L(\psi)$ in (3.3) for $n = 5$ and $n = 30$ observation points, and the mean likelihood function $\psi \to 1+(\psi_0-\psi)^2$. We see that for $n = 30$, the likelihood function realizations are much closer to the mean likelihood function than for $n = 5$, and that the mean likelihood function is minimized only by the true mean ψ_0.

Asymptotic distribution

The second objective of asymptotic theory is to give approximation of the distribution of an estimator $\hat{\boldsymbol{\psi}}$, in a given finite-sample situation. Indeed, for instance if $\hat{\boldsymbol{\psi}}$ is a M-estimator, it may not have an explicit expression, so that its finite-sample distribution may not be explicit.

Generally, but not always, the estimator will be proved consistent first, so that the question is to quantify its small deviation from the true parameter $\boldsymbol{\psi}^{(0)}$. In the case of M and Z-estimators, since the deviations are small, the asymptotic distribution can be deduced from the derivatives of the criterion, with respect to $\boldsymbol{\psi}$, around the true parameter $\boldsymbol{\psi}^{(0)}$. A result of this kind is presented in more details in chapter 5 for the ML and CV M-estimators in the Kriging case.

Here, we will give details about this principle in the case of the Maximum Likelihood estimator in the *iid* case. For this, let us first define the first and second derivatives of the logarithm of the likelihood.

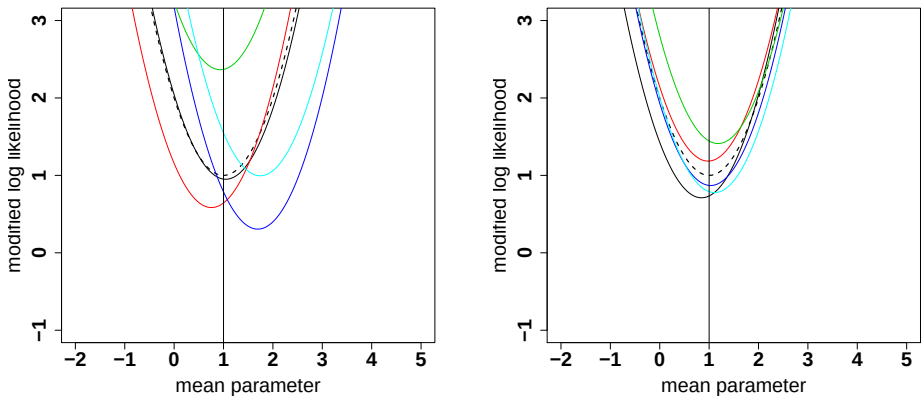

Figure 3.1: Illustration of convergence of the likelihood function in the *iid* case. Solid lines: plot of realizations of the modified log-likelihood function $\psi \to L(\psi)$ in (3.3) for *iid* Gaussian variables with known variance 1 and unknown mean. The true mean is $\psi_0 = 1$. Dashed lines: plot of the mean likelihood function $\psi \to \mathbb{E}((Y-\psi)^2) = 1 + (\psi_0 - \psi)^2$, where Y follows the $\mathcal{N}(\psi_0, 1^2)$ distribution. Left: $n = 5$ observation points. Right: $n = 30$ observation points.

In the two following definitions, and in the two following propositions, since n is fixed, we do not consider necessarily the *iid* case of definition 3.10.

Definition 3.13. *Assume all the distributions P_ψ in definition 3.1 have a probability density function l_ψ on $\mathbb{R}^n$. Then, the score vector is the $p \times 1$ random vector defined by*

$$\left(\frac{\partial}{\partial \boldsymbol{\psi}} \ln\left(l_{\boldsymbol{\psi}}\right)\right)_{\boldsymbol{\psi}^{(0)}}.$$

Definition 3.14. *Assume all the distributions P_ψ in definition 3.1 have a probability density function l_ψ on $\mathbb{R}^n$. Then, the random Fisher information matrix is the $p \times p$ random matrix defined by*

$$-\left(\frac{\partial^2}{\partial \boldsymbol{\psi}^2} \ln\left(l_{\boldsymbol{\psi}}\right)\right)_{\boldsymbol{\psi}^{(0)}}.$$

The moments of the score and of the random Fisher information define the (deterministic) Fisher information matrix, as presented in the following proposition.

Proposition 3.15. *Assume that $\sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial}{\partial \boldsymbol{\psi}}(l_{\boldsymbol{\psi}})|$ and $\sup_{\boldsymbol{\psi}\in\Psi} ||\frac{\partial^2}{\partial \boldsymbol{\psi}^2}(l_{\boldsymbol{\psi}})||_2$ are summable with respect to the Lebesque measure on $\mathbb{R}^n$. Then the $p\times p$ covariance matrix of the score of definition 3.13 and the $p \times p$ mean value matrix of the random Fisher information of definition 3.14 are equal. Their common value is denoted $\mathcal{I}_n$ and is called the (deterministic) Fisher information matrix.*

Proof. Let $1 \leq i, j \leq p$. Because of the dominated convergence theorem,

$$\begin{aligned}\mathbb{E}\left(\frac{\partial}{\partial\psi_i}\ln\left(l_{\boldsymbol{\psi}^{(0)}}\right)\right) &= \int_{\mathbb{R}^n}\frac{\frac{\partial}{\partial\psi_i}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))}{l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})}l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})\boldsymbol{z}\\ &= \frac{\partial}{\partial\psi_i}\int_{\mathbb{R}^n}l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})d\boldsymbol{z}\\ &= \frac{\partial}{\partial\psi_i}1\\ &= 0.\end{aligned}$$

We calculate

$$\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}\right) = \frac{\frac{\partial^2}{\partial\psi_i\partial\psi_j}(l_{\boldsymbol{\psi}^{(0)}})}{l_{\boldsymbol{\psi}^{(0)}}} - \frac{\frac{\partial}{\partial\psi_i}(l_{\boldsymbol{\psi}^{(0)}})\frac{\partial}{\partial\psi_j}(l_{\boldsymbol{\psi}^{(0)}})}{(l_{\boldsymbol{\psi}^{(0)}})^2}.$$

Integrating this relation, we obtain

$$\begin{aligned}\int_{\mathbb{R}^n}\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})\right)l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})d\boldsymbol{z} &= \int_{\mathbb{R}^n}\frac{\frac{\partial^2}{\partial\psi_i\partial\psi_j}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))}{l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})}l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})d\boldsymbol{z}\\ &\quad - \int_{\mathbb{R}^n}\frac{\frac{\partial}{\partial\psi_i}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))\frac{\partial}{\partial\psi_j}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))}{(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))^2}l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z})d\boldsymbol{z}.\end{aligned}$$

so that

$$\mathbb{E}\left\{\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}\right)\right\} = \int_{\mathbb{R}^n}\frac{\partial^2}{\partial\psi_i\partial\psi_j}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))d\boldsymbol{z} - \mathbb{E}\left\{\frac{\partial}{\partial\psi_i}\ln\left(l_{\boldsymbol{\psi}^{(0)}}\right)\frac{\partial}{\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}\right)\right\}.$$

Because of the dominated convergence theorem,

$$\int_{\mathbb{R}^n}\frac{\partial^2}{\partial\psi_i\partial\psi_j}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{z}))d\boldsymbol{z} = \frac{\partial^2}{\partial\psi_i\partial\psi_j}1 = 0,$$

which concludes the proof. □

From proposition 3.15, the Fisher information matrix $\mathcal{I}_n$ is a $p \times p$ covariance matrix. It is therefore a symmetric non-negative matrix.

Furthermore it can be written with both the expressions

$$(\mathcal{I}_n)_{i,j} = \int_{\mathbb{R}^n}\frac{\partial\ln\left(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{y})\right)}{\partial\psi_i}\frac{\partial\ln\left(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{y})\right)}{\partial\psi_j}l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{y})$$

and

$$(\mathcal{I}_n)_{i,j} = -\int_{\mathbb{R}^n}\frac{\partial^2\ln\left(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{y})\right)}{\partial^2\psi_i\psi_j}l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{y}).$$

The deterministic Fisher information matrix defines a lower-bound for the mean square error of all the unbiased estimators of $\boldsymbol{\psi}$. This inequality is called the Cramér Rao inequality, and is presented in the following proposition.

Proposition 3.16. *Let $\hat{\boldsymbol{\psi}}$ be an unbiased estimator of $\boldsymbol{\psi}$. Assume that for any j, $\boldsymbol{v} \to \sup_{\boldsymbol{\psi}\in\Psi}|\hat{\boldsymbol{\psi}}(\boldsymbol{v})\frac{\partial}{\partial\psi_j}(l_{\boldsymbol{\psi}}(\boldsymbol{v}))|$ is summable with respect to the Lebesque measure on $\mathbb{R}^n$. Let $\mathcal{I}_n$ be the deterministic Fisher information matrix of proposition 3.15, and assume that the matrix is positive. Then, for any $\boldsymbol{\alpha} \in \mathbb{R}^p$, we have the following Cramer-Rao inequality*

$$\mathbb{E}(|\alpha^t(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{(0)})|^2) \geq \boldsymbol{\alpha}^t(\mathcal{I}_n^{-1})\boldsymbol{\alpha}.$$

Proof. For $1 \leq i, j \leq p$,

$$\mathbb{E}(\hat{\psi}_i \frac{\partial}{\partial \psi_j} \ln(l_{\boldsymbol{\psi}^{(0)}})) = \int_{\mathbb{R}^n} \hat{\psi}_i(\boldsymbol{v}) \frac{\partial}{\partial \psi_j}(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}))d\boldsymbol{v}.$$

Because $\hat{\psi}_i(\boldsymbol{v})$ does not depend on $\boldsymbol{\psi}$, by using the dominated convergence theorem,

$$\begin{aligned}\mathbb{E}(\hat{\psi}_i \frac{\partial}{\partial \psi_j} \ln(l_{\boldsymbol{\psi}^{(0)}})) &= \frac{\partial}{\partial \psi_j} \int_{\mathbb{R}^n} \hat{\psi}_i(\boldsymbol{v})(l_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}))d\boldsymbol{v} \\ &= \frac{\partial}{\partial \psi_j}\psi_{0,i} \\ &= \delta_{i,j}.\end{aligned}$$

Hence, the covariance matrix of the size $p+1$ random vector

$$\begin{pmatrix} \sum_{i=1}^p \alpha_i \hat{\psi}_i \\ \frac{\partial}{\partial \psi_1} \ln(l_{\boldsymbol{\psi}^{(0)}})) \\ \vdots \\ \frac{\partial}{\partial \psi_p} \ln(l_{\boldsymbol{\psi}^{(0)}})) \end{pmatrix}$$

is the matrix

$$\begin{pmatrix} Var(\sum_{i=1}^p \alpha_i \hat{\psi}_i) & \boldsymbol{\alpha}^t \\ \boldsymbol{\alpha} & \mathcal{I}_n \end{pmatrix}.$$

Since this matrix is non-negative, for any $t \in \mathbb{R}$, considering the vector

$$\begin{pmatrix} t \\ -\mathcal{I}_n^{-1}\boldsymbol{\alpha} \end{pmatrix},$$

we have

$$0 \leq t^2 Var(\sum_{i=1}^p \alpha_i \hat{\psi}_i) - 2t\boldsymbol{\alpha}^t \mathcal{I}_n^{-1}\boldsymbol{\alpha} + \boldsymbol{\alpha}^t \mathcal{I}_n^{-1}\mathcal{I}_n\mathcal{I}_n^{-1}\boldsymbol{\alpha}.$$

Since this last term is non-negative for every t we obtain

$$Var\left(\sum_{i=1}^p \alpha_i \hat{\psi}_i\right) \boldsymbol{\alpha}^t \mathcal{I}_n^{-1}\boldsymbol{\alpha} \geq (\boldsymbol{\alpha}^t \mathcal{I}_n^{-1}\boldsymbol{\alpha})^2,$$

which proves the proposition. □

Let us consider the example 3.9 of the *iid* Gaussian variables with unknown mean and known variance. The log-likelihood at ψ is

$$-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\left(\sum_{i=1}^n (y_i - \psi)^2\right)$$

Differentiating two times with respect to ψ, we obtain that the random Fisher information matrix is in fact deterministic and is equal to the scalar n. Therefore, as we have seen that $\mathbb{E}(\hat{\psi}_{ML}) = \psi^{(0)}$ and $Var(\hat{\psi}_{ML}) = \frac{1}{n}$, we have shown that the ML estimator reaches the Cramér-Rao bound.

This is in fact a general result, as the following proposition shows: in the *iid* framework, the ML estimator is asymptotically normal, with mean vector zero and covariance matrix equal to the Cramér-Rao bound.

Proposition 3.17. *Consider the iid framework of definition 3.10, where $P_{\boldsymbol{\psi}}$ is a distribution on $\mathbb{R}$ having a probability density function $l_{\boldsymbol{\psi}}$ with respect to the Lebesgue measure. Assume that all parameters $\boldsymbol{\psi}$ give distinct distributions $P_{\boldsymbol{\psi}}$. Assume that Ψ is compact. Assume that, for any j,k,l, $\sup_{\boldsymbol{\psi}\in\Psi} |\ln(l_{\boldsymbol{\psi}}(y))|$, $\sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial}{\partial\psi_j}\ln(l_{\boldsymbol{\psi}}(y))|$, $\sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial^2}{\partial\psi_j\psi_k}\ln(l_{\boldsymbol{\psi}}(y))|$ and $\sup_{\boldsymbol{\psi}\in\Psi} |\frac{\partial^3}{\partial\psi_j\psi_k\psi_l}\ln(l_{\boldsymbol{\psi}}(y))|$ are summable with respect to the true probability density function $l_{\boldsymbol{\psi}^{(0)}}$ for y.*

Then the (deterministic) Fisher information matrix $\mathcal{I}_n$ is

$$n\mathbb{E}\left\{-\left(\frac{\partial^2}{\partial\boldsymbol{\psi}^2}\ln(l_{\boldsymbol{\psi}})\right)_{\boldsymbol{\psi}^{(0)}}\right\} := n\mathcal{I}_1.$$

Assuming that the matrix $\mathcal{I}_1$ is positive, the Maximum Likelihood estimator $\hat{\Psi}_{ML}$ verifies as $n\to+\infty$

$$\sqrt{n}(\hat{\boldsymbol{\psi}}_{ML}-\boldsymbol{\psi}^{(0)}) \to_{\mathcal{L}} \mathcal{N}(0,\mathcal{I}_1^{-1}).$$

Proof. As the observations are *iid*, the equality $\mathcal{I}_n = n\mathcal{I}_1$ is obtained by writing $\ln(\prod_{i=1}^n l_{\boldsymbol{\psi}}(y_i))$ as a sum, deriving it twice with respect to ψ_i and ψ_j and taking the mean value.

We verify the hypothesis of proposition 3.12 so the ML estimator is consistent.

Let us now address asymptotic normality. For all $1\le j\le p$,

$$\begin{aligned} 0 &= \sum_{i=1}^n \frac{\partial}{\partial\psi_j}\ln\left(l_{\hat{\boldsymbol{\psi}}_{ML}}(y_i)\right) \\ &= \sum_{i=1}^n \frac{\partial}{\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right) + \left(\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\psi}}\frac{\partial}{\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right)\right)^t \left(\hat{\boldsymbol{\psi}}_{ML}-\boldsymbol{\psi}^{(0)}\right) + r, \end{aligned}$$

with random r, so that $|r| \le \sum_{i=1}^n \sup_{\tilde{\boldsymbol{\psi}},j,k,l} \left|\frac{\partial^3}{\partial\psi_j\psi_k\psi_l}\ln\left(l_{\tilde{\boldsymbol{\psi}}}(y_i)\right)\right| \times |\hat{\boldsymbol{\psi}}_{ML}-\boldsymbol{\psi}^{(0)}|^2$. Hence $r = o_p\left(|\hat{\boldsymbol{\psi}}_{ML}-\boldsymbol{\psi}^{(0)}|\right)$. We then have

$$-\sum_{i=1}^n \frac{\partial}{\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right) = \left[\left(\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\psi}}\frac{\partial}{\partial\psi_j}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right)\right)^t + o_p(1)\right]\left(\hat{\boldsymbol{\psi}}_{ML}-\boldsymbol{\psi}^{(0)}\right),$$

and so

$$\left(\hat{\boldsymbol{\psi}}_{ML}-\boldsymbol{\psi}^{(0)}\right) = -\left\{\sum_{i=1}^n \frac{\partial^2}{\partial\boldsymbol{\psi}^2}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right) + o_p(1)\right\}^{-1}\left(\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\psi}}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right)\right).$$

Now, thanks to the strong law of large numbers, the central limit theorem and proposition 3.15, $\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\boldsymbol{\psi}^2}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right)$ converges in probability to $\mathcal{I}_1$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\psi}}\ln\left(l_{\boldsymbol{\psi}^{(0)}}(y_i)\right)$ converges in distribution to a $\mathcal{N}(0,\mathcal{I}_1)$ distribution. We conclude using Slutsky lemma. □

Proposition 3.17 is a justification of the use of the ML estimator, by showing that it is asymptotically unbiased, and that its asymptotic covariance matrix is equal to the Cramér-Rao lower-bound (in the convergence in distribution sense).

Looking back at the example 3.9 of the *iid* Gaussian variables with unknown mean and known variance, we see that $\hat{\psi}_{ML}-\psi^{(0)} \sim \mathcal{N}(0,\frac{1}{n})$. Hence, in this simple example, the asymptotic distribution of proposition 3.17 is in fact the exact distribution for any n.

The results of propositions 3.12 and 3.17 can not be used directly, in the Kriging framework, for the Maximum Likelihood estimator of section 3.2, because, as we will see, we are not in the *iid* framework. The existence, or the impossibility, of these kinds of asymptotic results for Kriging is the object of chapters 4 and 5. Before that, in the next section 3.2, we present the finite sample framework for the estimation of the covariance function for Kriging.

3.2 Estimation of the covariance function for Gaussian processes

In this section, we present the parametric estimation of the covariance function of a Gaussian process Y, from an observation vector $\boldsymbol{y}$. In subsection 3.2.1, we detail the framework for the covariance function estimation. In subsections 3.2.2 and 3.2.3, we present the Maximum Likelihood (ML) and Cross Validation (CV) estimators. In subsection 3.2.4, we provide the explicit gradients of the criteria for the ML and CV estimators. Finally, in subsection 3.2.5, we discuss the rather open problem of taking the covariance function estimation error into account in the Kriging predictions.

3.2.1 Parametric estimation of the covariance function

As discussed in section 3.1, it is unreasonable to consider all possible covariance functions as possible candidates for the Gaussian process at hand. Hence, similarly to definition 3.1, it is classical to assume a parametric family for the covariance function of a Gaussian process Y. Furthermore, in the present manuscript, we especially study the classical case of a family of stationary covariance functions. These two remarks motivate the following definition of a parametric family of stationary covariance functions.

Definition 3.18. *A parametric family of stationary covariance functions is of the form*

$$\{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\},$$

where $K_{\boldsymbol{\psi}}$ is a stationary covariance function, and Ψ is a subset of $\mathbb{R}^p$.

In definition 3.18, since $K_{\boldsymbol{\psi}}$ is a stationary covariance function for all $\boldsymbol{\psi}$, we have $K_{\boldsymbol{\psi}}(\boldsymbol{x}) \leq K_{\boldsymbol{\psi}}(0)$. We make the reasonable hypothesis that we have the strict inequality $K_{\boldsymbol{\psi}}(\boldsymbol{x}) < K_{\boldsymbol{\psi}}(0)$ for $\boldsymbol{x} \neq 0$. Without this hypothesis, for Y a centered Gaussian process with covariance function $K_{\boldsymbol{\psi}}$, we can have $\boldsymbol{x}^{(1)} \neq \boldsymbol{x}^{(2)}$ so that $Y(\boldsymbol{x}^{(1)}) = Y(\boldsymbol{x}^{(2)})$ almost-surely, which only holds in very particular situations.

Since the variance $K_{\boldsymbol{\psi}}(0)$ of the stationary Gaussian process is constant, it usually makes sense to consider it as an explicit parameter. Therefore, we shall consider the alternative parameterization of $K_{\boldsymbol{\psi}}$ in definition 3.18,

$$\{\sigma^2 R_{\boldsymbol{\theta}}, \sigma^2 > 0, \boldsymbol{\theta} \in \Theta\}, \tag{3.4}$$

where $R_{\boldsymbol{\theta}}$ is a correlation function and Θ is a subset of $\mathbb{R}^{p-1}$. The explicit separation of the variance hyper-parameter σ^2 and the correlation hyper-parameter $\boldsymbol{\theta}$ in (3.4) turns out to be useful when we address their estimation in subsections 3.2.2 and 3.2.3.

Similarly to definition 3.18, we assume $R_{\boldsymbol{\theta}}(\boldsymbol{x}) < R_{\boldsymbol{\theta}}(0)$ for $\boldsymbol{x} \neq 0$.

Remark 3.19. *In the manuscript, when the separation of the variance and correlation hyper-parameters is explicitly used, we will consider the parameterization* (3.4)*. When this separation is not used, we will rather consider the parameterization of definition 3.18.*

3.2.2 Maximum Likelihood for estimation

In all the subsection, $\boldsymbol{y}$ is the vector of observations of the Gaussian process Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. $\mathbf{K}_{\boldsymbol{\psi}}$ is the covariance matrix of $\boldsymbol{y}$ under covariance function $K_{\boldsymbol{\psi}}$ and $\mathbf{R}_{\boldsymbol{\theta}}$ is the correlation matrix of $\boldsymbol{y}$ under correlation function $R_{\boldsymbol{\theta}}$.

$\mathbf{K}_{\boldsymbol{\psi}}$ and $\mathbf{R}_{\boldsymbol{\theta}}$ are defined by $(\mathbf{K}_{\boldsymbol{\psi}})_{i,j} = K_{\boldsymbol{\psi}}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})$ and $(\mathbf{R}_{\boldsymbol{\theta}})_{i,j} = R_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})$. We assume that, when the points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ are distinct, the matrices $\mathbf{K}_{\boldsymbol{\psi}}$ and $\mathbf{R}_{\boldsymbol{\theta}}$ are invertible. This assumption is classical. For example, it is verified by all the covariance functions of subsection 2.1.2.

Maximum Likelihood

In the case of simple Kriging, the likelihood criterion of the observation vector $\boldsymbol{y}$ depends only on $\boldsymbol{\psi}$ in 3.18 and is,

$$L(\boldsymbol{\psi}) := \frac{1}{n}\left[\ln|\mathbf{K}_{\boldsymbol{\psi}}| + \boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y}\right] \tag{3.5}$$

Remark 3.20. *The criterion in 3.5 is not the likelihood, but it is a monotone transformation of it (it is* $-\frac{2}{n}\ln l(\boldsymbol{\psi}) - \ln(2\pi)$*, where* $l(\boldsymbol{\psi})$ *is the likelihood). Hence, the Maximum Likelihood estimator is of course preserved. The criterion in* (3.5) *gives the simplest expressions for the theoretical and practical development regarding Maximum Likelihood. Notice that, in* (3.5)*, we have changed the sign of the logarithm of the likelihood, so that the criterion* (3.5) *is to be minimized.*

In the case of simple Kriging, the Maximum Likelihood estimator of the covariance hyper-parameter $\boldsymbol{\psi}$ in definition 3.18 is

$$\hat{\boldsymbol{\psi}}_{ML} \in \underset{\boldsymbol{\psi}\in\Psi}{\operatorname{argmin}}\, L(\boldsymbol{\psi}). \tag{3.6}$$

Now, in the case of the separation of the variance and correlation hyper-parameters in (3.4), the likelihood criterion becomes

$$L(\sigma^2, \theta) := \frac{1}{n}\ln|\sigma^2\mathbf{R}_{\boldsymbol{\theta}}| + \frac{1}{n}\frac{1}{\sigma^2}\boldsymbol{y}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}, \tag{3.7}$$

Hence, the optimization with respect to σ^2, for fixed $\boldsymbol{\theta}$ can be carried out explicitly. This removes one dimension in the numerical optimization problem. This is summarized in the following proposition.

Proposition 3.21. *The Maximum Likelihood estimator of* $(\sigma^2, \boldsymbol{\theta})$ *is* $(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$*, with*

$$\hat{\boldsymbol{\theta}}_{ML} \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\theta}),$$

with

$$\mathcal{L}(\boldsymbol{\theta}) = \ln\left(\hat{\sigma}^2_{ML}(\boldsymbol{\theta})\right) + \frac{1}{n}\ln|\mathbf{R}_{\boldsymbol{\theta}}|,$$

$$\hat{\sigma}^2_{ML}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{y}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}$$

and

$$\hat{\sigma}^2_{ML} = \hat{\sigma}^2_{ML}(\hat{\boldsymbol{\theta}}_{ML}).$$

Proof. We show, by a simple zero-derivative condition that the minimizer of $L(\sigma^2, \boldsymbol{\theta})$, for fixed $\boldsymbol{\theta}$, is $\hat{\sigma}^2_{ML}(\boldsymbol{\theta})$. We conclude from $\min_{\sigma^2,\boldsymbol{\theta}} L(\sigma^2, \boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \left(\min_{\sigma^2} L(\sigma^2, \boldsymbol{\theta})\right)$. □

Consider now the case of ordinary or universal Kriging. In the case of ordinary or universal Kriging, we always use explicitly the decomposition σ^2, $\boldsymbol{\theta}$. Hence, we will present the Maximum Likelihood equations only in this case.

We denote by $\mathbf{H}$ the $n \times m$ regression matrix of subsection 2.2.2. The likelihood criterion now depends on $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ and is

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) := \frac{1}{n}\ln|\sigma^2\mathbf{R}_{\boldsymbol{\theta}}| + \frac{1}{n\sigma^2}(\boldsymbol{y} - \mathbf{H}\boldsymbol{\beta})^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \mathbf{H}\boldsymbol{\beta}), \tag{3.8}$$

Similarly to the case of simple Kriging, the likelihood criterion of (3.8) can be minimized explicitly with respect to $\boldsymbol{\beta}$ and σ^2, removing $m + 1 = \dim(\boldsymbol{\beta}) + 1$ dimensions in the numerical optimization problem. This is summarized in the following proposition.

Proposition 3.22. *The Maximum Likelihood estimator of $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ is $(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$, with*

$$\hat{\boldsymbol{\theta}}_{ML} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}),$$

with

$$\mathcal{L}(\boldsymbol{\theta}) = \ln\left(\hat{\sigma}^2_{ML}(\boldsymbol{\theta})\right)) + \frac{1}{n}\ln|\mathbf{R}_{\boldsymbol{\theta}}|,$$

$$\hat{\sigma}^2_{ML}(\boldsymbol{\theta}) = \frac{1}{n}(\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\theta}))^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\theta}))^t,$$

$$\hat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\theta}) = (\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}$$

and

$$\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{ML}(\hat{\boldsymbol{\theta}}_{ML}),$$

$$\hat{\sigma}^2_{ML} = \hat{\sigma}^2_{ML}(\hat{\boldsymbol{\theta}}_{ML}).$$

Furthermore, we can also write

$$\hat{\sigma}^2_{ML}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{y}^t\boldsymbol{\Pi}_{\boldsymbol{\theta}}\boldsymbol{y},$$

with

$$\boldsymbol{\Pi}_{\boldsymbol{\theta}} = \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}$$

Proof. Similar to the proof of proposition 3.21. □

Remark 3.23. *In proposition 3.22, if the matrix $\mathbf{H}$ is ill-conditioned, numerical issues can be avoided for the computation of $\hat{\boldsymbol{\theta}}_{ML}$ and $\hat{\sigma}^2_{ML}$. Indeed, let $\mathbf{U}, \mathbf{S}, \mathbf{V}$ be a Singular Value Decomposition of $\mathbf{H}$, with $\mathbf{U}$ of size $n \times m$ so that $\mathbf{U}^t\mathbf{U} = \mathbf{I}_{m,m}$, $\mathbf{S}$ a diagonal matrix of size m, with nonnegative numbers on the diagonal, and $\mathbf{V}$ an orthogonal matrix of size m, so that $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^t$. Then, we can show that the value of $\mathcal{L}(\boldsymbol{\theta})$ and $\hat{\sigma}^2_{ML}(\boldsymbol{\theta})$ are unchanged by replacing the matrix $\mathbf{H}$ by the matrix $\mathbf{U}$. The matrix $\mathbf{U}$ is of course perfectly conditioned.*

However, when the condition number of $\mathbf{H}$ *is large, there is an irreducible numerical imprecision when computing* $\hat{\boldsymbol{\beta}}_{ML}$. *We refer to the discussion of remark 2.31 on this subject. Roughly speaking, if* $\mathbf{H}$ *is ill-conditioned, the design of experiments is either incomplete, or the regression model is over-parameterized. If the design of experiments is incomplete, it has to be extended before considering using the Kriging model for prediction. If the regression model is over-parameterized, a minimal regression model can be obtained from it.*

Restricted Maximum Likelihood

The principle of Restricted Maximum Likelihood (REML) is to make the estimations of the regression coefficient vector $\boldsymbol{\beta}$ and of the covariance function hyper-parameters $\sigma^2, \boldsymbol{\theta}$ totally independent. This is of special interest when the Bayesian prior on $\boldsymbol{\beta}$ of subsection 2.2.2 is considered. Indeed, the estimation of σ^2 and θ is independent of this prior. In the FLICA IV application case of chapter 8, we use the Restricted Maximum Likelihood technique.

First consider the parameterization $K_{\boldsymbol{\psi}}$ of the covariance function in definition 3.18.

Let $\mathbf{W}$ be a $(n-m \times n)$ matrix of full rank so that $\mathbf{WH} = 0$. Note that if $\mathbf{H}$ is not of full rank, then m must be replaced by the rank of $\mathbf{H}$. Then

$$\boldsymbol{w} := \mathbf{W}\boldsymbol{y} \sim \mathcal{N}(0, \mathbf{W}\mathbf{K}_{\boldsymbol{\psi}}\mathbf{W}^t).$$

The law of $\boldsymbol{w}$ is independent of the value of $\boldsymbol{\beta}$. Hence the Restricted Maximum Likelihood Estimator $\hat{\boldsymbol{\psi}}_{REML}$ is the Maximum Likelihood estimator on the transformed observations $\boldsymbol{w}$. Hence, the restricted likelihood criterion is

$$L_R(\boldsymbol{\psi}) := \frac{1}{n}\ln|\mathbf{W}\mathbf{K}_{\boldsymbol{\psi}}\mathbf{W}^t| + \frac{1}{n}\boldsymbol{w}^t(\mathbf{W}\mathbf{K}_{\boldsymbol{\psi}}\mathbf{W}^t)^{-1}\boldsymbol{w}. \tag{3.9}$$

It is shown in [Har74] that changing $\mathbf{W}$ only adds a constant (with respect to $\boldsymbol{\psi}$) term to (3.9). It is also shown in [Har74] how we can avoid a matrix product with $\mathbf{W}$. Indeed let $\mathbf{W}$ so that $\mathbf{W}\mathbf{W}^t = \mathbf{I}_{n-m}$ and $\mathbf{W}^t\mathbf{W} = \mathbf{I}_n - \mathbf{H}(\mathbf{H}^t\mathbf{H})^{-1}\mathbf{H}^t$. Such a matrix $\mathbf{W}$ can be obtained as follows.

Consider a SVD decomposition of $\mathbf{H}$: $\mathbf{H} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^t$, with $\tilde{\mathbf{U}}$ an $n \times n$ orthogonal matrix, $\tilde{\mathbf{V}}$ an $m \times m$ orthogonal matrix and

$$\tilde{\mathbf{S}} = \begin{pmatrix} \mathbf{D} \\ 0_{n-m,m} \end{pmatrix},$$

with $\mathbf{D}$ a $m \times m$ diagonal matrix. Then, with u_i the i-th column of $\tilde{\mathbf{U}}$, with

$$\mathbf{W} = \begin{pmatrix} u_{m+1}^t \\ \vdots \\ u_n^t \end{pmatrix},$$

we have

$$\mathbf{W}\tilde{\mathbf{U}} = \begin{pmatrix} 0_{n-m,m} & \mathbf{I}_{n-m} \end{pmatrix},$$

so that $\mathbf{WH} = 0$. Furthermore we verify $\mathbf{W}\mathbf{W}^t = \mathbf{I}_{n-m}$.

With a matrix $\mathbf{W}$ verifying the conditions above, we have ([Har74])

$$L_R(\boldsymbol{\psi}) = -\frac{1}{n}\ln|\mathbf{H}^t\mathbf{H}| + \frac{1}{n}\ln|\mathbf{K}_{\boldsymbol{\psi}}| + \frac{1}{n}\ln|\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H}| + \frac{1}{n}\boldsymbol{y}^t\boldsymbol{\Pi}_{\boldsymbol{\psi}}\boldsymbol{y}, \tag{3.10}$$

with

$$\boldsymbol{\Pi}_{\boldsymbol{\psi}} = \mathbf{K}_{\boldsymbol{\psi}}^{-1} - \mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}.$$

The REML estimator $\hat{\boldsymbol{\psi}}_{REML}$ is hence

$$\hat{\boldsymbol{\psi}}_{REML} \in \operatorname*{argmin}_{\boldsymbol{\psi}\in\Psi} L_R(\boldsymbol{\psi}). \tag{3.11}$$

Remark 3.24. *Similarly to remark 3.23, it is not an issue for Restricted Maximum Likelihood if* $\mathbf{H}$ *is ill-conditioned. Let* $\mathbf{U}, \mathbf{S}, \mathbf{V}$ *be a Singular Value Decomposition of* $\mathbf{H}$*, with* $\mathbf{U}$ *of size* $n \times m$ *so that* $\mathbf{U}^t\mathbf{U} = \mathbf{I}_{m,m}$*,* $\mathbf{S}$ *a diagonal matrix of size* m*, with nonnegative numbers on the diagonal, and* $\mathbf{V}$ *an orthogonal matrix of size* m*, so that* $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^t$*. Then, we can show that, when replacing* $\mathbf{H}$ *by* $\mathbf{U}$*,* $L_R(\boldsymbol{\psi})$ *in* (3.10) *is unchanged. One can indeed see that, when* $\mathbf{H}$ *becomes ill-conditioned the two diverging terms -* $\ln|\mathbf{H}^t\mathbf{H}|$ *and* $\ln|\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H}|$ *in* (3.10) *actually compensate one another. Also, if* $\mathbf{H}$ *is singular, so that the* $m - m'$ *last diagonal values of* $\mathbf{S}$ *are zero, one can replace* $\mathbf{H}$ *in* (3.10) *by the* $n \times m'$ *matrix* $\mathbf{U}_{m'}$*, composed of the* m' *first columns of* $\mathbf{U}$*, similarly to remark 2.31.*

For the case of the decomposition σ^2, $\boldsymbol{\theta}$, once again, the optimization problem with respect to σ^2 for fixed $\boldsymbol{\theta}$ has an explicit solution, as shown in the following proposition.

Proposition 3.25. *The REML estimator of* $(\sigma^2, \boldsymbol{\theta})$ *is* $(\hat{\sigma}^2_{REML}, \hat{\boldsymbol{\theta}}_{REML})$*, with*

$$\hat{\boldsymbol{\theta}}_{REML} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_R(\boldsymbol{\theta}),$$

with

$$\mathcal{L}_R(\boldsymbol{\theta}) = \frac{n-m}{n}\ln\left(\hat{\sigma}^2_{REML}(\boldsymbol{\theta})\right) + \frac{1}{n}\ln|\mathbf{R}_{\boldsymbol{\theta}}| + \frac{1}{n}\ln|\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}|,$$
$$\hat{\sigma}^2_{REML}(\boldsymbol{\theta}) = \frac{1}{n-m}\boldsymbol{y}^t\boldsymbol{\Pi}_{\boldsymbol{\theta}}\boldsymbol{y},$$
$$\boldsymbol{\Pi}_{\boldsymbol{\theta}} = \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}$$

and

$$\hat{\sigma}^2_{REML} = \hat{\sigma}^2_{REML}(\hat{\boldsymbol{\theta}}_{REML}).$$

Proof. From (3.10) and similar to the proof of proposition 3.21. □

Remark 3.26. *Similarly to remark 3.23, it is not an issue for Restricted Maximum Likelihood if* $\mathbf{H}$ *is ill-conditioned. Let* $\mathbf{U}, \mathbf{S}, \mathbf{V}$ *be a Singular Value Decomposition of* $\mathbf{H}$*, with* $\mathbf{U}$ *of size* $n \times m$ *so that* $\mathbf{U}^t\mathbf{U} = \mathbf{I}_{m,m}$*,* $\mathbf{S}$ *a diagonal matrix of size* m*, with nonnegative numbers on the diagonal, and* $\mathbf{V}$ *an orthogonal matrix of size* m*, so that* $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^t$*. Then, we can show that, when replacing* $\mathbf{H}$ *by* $\mathbf{U}$*,* $\hat{\sigma}^2_{REML}(\boldsymbol{\theta})$ *in proposition 3.25 is unchanged. Furthermore, the marginal restricted likelihood function* $\mathcal{L}_R$ *in proposition 3.25 is changed only by an additive constant (with respect to* $\boldsymbol{\theta}$*).*

Let us discuss briefly the comparison between ML and REML. Both have the same computational cost and essentially require to compute $|\mathbf{R}_{\boldsymbol{\theta}}|$ and solve the linear systems $\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}$ and $\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}$. It is argued in [CL93] that ML has a larger small-sample bias than REML. Indeed, one can see, in the explicit case where $\boldsymbol{\theta}$ is known and σ^2 is estimated, that $\hat{\sigma}^2_{REML}$ is unbiased while $\hat{\sigma}^2_{ML}$ is biased.

3.2.3 Cross Validation for estimation

For this subsection on the Cross Validation estimation, we will make an explicit use of the $(\sigma^2, \boldsymbol{\theta})$ decomposition. Furthermore, we will not consider Cross Validation estimation for the Bayesian case on the regression coefficient vector $\boldsymbol{\beta}$.

Leave-One-Out Mean Square error

The Cross Validation procedure we study in the manuscript is based on the Leave-One-Out (LOO) Mean Square Error (MSE) criterion,

$$LOO(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{y}_{i,\boldsymbol{\theta}}\}^2, \tag{3.12}$$

where, for $1 \leq i \leq n$, $\hat{y}_{i,\theta}$ is the prediction, in (2.9) and (2.15), of y_i according to $y_1, ..., y_{i-1}, y_{i+1}, ..., y_n$, given the covariance function $\sigma^2 R_{\boldsymbol{\theta}}$. One sees that the predictions (2.9) and (2.15) do not depend on the variance hyper-parameter σ^2. This is emphasized by the notation $LOO(\boldsymbol{\theta})$, where the LOO MSE criterion explicitly does not depend on σ^2.

$\boldsymbol{\theta}$ is estimated by minimizing the LOO MSE criterion,

$$\hat{\boldsymbol{\theta}}_{LOO} \in \underset{\theta\in\Theta}{\operatorname{argmin}}\, LOO(\boldsymbol{\theta}). \tag{3.13}$$

Leave-One-Out Predictive variance criterion

The variance hyper-parameter σ^2 can not be estimated using the LOO MSE criterion. This criterion reflects the quality of the point wise prediction of (2.9) and (2.15). The other intuitive criterion for a Kriging model would be a criterion reflecting the quality of the predictive variances for these pointwise predictions. We study the criterion based on

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_{i,\hat{\boldsymbol{\theta}}_{LOO}})^2}{\sigma^2\hat{c}^2_{i,\hat{\boldsymbol{\theta}}_{LOO}}}, \tag{3.14}$$

where $\hat{c}^2_{i,\boldsymbol{\theta}}$ and $\sigma^2\hat{c}^2_{i,\boldsymbol{\theta}}$ are the predictive variances of y_i according to $y_1, ..., y_{i-1}, y_{i+1}, ..., y_n$, given the covariance functions $R_{\boldsymbol{\theta}}$ and $\sigma^2 R_{\boldsymbol{\theta}}$. It is noted in [Cre93] p.102 that if σ^2 is a correct estimate of the variance parameter, then we should expect (3.14) to be close to 1. The principle of the CV estimation of σ^2 is to set this criterion equal to 1 exactly. Thus, the CV estimation of σ^2 is

$$\hat{\sigma}^2_{LOO} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_{i,\hat{\boldsymbol{\theta}}_{LOO}})^2}{\hat{c}^2_{i,\hat{\boldsymbol{\theta}}_{LOO}}}, \tag{3.15}$$

with $\hat{\boldsymbol{\theta}}_{LOO}$ as in (3.13).

To summarize, the general CV procedure we study is a two-step procedure. In a first step, the correlation hyper-parameters are selected according to a mean square error criterion. In a second step, the global variance hyper-parameter is selected, so that the predictive variances are adapted to the Leave-One-Out prediction errors.

Matrix form criteria

Using the virtual Cross Validation formulas of proposition 2.35, we can write the estimators $\hat{\boldsymbol{\theta}}_{LOO}$ and $\hat{\sigma}^2_{LOO}$ of (3.13) and (3.15) with explicit quadratic forms.

First, the LOO MSE criterion for $\hat{\boldsymbol{\theta}}_{LOO}$ can be written as

$$LOO_{\boldsymbol{\theta}} = \frac{1}{n}\boldsymbol{y}^t\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}Diag(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-})^{-2}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\boldsymbol{y}, \tag{3.16}$$

with $\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}$ being $\mathbf{R}_{\boldsymbol{\theta}}^{-1}$ in the simple Kriging case and $\mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}$ in the (frequentist) ordinary or universal Kriging cases. The expression (3.16) allows to estimate $\boldsymbol{\theta}$ by CV by minimizing a criterion that has the same computational complexity of $O(n^3)$ as ML.

The explicit estimation of σ^2 by CV in (3.15) also has the explicit quadratic form expression

$$\hat{\sigma}^2_{LOO} = \frac{1}{n}\boldsymbol{y}^t\tilde{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{LOO}}^{-}Diag(\tilde{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{LOO}}^{-})^{-1}\tilde{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{LOO}}^{-}\boldsymbol{y}. \tag{3.17}$$

Remark 3.27. *Consider the universal Kriging case. Similarly to the ML and REML estimators, if the matrix* $\mathbf{H}$ *is ill-conditioned, numerical issues can be avoided for the computation of* $\hat{\boldsymbol{\theta}}_{LOO}$ *and* $\hat{\sigma}^2_{LOO}$*. Indeed, let* $\mathbf{U}, \mathbf{S}, \mathbf{V}$ *be a Singular Value Decomposition of* $\mathbf{H}$*, with* $\mathbf{U}$ *of size* $n \times m$ *so that* $\mathbf{U}^t\mathbf{U} = \mathbf{I}_{m,m}$*,* $\mathbf{S}$ *a diagonal matrix of size* m*, with nonnegative numbers on the diagonal, and* $\mathbf{V}$ *an orthogonal matrix of size* m*, so that* $\mathbf{H} = \mathbf{USV}^t$*. Then, we can show that the values of* (3.16) *and* (3.17) *are unchanged by replacing the matrix* $\mathbf{H}$ *by the matrix* $\mathbf{U}$*.*

Discussion on the Leave-One-Out criteria studied

The CV procedure of (3.12) and (3.15) gives the priority first to the point wise prediction at a new point, and second to the predictive variance for this new point. Furthermore, it addresses this double objective by using criteria that are the direct empirical counterpart of this double objective.

The double remark above may raise two interrogations on the CV estimation. First, on can argue that the predictive means and variances may not always be the priority for a Kriging model. For instance, we can be more interested in the estimation of the conditional correlation between $Y(\boldsymbol{x}^{(new,1)})$ and $Y(\boldsymbol{x}^{(new,2)})$ at two different new points. We will not discuss this point any longer, since it is intrinsically dependent on the particular application of the Kriging model at hand. We will just mention that, in many application cases, priority is given to pointwise predictive means and variances.

Second, when established that the priority is given to having accurate point-wise predictive means and variances, the procedure of (3.12) and (3.15) also constitutes a particular strategy for this priority. More precisely, the LOO criterion (3.12) is interpreted as a direct approximation of an underlying integrated prediction Mean Square Error. Another criterion that appears frequently in the literature ([RW06], chapter 5, [ZW10], [SK01]) is the LOO log predictive probability, which is

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\ln\left(\sigma^2\hat{c}^2_{i,\boldsymbol{\theta}}\right) + \frac{(y_i - \hat{y}_{i,\boldsymbol{\theta}})^2}{\sigma^2\hat{c}^2_{i,\boldsymbol{\theta}}}\right\}, \tag{3.18}$$

and is minimized jointly w.r.t σ^2 and $\boldsymbol{\theta}$. The LOO log predictive probability criterion consists in maximizing the product of the conditional likelihoods of each of the LOO observations according to the remaining ones. It could be argued that doing so can also improve the accuracy of the predictions $\hat{y}_{i,\hat{\boldsymbol{\theta}}}$. Indeed, for instance, in (3.18), large prediction errors are divided by predictive variances that are more likely to be large as well. This results in homogenizing the terms in (3.18), thus potentially reducing the variance of the LOO log predictive probability estimator

minimizing their sum. Furthermore, let us note that the criterion in (3.18) is minimized jointly with respect to σ^2 and $\boldsymbol{\theta}$ and hence does not need this separation, contrary to the procedure based on (3.12) and (3.15).

We essentially believe that the choice of the LOO procedure to be used is still an open problem. In chapter 6, we show that the LOO procedure of (3.12) and (3.15) is more robust than ML, when the family of covariance functions in which the selection is carried out is far from the true covariance function of the Gaussian process. It is unclear yet whether the procedure based on (3.18) would be as robust. Finally, the questions related to the choice of the LOO procedure are further discussed in the perspectives in chapter 10.

3.2.4 Gradients of the different criteria

In this subsection 3.2.4, we give the expressions of the gradients of the criteria that need to be minimized numerically. Explicit expressions of the gradient are indeed useful for gradient-based optimization algorithms.

In all subsection 3.2.4, let ψ_i , $i \in \{1, ..., p\}$ be a component of $\boldsymbol{\psi}$. Let also, depending on the situation, θ_i, $i \in \{1, ..., p-1\}$ be a component of $\boldsymbol{\theta}$.

Proposition 3.28 gives the gradient of the likelihood criterion in the simple Kriging case.

Proposition 3.28. *Let $L(\boldsymbol{\psi})$ be the likelihood criterion of* (3.5)*. Then*

$$\frac{\partial}{\partial \psi_i} L(\boldsymbol{\psi}) = \frac{1}{n} Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i}\right) - \frac{1}{n} \boldsymbol{y}^t \mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \boldsymbol{y}.$$

Proof. The proof can be found in [MM84]. It is a straightforward calculation based on $\frac{\partial}{\partial \psi_i} \ln |\mathbf{M}_{\boldsymbol{\psi}}|$ $= Tr\left(\mathbf{M}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{M}_{\boldsymbol{\psi}}}{\partial \psi_i}\right)$ and $\frac{\partial}{\partial \psi_i} \mathbf{M}_{\boldsymbol{\psi}}^{-1} = -\mathbf{M}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{M}_{\boldsymbol{\psi}}}{\partial \psi_i} \mathbf{M}_{\boldsymbol{\psi}}^{-1}$. □

For all the criteria in the universal Kriging case, expressions are simplified by making use of the following lemma.

Lemma 3.29. *Let*

$$\boldsymbol{\Pi}_{\boldsymbol{\theta}} = \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1} \mathbf{H} (\mathbf{H}^t \mathbf{R}_{\boldsymbol{\theta}}^{-1} \mathbf{H})^{-1} \mathbf{H}^t \mathbf{R}_{\boldsymbol{\theta}}^{-1}.$$

Then,

$$\frac{\partial \boldsymbol{\Pi}_{\boldsymbol{\theta}}}{\partial \theta_i} = -\boldsymbol{\Pi}_{\boldsymbol{\theta}} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Pi}_{\boldsymbol{\theta}}$$

Proof. Straightforward calculation based on $\frac{\partial}{\partial \theta_i} \mathbf{M}_{\boldsymbol{\theta}}^{-1} = -\mathbf{M}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{M}_{\boldsymbol{\theta}}}{\partial \theta_i} \mathbf{M}_{\boldsymbol{\theta}}^{-1}$. □

Proposition 3.30 gives the gradient of the marginal likelihood criterion in the simple and universal Kriging cases.

Proposition 3.30. *Let $\mathcal{L}(\boldsymbol{\theta})$ be the marginal likelihood criterion of proposition 3.21 for the simple Kriging case and proposition 3.22 in the universal Kriging case. Then*

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\theta) = \frac{1}{n} Tr\left(\mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i}\right) - \frac{\boldsymbol{y}^t \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}}{\boldsymbol{y}^t \mathbf{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}},$$

in the simple Kriging case, and

$$\frac{\partial}{\partial \theta_i} \mathcal{L}_{\theta} = \frac{1}{n} Tr\left(\mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i}\right) - \frac{\boldsymbol{y}^t \boldsymbol{\Pi}_{\boldsymbol{\theta}} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \boldsymbol{\Pi}_{\boldsymbol{\theta}} \boldsymbol{y}}{\boldsymbol{y}^t \boldsymbol{\Pi}_{\boldsymbol{\theta}} \boldsymbol{y}},$$

with

$$\mathbf{\Pi}_{\boldsymbol{\theta}} = \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1},$$

in the universal Kriging case.

Proof. Straightforward calculation based on lemma 3.29, $\frac{\partial}{\partial\theta_i}\mathbf{M}_{\boldsymbol{\theta}}^{-1} = -\mathbf{M}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{M}_{\boldsymbol{\theta}}^{-1}$ and $\frac{\partial}{\partial\theta_i}\ln|\mathbf{M}_{\boldsymbol{\theta}}| = Tr\left(\mathbf{M}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\theta}}}{\partial\theta_i}\right)$. □

Proposition 3.31 gives the gradient of the restricted likelihood criterion.

Proposition 3.31. *Let $L_R(\boldsymbol{\psi})$ be the restricted likelihood criterion of* (3.10)*. Then*

$$\frac{\partial}{\partial\psi_i}L_R(\boldsymbol{\psi}) = \frac{1}{n}Tr\left(\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{\Pi}_{\boldsymbol{\psi}}\right) - \frac{1}{n}\boldsymbol{y}^t\mathbf{\Pi}_{\boldsymbol{\psi}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{\Pi}_{\boldsymbol{\psi}}\boldsymbol{y},$$

with

$$\mathbf{\Pi}_{\boldsymbol{\psi}} = \mathbf{K}_{\boldsymbol{\psi}}^{-1} - \mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}.$$

Proof. Straightforward calculation based on lemma 3.29, $\frac{\partial}{\partial\psi_i}\mathbf{M}_{\boldsymbol{\psi}}^{-1} = -\mathbf{M}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{M}_{\boldsymbol{\psi}}^{-1}$ and $\frac{\partial}{\partial\psi_i}\ln|\mathbf{M}_{\boldsymbol{\psi}}| = Tr\left(\mathbf{M}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\psi}}}{\partial\psi_i}\right)$. □

Proposition 3.32 gives the gradient of the marginal restricted likelihood criterion.

Proposition 3.32. *Let $\mathcal{L}_R(\boldsymbol{\theta})$ be the marginal restricted likelihood criterion of proposition 3.25. Then*

$$\frac{\partial}{\partial\theta_i}\mathcal{L}_R(\boldsymbol{\theta}) = \frac{1}{n}Tr\left(\frac{\partial\mathbf{R}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{\Pi}_{\boldsymbol{\theta}}\right) - \frac{n-m}{n}\frac{\boldsymbol{y}^t\mathbf{\Pi}_{\boldsymbol{\theta}}\frac{\partial\mathbf{R}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{\Pi}_{\boldsymbol{\theta}}\boldsymbol{y}}{\boldsymbol{y}^t\mathbf{\Pi}_{\boldsymbol{\theta}}\boldsymbol{y}},$$

with

$$\mathbf{\Pi}_{\boldsymbol{\theta}} = \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}.$$

Proof. Straightforward calculation based on lemma 3.29, $\frac{\partial}{\partial\theta_i}\mathbf{M}_{\boldsymbol{\theta}}^{-1} = -\mathbf{M}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{M}_{\boldsymbol{\theta}}^{-1}$ and $\frac{\partial}{\partial\theta_i}\ln|\mathbf{M}_{\boldsymbol{\theta}}| = Tr\left(\mathbf{M}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\theta}}}{\partial\theta_i}\right)$. □

Proposition 3.33 gives the gradient of the CV criterion for the simple and universal Kriging cases.

Proposition 3.33. *Let $LOO(\boldsymbol{\theta})$ be the LOO criterion of* (3.16)*. Then,*

$$\begin{aligned}\frac{\partial}{\partial\theta_i}LOO(\boldsymbol{\theta}) = & -\tfrac{2}{n}\boldsymbol{y}^t\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}Diag(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-})^{-2}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\frac{\partial\mathbf{R}_{\boldsymbol{\theta}}}{\partial\theta_i}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\boldsymbol{y} \\ & +\tfrac{2}{n}\boldsymbol{y}^t\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}Diag(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-})^{-2}Diag\left(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\frac{\partial\mathbf{R}_{\boldsymbol{\theta}}}{\partial\theta_i}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\right)Diag(\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-})^{-1}\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}\boldsymbol{y},\end{aligned}$$

with $\tilde{\mathbf{R}}_{\boldsymbol{\theta}}^{-}$ being $\mathbf{R}_{\boldsymbol{\theta}}^{-1}$ in the simple Kriging case and $\mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H}(\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}$ in the universal Kriging case.

Proof. The proof for the simple Kriging case can be found in [Bac13]. The proof for the universal Kriging case is a straightforward but rather long calculation based on lemma 3.29, $\frac{\partial}{\partial\theta_i}\mathbf{M}_{\boldsymbol{\theta}}^{-1} = -\mathbf{M}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{M}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{M}_{\boldsymbol{\theta}}^{-1}$ and $\frac{\partial}{\partial\theta_i}Diag\left(\mathbf{M}_{\boldsymbol{\theta}}\right) = Diag\left(\frac{\partial\mathbf{M}_{\boldsymbol{\theta}}}{\partial\theta_i}\right)$. □

Let us discuss briefly the computational costs of the explicit gradients above. In the universal Kriging case, we consider the case where m is small compared to n, which holds for all the applications treated in the manuscript. For all the likelihood-oriented criteria, if the inverse of the covariance, or correlation, matrix is calculated and stored, then computing the gradient can be done with $O(n^2)$ operations. Indeed, a term like $Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i}\right) = \sum_{j=1}^{n}\sum_{k=1}^{n}\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)_{j,k}\left(\frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i}\right)_{k,j}$ can be calculated in $O(n^2)$ if $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ is already calculated. Hence the computational cost for calculating a likelihood criterion and its gradient is a $O(n^3)$, and is independent of the dimension of $\boldsymbol{\theta}$ or $\boldsymbol{\psi}$. However, for CV, we have to compute e.g. $Diag\left(\mathbf{R}_{\boldsymbol{\theta}}^{-1}\frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\right)$ for each component of $\boldsymbol{\theta}$. Hence, the computational cost for calculating a CV criterion and its gradient is $O(n^3)$, like ML, but is proportional to the dimension of $\boldsymbol{\theta}$ or $\boldsymbol{\psi}$.

3.2.5 The challenge of taking into account the uncertainty on the covariance function

The Kriging equations of subsection 2.2.2 assume that the covariance function of the Gaussian process Y is known. In practice, this function is estimated beforehand, yielding plug-in ([Ste99], chapter 6.8) prediction equations. The plug-in approach does not take into account the randomness of the covariance function estimator. Let $\hat{\boldsymbol{\psi}}$ be an estimator of the covariance hyper-parameter $\boldsymbol{\psi}$ that verifies, for any two $m \times 1$ vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, $\hat{\boldsymbol{\psi}}(\boldsymbol{v}) = \hat{\boldsymbol{\psi}}(\boldsymbol{v} + \mathbf{H}\boldsymbol{w})$, with $\mathbf{H}$ the regression matrix. In the simple Kriging case, this conventionally adds no condition on the estimator $\hat{\boldsymbol{\psi}}$. In [Ste99] p.201, the estimator $\hat{\boldsymbol{\psi}}$ is said to depend only on the contrasts of $\boldsymbol{y}$. Note that all the estimators studied in the manuscript do depend only on the contrasts of $\boldsymbol{y}$. Indeed, for example, the likelihood criterion in proposition 3.22 is written as a function of $\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{\beta}}(\boldsymbol{y})$, with $\hat{\boldsymbol{\beta}}(\boldsymbol{y} + \mathbf{H}\boldsymbol{v}) = \hat{\boldsymbol{\beta}}(\boldsymbol{y}) + \boldsymbol{v}$. Similarly, for CV in (3.16), the LOO criterion is written $\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}$, with $\mathbf{M}$ a matrix so that $\mathbf{MH} = 0$. It is shown in [ZC92], and discussed in [Ste99] p.201 that, in this case, the estimator $\hat{\boldsymbol{\psi}}$ is independent of the prediction error, with the true covariance hyper-parameter, $\hat{y}_{\boldsymbol{\psi}^{(0)},0} - y_0$ at a new point $\boldsymbol{x}^{(0)}$. As a result, we have that $\hat{y}_{\boldsymbol{\psi}^{(0)},0} - y_0$ is independent of $\hat{y}_{\boldsymbol{\psi}^{(0)},0} - \hat{y}_{\hat{\boldsymbol{\psi}},0}$. Hence

$$\mathbb{E}((\hat{y}_{\hat{\boldsymbol{\psi}},0} - y_0)^2) = \mathbb{E}((\hat{y}_{\boldsymbol{\psi}^{(0)},0} - y_0)^2) + \mathbb{E}((\hat{y}_{\boldsymbol{\psi}^{(0)},0} - \hat{y}_{\hat{\boldsymbol{\psi}},0})^2), \tag{3.19}$$

so that the prediction MSE is always larger when using an estimated covariance hyper-parameter than when using the true hyper-parameter. As a result, the predictive variances obtained from the plug-in approach may be overoptimistic.

This issue is well known in Kriging models, and is difficult to address. Until now, the majority of the research on Kriging models adopt the plug-in approach, unless being explicitly oriented toward, for instance, computing predictive variances explicitly taking the uncertainty on the covariance function into account. In this manuscript, we also adopt the plug-in approach.

We now mention some alternatives to the plug-in approach, to be found in the literature.

In [ZZ06], the distribution of the estimation error is approximated by a centered Gaussian distribution with covariance matrix equal to the inverse of the Fisher information matrix. Then, this approximated distribution is propagated in (3.19), by using a first order Taylor series expansion, yielding an estimated predictive variance that is larger than the plug-in one.

Using a Bayesian prior on the covariance hyper-parameters will, in nature, yield predictive means and variances taking the posterior distribution of the covariance hyper-parameter into account (in fact the conditional distribution of $Y(\boldsymbol{x}^{(new)})$ will not be Gaussian anymore). This is done for instance in [BBV11] in a Kriging-based optimization context. Nevertheless, it is worth mentioning that the Bayesian approach yields an increased computational cost.

A parametric bootstrap approach is also presented in [Ste99], p.202. In essence, it consists in exploiting the independence between the two random variables on the right-hand side of (3.19). The principle is to assume that the estimator $\hat{\boldsymbol{\psi}}$ obtained is the true one, to sample n_b new observation vectors with the distribution obtained from it, and to obtain from them a sample $(\hat{\boldsymbol{\psi}}^{(i)})_{1\leq i\leq n_b}$ of estimated covariance hyper-parameters. Then, the distribution of $(\hat{y}_{\boldsymbol{\psi}^{(0)},0} - y_0)$, obtained from the Kriging equations with hyper-parameter $\hat{\boldsymbol{\psi}}$, is convolved with the bootstrap empirical distribution of $(\hat{y}_{\boldsymbol{\psi}^{(0)},0} - \hat{y}_{\hat{\boldsymbol{\psi}}^{(i)},0})_{1\leq i\leq n_b}$. From the independence between the two random variables on the right-hand side of (3.19), this procedure yields an approximate distribution of $(\hat{y}_{\hat{\boldsymbol{\psi}},0} - y_0)$.

Chapter 4

Asymptotic results for Kriging

In this chapter 4, we review some existing asymptotic results for Kriging. We first present in section 4.1 the two classical asymptotic frameworks: increasing-domain asymptotics and fixed-domain asymptotics. Then, in subsection 4.2.1 we review the existing fixed-domain asymptotic results for the consistency of the Kriging predictions, in both the cases where the Gaussian process assumption is well-specified or misspecified. In subsection 4.2.2, we review the results of [Ste99] for asymptotic optimality of Kriging predictions with misspecified mean and covariance functions. In section 4.3, we present the asymptotic results for estimation, in both asymptotic frameworks. Subsection 4.3.1 is dedicated to ML in the increasing-domain asymptotics framework. Subsection 4.3.2 addresses various estimators in fixed-domain asymptotics.

4.1 Two asymptotic frameworks

There is a fundamental difference between the asymptotic framework in the *iid* case of subsection 3.1.2 and the asymptotic framework for Kriging. In the *iid* case, letting $n \to +\infty$ defines the asymptotic framework without ambiguity. However, for Kriging, when letting the number of observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ grow to $+\infty$, the position of the points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ still remains to be set.

Remark 4.1. *In an asymptotic framework for Kriging, it is not necessary that the observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ be part of a sequence $(\boldsymbol{x}^{(i)})_{i \in \mathbb{N}^*}$. For instance, we may consider, for each $n \in \mathbb{N}^*$, observing Y on $[0,1]$ at the regular grid $\{\frac{i}{n}, 1 \leq i \leq n\}$. Hence, we can write the observation points at step n $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$. Nevertheless, for concision, we write the observation points at step n $\{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}\}$, even when there is no sequence $(\boldsymbol{x}^{(i)})_{i \in \mathbb{N}^*}$.*

In an asymptotic framework for Kriging, even the observation domain $\mathcal{D}$ may depend on n. We may emphasize it by writing it $\mathcal{D}_n$. Informally, the domain $\mathcal{D}_n$, for a fixed n, corresponds to the region of $\mathbb{R}^d$ where we are interested in predicting the Gaussian process Y. Hence, the observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ should cover all the domain $\mathcal{D}_n$ and only the domain $\mathcal{D}_n$.

Two main asymptotic frameworks exist, characterized by the variation of $\mathcal{D}_n$ with respect to n. In the fixed-domain asymptotic framework ([Ste99], p62), $\mathcal{D}_n$ is independent of n and corresponds to a compact set $\mathcal{D}$. Because the goal is to predict Y in all $\mathcal{D}$, it is assumed

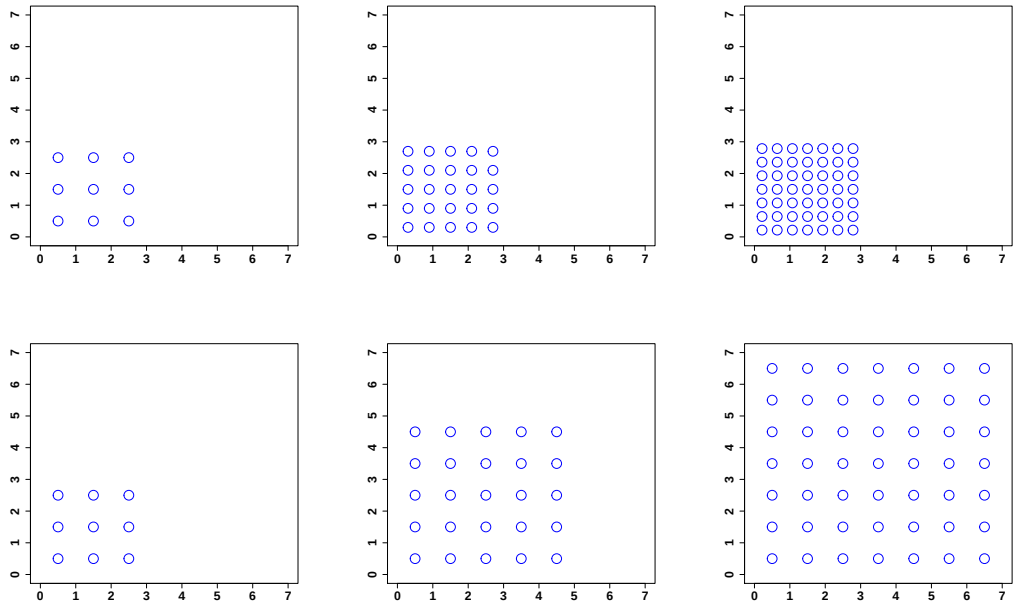

Figure 4.1: Illustration of fixed and increasing-domain asymptotics. Top: plot of 3×3, 5×5 and 7×7 regular grids in the fixed-domain asymptotic framework. Bottom: plot of 3×3, 5×5 and 7×7 regular grids in the increasing-domain asymptotic framework.

that the observation points become dense in $\mathcal{D}$. If the observation points are not taken from a sequence, we mean by dense that, with d_n the maximum over $\boldsymbol{x} \in \mathcal{D}$ of the distance between $\boldsymbol{x}$ and $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$, d_n vanishes to zero when $n \to +\infty$. One classical example of fixed-domain asymptotic framework is a dense regular grid for the observation points on $\mathcal{D} = [0, 1]$, i.e. $\boldsymbol{x}^{(i,n)} = \frac{i}{n}$.

In the increasing-domain asymptotic framework, $\mathcal{D}_n \subset \mathcal{D}_{n+p}$ for $n, p \in \mathbb{N}^*$, and $\bigcup_{n \in \mathbb{N}^*} \mathcal{D}_n = \mathbb{R}^d$ or $\bigcup_{n \in \mathbb{N}^*} \mathcal{D}_n = (\mathbb{R}^+)^d$. Furthermore, it is mentioned in [Ste99] p62 that, in increasing-domain asymptotics, the ratio of n on the volume of $\mathcal{D}_n$ is bounded. This implies that the observation points do not become dense in $\mathcal{D}_n$. This is the case when it is assumed that there exists a positive minimal distance between two different points $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$. In the manuscript, we make this assumption for increasing-domain asymptotics. An example of increasing-domain asymptotic framework, that we treat in chapter 5, is when for all $N \in \mathbb{N}^*$, $\mathcal{D}_{N^d} = [0, N]^d$ and the $(\boldsymbol{x}^{(n)})_{n \in \mathbb{N}^*}$ constitute a sequence so that for all $N \in \mathbb{N}^*$ $\{\boldsymbol{x}^{(i)}, 1 \leq i \leq N^d\} = \{1, ..., N\}^d$. Roughly speaking, this is a tensorized regular grid, with inter-point spacing 1.

In figure 4.1, we plot three regular grids with $n = 3^2$, $n = 5^2$ and $n = 7^2$ observation points, in both the fixed-domain and increasing-domain asymptotic frameworks. We clearly see the two different asymptotic behaviors: in fixed-domain asymptotics, the first 3×3 regular grid already covers all the prediction domain, and the two other regular grids cover it more densely. In increasing-domain asymptotics, the size of the prediction domain, covered by the regular grid, increases, but the density of the prediction points is constant.

Finally, let us mention that we can also study an asymptotic framework when the domain $\mathcal{D}_n$ grows to $\mathbb{R}^d$, but more slowly, so that the set of observation points $\{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}\}$ becomes

dense in $\mathcal{D}_n$. For instance, consider $\mathcal{D}_n = [-\sqrt{n}, \sqrt{n}]$ and $x^{(i,n)} = -\sqrt{n} + \frac{2i}{\sqrt{n}}$. This asymptotic framework is called hybrid asymptotics in [Ste99] and mixed increasing-domain asymptotics in [LM04]. We will not treat it in the manuscript.

4.2 Asymptotic results for prediction with fixed covariance function

4.2.1 Consistency

Consider the fixed-domain asymptotic framework and consider a fixed point $\boldsymbol{x} \in \mathcal{D}$. The goal of this subsection 4.2.1 is to answer the question: when $n \to +\infty$ does the prediction error of $Y(\boldsymbol{x})$ given $y_1, ..., y_n$ at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ go to zero?

Note that Kriging predictions are not expected to be consistent in the increasing-domain asymptotic framework. Indeed, the interpoint distance is bounded away from zero when $n \to +\infty$, so that most of the points in the prediction domain remain isolated.

Naturally, in the fixed-domain asymptotic framework, it is desirable that Kriging predictions are consistent. Indeed, when predicting a continuous function on a fixed bounded domain $\mathcal{D}$, many simple approximation methods have their prediction error vanishing when the number of observations goes to $+\infty$.

We will first answer the question when Y is a Gaussian process with known mean structure and covariance function. This corresponds to the question of the consistency of Kriging, when the Gaussian process assumption is correct and the mean structure and the covariance function are well-specified.

Second, we will consider the question when the observations stem from a deterministic continuous function f, which is modeled as a trajectory of a Gaussian process Y with fixed mean structure and covariance function. This second case can include the case of a misspecification of the mean structure or covariance function of the Gaussian process Y, when this Gaussian process does yield continuous trajectories. This second question corresponds to the robustness of Kriging in the case where the Gaussian process assumption is wrong. This has an important practical influence, since Kriging models are often applied, for instance, to approximate deterministic computer models. We will review some results in the literature, but we will also see that this question is not fully solved yet, to the best of our knowledge.

Consistency when the Gaussian process assumption is correct

Consistency is proved in proposition 4.2, in the case where the Gaussian process is observed without measurement error.

Proposition 4.2. *Consider the universal Kriging framework with a Gaussian process Y on $\mathcal{D} \subset \mathbb{R}^d$. Assume that the mean function and the covariance function of Y are continuous. Consider a fixed point $\boldsymbol{x} \in \mathcal{D}$. Assume that Y is observed exactly at $\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}$ and that the distance between $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$ and $\boldsymbol{x}$ goes to zero. Then*

$$\mathbb{E}((\hat{y}(\boldsymbol{x}) - Y(\boldsymbol{x}))^2) \to_{n \to +\infty} 0,$$

where $\hat{y}(\boldsymbol{x})$ is the universal Kriging prediction of (2.15).

Proof. Consider a sequence p_n so that $\boldsymbol{x}^{(p_n,n)}$ goes to $\boldsymbol{x}$ as $n \to +\infty$. Then, because the mean and covariance functions of Y are continuous, the linear predictor $\tilde{y}(\boldsymbol{x}) = Y(\boldsymbol{x}^{(p_n,n)})$ verifies (proposition 2.20)

$$\mathbb{E}((\tilde{y}(\boldsymbol{x}) - Y(\boldsymbol{x}))^2) \to_{n\to+\infty} 0.$$

Since $\hat{y}(\boldsymbol{x})$ minimizes the MSE among all linear predictors, it also verifies

$$\mathbb{E}((\hat{y}(\boldsymbol{x}) - Y(\boldsymbol{x}))^2) \to_{n\to+\infty} 0.$$

□

Remark 4.3. *The condition that the distance between $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$ and $\boldsymbol{x}$ goes to zero is meant naturally to hold for all $\boldsymbol{x} \in \mathcal{D}$. Hence, proposition 4.2 does hold only in the fixed-domain asymptotic framework, as discussed above.*

Proposition 4.4 consider the case where noisy observations of the Gaussian process are made. The proposition assesses that, in the fixed-domain asymptotic framework, the prediction will be consistent despite the measurement errors.

Proposition 4.4. *Consider the universal Kriging framework with a Gaussian process Y on $\mathcal{D} \subset \mathbb{R}^d$. Assume that the mean function and the covariance function of Y are continuous. Assume that Y is observed at $\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}$ with observed value $\boldsymbol{y}_{i,n} = Y(\boldsymbol{x}^{i,n}) + \epsilon_{i,n}$ where the $\epsilon_{i,n}$ are iid and follow a $\mathcal{N}(0, \sigma^2_{mes})$ distribution. Consider a fixed point $\boldsymbol{x} \in \mathcal{D}$. Assume that, for any open ball with center $\boldsymbol{x}$ and positive radius, the number of points in $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$ belonging to the ball goes to $+\infty$ when n goes to $+\infty$. Then*

$$\mathbb{E}((\hat{y}(\boldsymbol{x}) - Y(\boldsymbol{x}))^2) \to_{n\to+\infty} 0,$$

where $\hat{y}(\boldsymbol{x})$ is the Kriging prediction of (2.15)*, in the noisy case.*

Proof. There exists a sequence of radius $r_n \to 0$ so that the number $n_{b,n}$ of points of $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$ belonging to the ball with center $\boldsymbol{x}$ and radius $\boldsymbol{r}_n$ goes to $+\infty$. Consider the linear predictor

$$\tilde{y}(\boldsymbol{x}) = \frac{1}{n_{b,n}} \sum_{i||\boldsymbol{x}^{(i,n)} - \boldsymbol{x}| \leq r_n} y_{i,n}.$$

Basically this predictor is the empirical mean of a large enough number of observations whose

observation points are close enough to the prediction point. Then

$$
\begin{aligned}
&\mathbb{E}\left(\left(Y(\boldsymbol{x})-\tilde{y}(\boldsymbol{x})\right)^2\right)\\
&=\mathbb{E}\left(\left(Y(\boldsymbol{x})-\frac{1}{n_{b,n}}\sum_{i||\boldsymbol{x}^{(i,n)}-\boldsymbol{x}|\leq r_n} y_{i,n}\right)^2\right)\\
&=\mathbb{E}\left(\left(Y(\boldsymbol{x})-\frac{1}{n_{b,n}}\sum_{i||\boldsymbol{x}^{(i,n)}-\boldsymbol{x}|\leq r_n} Y(\boldsymbol{x}^{(i,n)})-\frac{1}{n_{b,n}}\sum_{i||\boldsymbol{x}^{(i,n)}-\boldsymbol{x}|\leq r_n} \epsilon_{i,n}\right)^2\right)\\
&=\mathbb{E}\left(\left(Y(\boldsymbol{x})-\frac{1}{n_{b,n}}\sum_{i||\boldsymbol{x}^{(i,n)}-\boldsymbol{x}|\leq r_n} Y(\boldsymbol{x}^{(i,n)})\right)^2\right)+\frac{\sigma^2_{mes}}{n_{b,n}}\\
&\to_{n\to+\infty} 0.
\end{aligned}
$$

We conclude by mentioning that the MSE of $\hat{y}(\boldsymbol{x})$ is smaller than the MSE of $\tilde{y}(\boldsymbol{x})$. □

Note that in proposition 4.4 the condition that for any open ball with center $\boldsymbol{x}$ and positive radius, the number of points in $\{\boldsymbol{x}^{(1,n)},...,\boldsymbol{x}^{(n,n)}\}$, belonging to the ball, goes to $+\infty$ holds in the fixed-domain asymptotic framework.

Consistency when the Gaussian process assumption is incorrect

The following proposition shows that the universal Kriging equation (2.15) gives a consistent prediction, when the observations stem from a deterministic smooth function. This deterministic smooth function can be the trajectory of a random process with almost surely smooth trajectories. In practice it can also be a deterministic computer model with a smooth relation between its inputs and its output.

Thus, proposition 4.5 assesses the robustness of Kriging to the misspecifications of the Gaussian process assumption. It is hence complementary to propositions 4.2 and 4.4, which assess the efficiency of Kriging in the "favorable" case when the Gaussian process assumption holds.

Proposition 4.5. *Consider the universal Kriging framework with a Gaussian process Y on a compact $\mathcal{D}\subset\mathbb{R}^d$, with a fixed continuous mean structure and continuous stationary covariance function K. Assume that there exists $k<+\infty$ so that the Fourier transform $\hat{K}$ of K is positive-valued and verifies $\hat{K}(\boldsymbol{\omega})|\boldsymbol{\omega}|^k\to+\infty$ when $|\boldsymbol{\omega}|\to+\infty$. Consider a fixed point $\boldsymbol{x}\in\mathcal{D}$. Assume that an infinitely differentiable function f is observed exactly at $\boldsymbol{x}^{(1,n)},...,\boldsymbol{x}^{(n,n)}\in\mathcal{D}$. Assume that, for any open ball with center $\boldsymbol{x}$ and positive radius, the number of points in $\{\boldsymbol{x}^{(1,n)},...,\boldsymbol{x}^{(n,n)}\}$ belonging to the ball goes to $+\infty$ when n goes to $+\infty$. Let $\hat{y}(\boldsymbol{x})$ be the Kriging prediction (2.15), with possibly an inappropriately assumed iid measurement error with variance $\sigma^2_{mes}\geq 0$, of $f(\boldsymbol{x})$ with observations $y_1=f(\boldsymbol{x}^{(1)})$,..., $y_n=f(\boldsymbol{x}^{(n)})$. Then $\hat{y}(\boldsymbol{x})\to f(\boldsymbol{x})$ when $n\to+\infty$.*

Proof. In [YS85], it is claimed that the proposition holds with the significantly less restrictive condition that f is continuous. However, the proof given is flawed, as explained in [VB10], and as we discuss below. Nevertheless, the first step of the proof given in [YS85] proves the

proposition for smooth functions f, as stated here. We find this part of the proof instructive and hence we reproduce it. Note also, in our case, the slight modification of taking into account an inappropriate measurement error assumption.

Let $\boldsymbol{y}$ denote the observations, $y_i = f(\boldsymbol{x}^{(i,n)})$. The Kriging prediction with the mean structure and covariance K as described in the proposition is $\hat{y}(\boldsymbol{x}) = \boldsymbol{\lambda}^t \boldsymbol{y}$, with (see (2.15)),

$$\begin{aligned} \boldsymbol{\lambda} &= (\boldsymbol{h}(\boldsymbol{x}))^t (\mathbf{H}^t \mathbf{K}_{obs}^{-1} \mathbf{H})^{-1} \mathbf{H}^t \mathbf{K}_{obs}^{-1} \\ &\quad + (\boldsymbol{r}(\boldsymbol{x}))^t \left(\mathbf{K}_{obs}^{-1} - \mathbf{K}_{obs}^{-1} \mathbf{H} (\mathbf{H}^t \mathbf{K}_{obs}^{-1} \mathbf{H})^{-1} \mathbf{H}^t \mathbf{K}_{obs}^{-1} \right), \end{aligned} \tag{4.1}$$

where $\mathbf{K}_{obs} = (\mathbf{K} + \sigma_{mes}^2 \mathbf{I}_n)^{-1}$.

The proof is based on considering first the abstract case of a Gaussian process Y with the mean structure and covariance K as described in the proposition. This will enable us to derive a property of the $\boldsymbol{\lambda}$ vector sequence only, that can be used also in the case of the proposition, where the observations stem from a smooth function.

Thus, consider a Gaussian process Y with the mean structure and covariance K as described in the proposition. We have shown in proposition 4.4 that the Kriging prediction is consistent and

$$\begin{aligned} 0 &= \lim_{n \to +\infty} \mathbb{E}\left(\left((Y(\boldsymbol{x}) - \boldsymbol{\lambda}^t \boldsymbol{y})^2 \right)^2 \right) \\ &= \lim_{n \to +\infty} \mathbb{E}\left(\left((Y(\boldsymbol{x}) - \sum_{i=1}^n \lambda_i Y(\boldsymbol{x}^{(i,n)}) - \sum_{i=1}^n \lambda_i \epsilon_i \right)^2 \right) \\ &= \lim_{n \to +\infty} \mathbb{E}\left(\left((Y(\boldsymbol{x}) - \sum_{i=1}^n \lambda_i Y(\boldsymbol{x}^{(i,n)}) \right)^2 \right) + \sigma_{mes}^2 \sum_{i=1}^n \lambda_i^2. \end{aligned}$$

Let $\lambda_0 = 1$ and $\boldsymbol{x}^{(0,n)} = \boldsymbol{x}$. We have

$$\begin{aligned} 0 &= \lim_{n \to +\infty} \mathbb{E}\left(\left((\lambda_0 Y(\boldsymbol{x}) - \sum_{i=1}^n \lambda_i Y(\boldsymbol{x}^{(i,n)}) \right)^2 \right) \\ &= \lim_{n \to +\infty} \sum_{i=0}^n \sum_{j=0}^n \lambda_i \lambda_j K(\boldsymbol{x}^{(i,n)}, \boldsymbol{x}^{(j,n)}) \\ &= \lim_{n \to +\infty} \int_{\mathbb{R}^d} \hat{K}(\boldsymbol{\omega}) \left| \sum_{i=0}^n \lambda_i e^{\mathrm{i}\boldsymbol{\omega}.\boldsymbol{x}^{(i,n)}} \right|^2 d\boldsymbol{\omega}. \end{aligned}$$

Let us now adopt a distribution framework. We consider $F_{\boldsymbol{\lambda}}$ as the distribution $\sum_{i=0}^n \lambda_i \delta_{\boldsymbol{x}^{(i,n)}}$ with $\delta_{\boldsymbol{x}^{(i,n)}}$ the Dirac distribution at $\boldsymbol{x}^{(i,n)}$. Then, in the distribution sense, $\hat{F}_{\boldsymbol{\lambda}}(\boldsymbol{\omega}) = \sum_{i=0}^n \lambda_i$ $e^{\mathrm{i}\boldsymbol{\omega}.\boldsymbol{x}^{(i,n)}}$. Hence, we have

$$\int_{\mathbb{R}^d} \hat{K}(\boldsymbol{\omega}) |\hat{F}_{\boldsymbol{\lambda}}(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \to_{n \to +\infty} 0.$$

Now, consider a rapidly decaying test function g, that is a C^∞ function so that, for any $k > 0$, $|g(\boldsymbol{\omega})||\boldsymbol{\omega}|^k \to_{|\boldsymbol{\omega}| \to +\infty} 0$. Then, there exists $C > 0$ so that $\hat{K}(\boldsymbol{\omega}) \geq C|g(\boldsymbol{\omega})|$ for any $\boldsymbol{\omega}$. Hence, we have shown that for any rapidly decaying test function

$$\int_{\mathbb{R}^d} |g(\boldsymbol{\omega})||\hat{F}_{\boldsymbol{\lambda}}(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \to_{n \to +\infty} 0,$$

so that (using Cauchy Schwartz) for any rapidly decaying test function $\tilde{g}$,

$$\int_{\mathbb{R}^d} \tilde{g}(\boldsymbol{\omega})\hat{F}_{\boldsymbol{\lambda}}(\boldsymbol{\omega})d\boldsymbol{\omega} \to_{n\to+\infty} 0.$$

Hence, because of a continuity theorem for the inverse of the Fourier transform ([Zem65] p187), we have, for any test function g,

$$\int_{\mathbb{R}^d} g(\boldsymbol{x})F_{\boldsymbol{\lambda}}(\boldsymbol{x})d\boldsymbol{x} \to_{n\to+\infty} 0. \tag{4.2}$$

The relation (4.2) only depends on the $\boldsymbol{\lambda}$ sequence, and not on the values of the abstract Gaussian process Y. It can hence be used in the case in which the observations are $y_i = f(\boldsymbol{x}^{(i,n)})$. We consider this case in the rest of the proof.

The true function f, being infinitely differentiable and defined on the compact $\mathcal{D}$, is a rapidly decaying test function. Hence,

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})F_{\boldsymbol{\lambda}}(\boldsymbol{x})d\boldsymbol{x} \to_{n\to+\infty} 0,$$

which is exactly

$$\sum_{i=0}^{n} \lambda_i f(\boldsymbol{x}^{(i,n)}) \to_{n\to+\infty} 0,$$

so that

$$\hat{y}(\boldsymbol{x}) \to_{n\to+\infty} f(\boldsymbol{x}).$$

□

As we have said in the proof above, proposition 4.5 is proved only for smooth functions f. As noted in [VB10], the generalization of proposition 4.5 to continuous functions f, proposed in [YS85], is not valid. The question of this generalization is of strong practical interest, since many simple prediction methods (e.g., a nearest neighbor method) are consistent for predicting continuous functions. To the best of our knowledge this question remains an open problem. In [VB10], an equivalent formulation of it is given, in term of the Lebesgue constant, but the equivalent formulation is unsolved either.

In proposition 4.5, note that the Kriging prediction is consistent even if a measurement error is inappropriately assumed.

In proposition 4.5, note also the important condition $\hat{K}(\boldsymbol{\omega})|\boldsymbol{\omega}|^k \to +\infty$. This means that the assumed Gaussian process is not infinitely differentiable (proposition 2.21). Looking at the Matérn model of subsection 2.1.2, the covariance functions of this model verify proposition 4.5 for finite smoothness parameter ν. However the Gaussian covariance function ($\nu = +\infty$) does not verify proposition 4.5. We are not aware of results in the literature on the consistency of Kriging with a Gaussian covariance function, with a dense sequence of observation points on a bounded domain and when a smooth function is predicted.

The Gaussian covariance function gives a.s analytic trajectories, so that, for instance, the associated Gaussian process Y on $[0, 1]$ can be predicted exactly from observing Y only on $[0, \epsilon]$ with $\epsilon > 0$ ([Ste99], p30). Similarly, it is shown in [VB10] that the Gaussian covariance function can yield a conditional variance going to zero, when there exists a positive minimum distance between the prediction points and all the observation points. These two facts may seem counter

intuitive when applying Kriging models in practical situations. Hence, it is recommended in several references (e.g. [Ste99]) not to use the Gaussian covariance function. An alternative to the Gaussian covariance function is the Matérn covariance function, whose smoothness parameter can be estimated from data.

In proposition 4.6, the case of a smooth function observed with measurement errors is addressed. Kriging is consistent in this case when the prediction incorporates an assumed measurement error with positive variance. The condition that the covariance function is not infinitely differentiable remains present.

Proposition 4.6. *Consider the universal Kriging framework with a Gaussian process Y on a compact $\mathcal{D} \subset \mathbb{R}^d$, with fixed continuous mean structure and continuous stationary covariance function K. Assume that there exists $k < +\infty$ so that the Fourier transform $\hat{K}$ of K is positive-valued and verifies $\hat{K}(\boldsymbol{\omega})|\boldsymbol{\omega}|^k \to +\infty$ when $|\boldsymbol{\omega}| \to +\infty$. Consider a fixed point $\boldsymbol{x} \in \mathcal{D}$. Assume that an infinitely differentiable function f is observed at $\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)} \in \mathcal{D}$, with observed values $y_{i,n} = f(\boldsymbol{x}^{(i,n)}) + \epsilon_{i,n}$, for $1 \leq i \leq n$, where the $\epsilon_{i,n}$ are iid and follow a $\mathcal{N}(0, \sigma^2_{mes,1})$ distribution. Assume that, for any open ball with center $\boldsymbol{x}$ and positive radius, the number of points in $\{\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}\}$ belonging to the ball goes to $+\infty$ when n goes to $+\infty$. Let $\hat{y}(\boldsymbol{x})$ be the Kriging prediction (2.15) of $f(\boldsymbol{x})$ with observations at $\boldsymbol{x}^{(1,n)}, ..., \boldsymbol{x}^{(n,n)}$, where an iid Gaussian measurement error is assumed, with mean zero and variance $\sigma^2_{mes,2} > 0$. Then, as $n \to +\infty$, $\hat{y}(\boldsymbol{x})$ goes to $f(\boldsymbol{x})$ in the mean square sense (w.r.t the measurement errors $\epsilon_{i,n}$ of the true function).*

Proof. The Kriging prediction is $\hat{y}(\boldsymbol{x}) = \boldsymbol{\lambda}^t \boldsymbol{y}$ where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)^t$ and $y_i = f(\boldsymbol{x}^{(i,n)}) + \epsilon_{i,n}$. In the proof of proposition 4.5, we have shown that, under assumed covariance function K and assumed measurement error variance $\sigma^2_{mes,2}$,

$$\sigma^2_{mes,2} \sum_{i=1}^{n} \lambda_i^2 \to_{n \to +\infty} 0. \tag{4.3}$$

Furthermore, under the distribution of the true measurement error, with variance $\sigma^2_{mes,1}$, we get

$$\mathbb{E}\left((f(\boldsymbol{x}) - \hat{y}(\boldsymbol{x}))^2 \right) = \left(f(\boldsymbol{x}) - \sum_{i=1}^{n} \lambda_i f(\boldsymbol{x}^{(i,n)}) \right)^2 + \sigma^2_{mes,1} \sum_{i=1}^{n} \lambda_i^2.$$

From (4.3), since $\sigma_{mes,2} > 0$, $\sum_{i=1}^{n} \lambda_i^2 \to 0$. We have seen in the proof of proposition 4.5 that $\sum_{i=1}^{n} \lambda_i f(\boldsymbol{x}^{(i,n)}) \to f(\boldsymbol{x})$. This concludes the proof. □

Looking at propositions 4.5 and 4.6, we see that the Kriging prediction with an assumed measurement error is consistent, whether or not the observations of the smooth function actually come with measurement errors. On the contrary, Kriging prediction without assumed measurement error would not be consistent in the case where the observations of the smooth function would come with measurement errors.

To see this, consider the prediction of $f(1)$, based on noisy observations at $\{f(\frac{i}{n}), 0 \leq i \leq n-1\}$, using an exponential covariance function with $\sigma^2 = 1$ and $\ell = 1$. A Gaussian process Y on $\mathbb{R}$, with the exponential covariance function, is a Markov process: for any $x_1 < ... <$

$x_n < x$, $\mathcal{L}(Y(x)|Y(x_1),...,Y(x_n)) = \mathcal{L}(Y(x)|Y(x_n))$ (see e.g. [Yin91]). Hence, $f(1)$ would be inconsistently predicted, using the single observation at $\frac{n-1}{n}$, by

$$\hat{y}(1) = e^{-\frac{1}{n}} y_n,$$

with $y_n = Y(\frac{n-1}{n}) + \epsilon_n$, with ϵ_n the measurement error at x_n.

The discussion above is an argument in favor of systematically incorporating a positive nugget effect (we talk of numerical nugget effect) in the Kriging prediction (2.15).

Finally, in this subsection 4.2.1, we have addressed consistency qualitatively. Quantitative results for a rate of convergence of the Kriging prediction do not exist, to the best of our knowledge, in a general framework when Y is observed exactly, even when the prediction is done with the true distribution of Y. In the case of measurement errors, [GG12] recently provided results for the rate of convergence of Kriging prediction, with the true distribution of Y.

4.2.2 Asymptotic influence of a misspecified covariance function

In the previous subsection 4.2.1, we have studied the consistency of the Kriging predictions, with a well-specified or ill-specified Gaussian process model.

When the Gaussian process model is ill-specified, but the observations still stem from a Gaussian process, with a different covariance function, we have seen than the question of Kriging-prediction consistency is still open. Another relevant question in this case is also: if the Kriging prediction is consistent, can we quantify the loss compared to the prediction with the correct Gaussian process model. The present subsection 4.2.2 gives some elements on this question. The asymptotic framework followed is the fixed-domain asymptotic framework.

Orthogonal and equivalent Gaussian measures

In this subsection 4.2.2, we consider a Gaussian process Y, on a compact set $\mathcal{D} \subset \mathbb{R}^d$. Y has mean function m_1 and has covariance function K_1. We assume that the Kriging predictions are carried out with the Kriging formulas obtained when considering that Y is a Gaussian process with mean function m_2 and covariance function K_2.

To compare (m_1, K_1) and (m_2, K_2), we first define the two Gaussian measures yielded by (m_1, K_1) and (m_2, K_2) in definition 4.7.

Definition 4.7. *Consider a measurable space $(\Omega, \mathcal{F})$, equipped with two probability measures P_1 and P_2. Assume that there exist two stochastic processes Y_1 and Y_2 on $\mathcal{D}$, where Y_i has probability space $(\Omega, \mathcal{F}, P_i)$. Assume also that Y_i is a Gaussian process with mean function m_i and covariance function K_i. Then P_1 and P_2 are called two Gaussian measures yielded by (m_1, K_1) and (m_2, K_2).*

Remark 4.8. *In definition 4.7, consider two mean and covariance functions (m_1, K_1) and (m_2, K_2). In order to define two Gaussian measures they yield, it is necessary to define two Gaussian processes Y_1, Y_2 which have the same measurable space $(\Omega, \mathcal{F})$ with two different probability measures. This is in fact always possible, because the probability space of a stochastic process on $\mathcal{D}$ can always be defined as $(\tilde{\Omega}, \tilde{\mathcal{F}}, P)$, where $\tilde{\Omega}$ and $\tilde{\mathcal{F}}$ only depend on $\mathcal{D}$ and P only*

depends on the finite-dimensional distributions of the stochastic process. We refer to e.g. chapter 2.1.1 of [Vaz05] or chapter I.2 of [IR78] for details on this point.

From definition 4.7, we see that we can compare pairs (m_1, K_1) and (m_2, K_2) of mean and covariance functions, for a Gaussian process Y, by comparing the two probability measures P_1 and P_2 that they yield on the abstract probability space Ω.

The criterion for comparing P_1 and P_2, used in [Ste99], is the criterion of their equivalence or their orthogonality, as presented in the following definition.

Definition 4.9. *Consider the framework of definition 4.7. P_1 and P_2 are equivalent if, for any $E \subset \mathcal{F}$, $P_1(E) = 0$ if and only if $P_2(E) = 0$. P_1 and P_2 are orthogonal if there exists $E \subset \mathcal{F}$, so that $P_1(E) = 0$ and $P_2(E) = 1$.*

It is shown in [Ste99] p.117, following [IR78], p.74-77, that two Gaussian measures yielded by two pairs (m_1, K_1) and (m_2, K_2) are either equivalent or orthogonal. This is stated in the following proposition.

Proposition 4.10. *Consider two pairs of mean and covariance functions (m_i, K_i), $i = 1, 2$ for a Gaussian process Y on a compact set $\mathcal{D} \subset \mathbb{R}^d$. Assume that, for $i = 1, 2$, the mean function m_i is continuous on $\mathcal{D}$ and that the covariance function K_i is continuous and positive definite on $\mathcal{D} \times \mathcal{D}$. Then, in the context of definition 4.7, the two measures P_1 and P_2 are either equivalent or orthogonal.*

In subsection 4.3.2, we give some explicit relations between the covariance hyper-parameters in the Matérn family of subsection 2.1.2 and the equivalence or orthogonality of the obtained covariance functions.

Proposition 4.10 shows that we can compare Gaussian process measures in a binary way, because they are either equivalent or orthogonal. We will see that this binary distinction has a great impact, for Kriging prediction in the sequel of subsection 4.2.2, and for covariance function estimation in subsection 4.3.2 .

We will hence start by considering Kriging prediction and showing that if (m_1, K_1) and (m_2, K_2) are equivalent, then there is asymptotically no loss in using incorrectly (m_2, K_2) for prediction.

Case of a misspecified but equivalent Gaussian measure

We consider now the case when P_1 and P_2 of definition 4.7 are equivalent. We will describe the results in [Ste88, Ste90a, Ste90c], stating that, in the equivalence case, there is asymptotically no loss using the incorrect pair (m_2, K_2) compared to using the correct pair (m_1, K_1). The asymptotic optimality concerns predictions as well as correct assessments of prediction errors.

This result was first shown for a fixed predictand point in [Ste88], theorems 1 and 2. The following theorem directly follows from the reference hereabove.

Theorem 4.11. *Consider a dense sequence of observation points $(\boldsymbol{x}^{(i)})_{i \in \mathbb{N}^*}$, in the compact set $\mathcal{D} \subset \mathbb{R}^d$, where the Gaussian process Y is observed exactly. Let E_i, $\hat{y}_i(\boldsymbol{x})$ and $\hat{\sigma}_i^2(\boldsymbol{x})$ be the mean value, the prediction* (2.9) *and the predictive variance* (2.10) *under Gaussian process*

structure (m_i, K_i), $i = 1, 2$, *for* Y. *Assume that* (m_i, K_i), $i = 1, 2$, *are continuous and that* P_1 *and* P_2 *(of definition 4.7) are equivalent. Then, with* $\boldsymbol{x}$ *a fixed predictand in* $\mathcal{D}$, *different from the* $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$,

$$\frac{\mathbb{E}_1\left(\left(\hat{y}_2(\boldsymbol{x}) - Y(\boldsymbol{x})\right)^2\right)}{\mathbb{E}_1\left(\left(\hat{y}_1(\boldsymbol{x}) - Y(\boldsymbol{x})\right)^2\right)} \to_{n\to+\infty} 1 \tag{4.4}$$

and

$$\frac{\hat{\sigma}_2^2(\boldsymbol{x})}{\mathbb{E}_1\left(\left(\hat{y}_2(\boldsymbol{x}) - Y(\boldsymbol{x})\right)^2\right)} \to_{n\to+\infty} 1. \tag{4.5}$$

In theorem 4.11, the ratio in (4.4) is the ratio of the MSE of the sub-optimal prediction given by (m_2, K_2) on the MSE of the optimal prediction given by (m_1, K_1). It is larger than 1, and the fact that it goes to 1 is the mathematical translation of the sentence "no asymptotic loss for the prediction MSE in using the incorrect mean and covariance functions".

The ratio in (4.5) is the ratio, for the sub-optimal prediction given by (m_2, K_2), of the incorrect estimation of its MSE given by (m_2, K_2) on its true MSE given by (m_1, K_1). The fact that it goes to 1 is the mathematical translation of the sentence "asymptotically correct assessment of prediction errors".

The following theorem, obtained by [Ste90c], gives a uniform version of theorem 4.11.

Theorem 4.12. *Consider a dense sequence of observation points* $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$, *in the compact set* $\mathcal{D} \subset \mathbb{R}^d$, *where the Gaussian process* Y *is observed exactly. Let* E_i, $\hat{y}_i(\boldsymbol{x})$ *and* $\hat{\sigma}_i^2(\boldsymbol{x})$ *be the mean value, the prediction and the predictive variance under Gaussian process structure* (m_i, K_i), $i = 1, 2$, *for* Y. *Assume that* (m_i, K_i), $i = 1, 2$, *are continuous and that* P_1 *and* P_2 *(of definition 4.7) are equivalent. Define, for* $i = 1, 2$, H_i *as the Hilbert space equal to the adherence of the linear span of the* $Y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$, *with the norm induced by the dot product* $z_1, z_2 \to \mathbb{E}_i(z_1, z_2)$. *Then the Hilbert spaces* H_1 *and* H_2 *are the same and are denoted* H. *Furthermore, letting* $\hat{y}_i(h)$ *and* $\hat{\sigma}_i^2(h)$ *be the predictions* (2.9) *and predictive variances* (2.10), *from* $(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))$, *of each random variable* $h \in H$, *under mean and covariance function* m_i, K_i, *we have*

$$\sup_{h\in H} \frac{\mathbb{E}_1\left(\left(\hat{y}_2(h) - h\right)^2\right) - \mathbb{E}_1\left(\left(\hat{y}_1(h) - h\right)^2\right)}{\mathbb{E}_1\left(\left(\hat{y}_1(h) - h\right)^2\right)} \to_{n\to+\infty} 0 \tag{4.6}$$

and

$$\sup_{h\in H} \frac{\left|\hat{\sigma}_2^2(h) - \mathbb{E}_1\left(\left(\hat{y}_2(h) - h\right)^2\right)\right|}{\mathbb{E}_1\left(\left(\hat{y}_2(h) - h\right)^2\right)} \to_{n\to+\infty} 0, \tag{4.7}$$

with the convention $\frac{0}{0} = 0$.

Remark 4.13. *In theorem 4.12, the Hilbert space* H *is basically composed of all the linear functionals of* Y. *It hence includes all the random variables* $Y(\boldsymbol{x})$, *but also integral terms like* $\int_{\mathcal{D}} Y(\boldsymbol{x})d\boldsymbol{x}$, *or, in the case where* Y *is mean-square differentiable, derivative terms like* $\frac{\partial Y(\boldsymbol{x})}{\partial x_i}$.

In theorem 4.12, as in theorem 4.11, the ratios (4.4) and (4.6) correspond to the asymptotic optimality (in terms of MSE) of the prediction using incorrectly (m_1, K_1) and the ratios (4.5) and (4.7) correspond to the asymptotically correct assessment of the predictive variance. Theorem 4.11 can be seen as a particular case of theorem 4.12. Theorem 4.12 shows that the asymptotic

optimality at $\boldsymbol{x}$ is actually uniform over all $\boldsymbol{x} \in \mathcal{D}$. Furthermore, the uniformity also holds for the prediction of all linear functionals of Y, such as $\int_{\mathcal{D}} Y(\boldsymbol{x})d\boldsymbol{x}$.

Let us also mention that the analysis of bounds for the asymptotic optimality in theorems 4.11 and 4.12 is performed in [Ste90a, Ste90c].

Finally, let us mention that, in theorems 4.11 and 4.12, it is assumed that there is a single sequence $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$ of observation points. This excludes some cases when the sets of n observation points, $n \in \mathbb{N}^*$, are not part of a single sequence $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$. Consider for instance observing Y, at step n, on $\{\frac{i}{n}, i \in \{1, ..., n\}\}$. [Ste99], p.132, theorem 10 gives the theorem corresponding to theorem 4.12 in this case. [Ste99], p.132, theorem 10 also considers observing Y on non-numerable subsets of $\mathcal{D}$. For instance the theorem applies to the case, in dimension 1, of the prediction of $Y(1)$ from $\{Y(t), 0 \leq t \leq 1-\epsilon\}$, for $\epsilon > 0$ ([Ste99], p132).

In [Vaz05], numerical illustrations of theorems 4.11 and 4.12 are presented. It is confirmed numerically that using misspecified but equivalent mean and covariance functions results in almost optimal predictions when the number of observation points is large compared to the dimension. However, in the complementary case where the number of points is not large compared to the dimension, situations are presented where equivalent but misspecified mean and covariance functions yield considerably sub-optimal predictions. Hence, for moderate sample size, or for high-dimensional cases, the question of the choice of the mean and the covariance function goes beyond the question of equivalence or orthogonality.

A proof of theorem 4.12

We give a proof of theorem 4.12, that is also given in [Ste99] p 135. The objective is to give a pedagogical proof, highlighting that theorem 4.12 can be seen as a particular case of a general theorem treating asymptotic equivalence of conditional distributions. This is also underlined in [Ste99], p135.

The general theorem on asymptotic equivalence of conditional distributions is the main theorem in [BD62]. In the following theorem, we present an adaptation of this main theorem in the context of theorem 4.12.

Theorem 4.14. *Consider a compact set $\mathcal{D} \subset \mathbb{R}^d$. For $i = 1, 2$, let $(\Omega, \mathcal{F}, P_i)$ be the probability space associated to the Gaussian process Y on $\mathcal{D}$, with mean function m_i and covariance function K_i. Assume that $\mathcal{F}$ is the smallest sigma-algebra on Ω for which the $Y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$, are measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with $\mathbb{R}, \mathcal{B}(\mathbb{R})$ the Borel sigma-algebra on $\mathbb{R}$.*

Consider a dense sequence of observation points $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$, in $\mathcal{D}$, where Y is observed exactly.

Assume that (m_i, K_i), $i = 1, 2$, are continuous and that P_1 and P_2 are equivalent.

Let, for $i = 1, 2$, $P_{i|n}$ be the distribution P_i on $(\Omega, \mathcal{F})$, conditionally to $Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)})$. Let us define the distance between two distributions $\tilde{P}_1$ and $\tilde{P}_2$ on $(\Omega, \mathcal{F})$ by $|\tilde{P}_1 - \tilde{P}_2| = \sup_{F\in\mathcal{F}} |(\tilde{P}_1(F) - \tilde{P}_2(F)|$. Then, P_1-almost surely,

$$|P_{1|n} - P_{2|n}|$$

goes to zero when $n \to +\infty$.

In theorem 4.14, $P_{i|n}$ is interpreted as the conditional distribution of Y, when $Y(\boldsymbol{x}^{(1)}), ...,$ $Y(\boldsymbol{x}^{(n)})$ are fixed (like in figure 2.8), under mean and covariance functions m_i, K_i.

We recall that, for $i = 1, 2$, H_i is the Hilbert space equal to the adherence of the linear span of the $Y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$, with the norm induced by the dot product $z_1, z_2 \to \mathbb{E}_i(z_1, z_2)$ and that the Hilbert spaces H_1 and H_2 are the same and are denoted H. All the random variables in H are measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Hence, the consequence of theorem 4.14 is that all the elements of the Hilbert space H have asymptotically the same conditional distribution under P_1 and P_2. Let $\tilde{\mathcal{L}}_1$ and $\tilde{\mathcal{L}}_2$ be two distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and define their distance as $|\tilde{\mathcal{L}}_1 - \tilde{\mathcal{L}}_2| = \sup_{A \in \mathcal{B}(\mathbb{R})} |\tilde{\mathcal{L}}_1(A) - \tilde{\mathcal{L}}_2(A)|$, where $\mathcal{B}(\mathbb{R})$ is the Borel sigma-algebra on $\mathbb{R}$.

Let, for $h \in H$, $\mathcal{L}^h_{i|n}$ be the distribution of h conditionally to $(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))$.

Then, we have, P_1-almost surely,

$$\sup_{h \in H} \left| \mathcal{L}^h_{1|n} - \mathcal{L}^h_{2|n} \right| \tag{4.8}$$

goes to zero when $n \to +\infty$.

If in (4.6) and (4.7), for some $h \in H$, one of the two denominators is zero, then, by the equivalence between P_1 and P_2, the two numerators are also zero. This is because, if $\hat{\sigma}^2_1(h) = 0$, then the event $\{\hat{y}_1(h) = h\}$ has probability 1 under P_1. Since P_1 and P_2 are equivalent, it has probability 1 under P_2. Since $\hat{y}_1(h)$ minimizes the prediction MSE under P_2 it also verifies $P_2(\{\hat{y}_2(h) = h\}) = 1$ and hence $P_1(\{\hat{y}_2(h) = h\}) = 1$, so that $\mathbb{E}_1((\hat{y}_2(h) - h)^2) = 0$. This is the same for addressing the case $\mathbb{E}_1((\hat{y}_2(h) - h)^2) = 0$.

We hence consider the case where in (4.6) and (4.7), $\hat{\sigma}^2_1(h) > 0$. In this case, we have shown $\hat{\sigma}^2_2(h) > 0$ and the last step of the proof is to show that (4.8) implies (4.6) and (4.7). This is done using the two following lemmas, which are proved below.

Lemma 4.15. *Let Φ_{m,σ^2} be the Gaussian cumulative distribution function on $\mathbb{R}$ with mean m and variance σ^2. Then there exists $0 < K < +\infty$ and $0 < \epsilon < +\infty$ so that, for $\sigma^2_1 > 0$,*

$$\sup_{t \in \mathbb{R}} \left(\Phi_{m_1,\sigma^2_1}(t) - \Phi_{m_2,\sigma^2_2}(t) \right)^2 \geq min \left(\epsilon, K \max \left(\frac{|m_1 - m_2|}{\sigma_1}, \frac{|\sigma_1 - \sigma_2|}{\sigma_1} \right)^2 \right).$$

Lemma 4.16. *Consider a family of sequences $(X_{h,n})_{n \in \mathbb{N}^*, h \in H}$ of real-valued Gaussian variables. If for all $0 < t < +\infty$, $\sup_{h \in H} P(|X_{h,n}| \geq t) \to_{n \to +\infty} 0$ then $\sup_{h \in H} \mathbb{E}(X^2_{h,n}) \to_{n \to +\infty} 0$.*

Now, with $\tilde{H} := \{h \in H | \hat{\sigma}^2_1(h) > 0\}$ (4.8) implies

$$\sup_{h \in \tilde{H}} \sup_{t \in \mathbb{R}} \left| \Phi_{\hat{y}_1(h), \hat{\sigma}^2_1(h)}(t) - \Phi_{\hat{y}_2(h), \hat{\sigma}^2_2(h)}(t) \right| \tag{4.9}$$

goes to zero P_1-almost surely.

Using lemma 4.15, we obtain that

$$\sup_{h \in \tilde{H}} \frac{|\hat{\sigma}_1(h) - \hat{\sigma}_2(h)|}{\hat{\sigma}_1(h)} \tag{4.10}$$

goes P_1-almost surely to zero. Since it is actually non random (because is is composed of two Gaussian conditional standard deviations, see (2.10)), it goes to zero.

Now, using lemma 4.15 again, we obtain that

$$\sup_{h \in \tilde{H}} \left(\frac{\hat{y}_1(h)}{\hat{\sigma}_1(h)} - \frac{\hat{y}_2(h)}{\hat{\sigma}_1(h)} \right)^2 \tag{4.11}$$

goes to zero P_1-almost-surely. From, (4.11), we obtain, for all $t > 0$

$$\sup_{h \in \tilde{H}} P_1 \left(\left(\frac{\hat{y}_1(h)}{\hat{\sigma}_1(h)} - \frac{\hat{y}_2(h)}{\hat{\sigma}_1(h)} \right)^2 \geq t \right) \tag{4.12}$$

goes to zero.

Finally, the variables $\left(\frac{\hat{y}_1(h)}{\hat{\sigma}_1(h)} - \frac{\hat{y}_2(h)}{\hat{\sigma}_1(h)} \right)$, for $h \in H$ so that $\hat{\sigma}_1^2(h) > 0$, are Gaussian, so that applying lemma 4.16 to them yields

$$\sup_{h \in \tilde{H}} \mathbb{E}_1 \left(\left(\frac{\hat{y}_1(h)}{\hat{\sigma}_1(h)} - \frac{\hat{y}_2(h)}{\hat{\sigma}_1(h)} \right)^2 \right), \tag{4.13}$$

goes to 0.

Hence using the classical bias variance decomposition $\mathbb{E}_1((\hat{y}_2(h) - h)^2) = \hat{\sigma}_1^2(h) + \mathbb{E}_1((\hat{y}_2(h) - \hat{y}_1(h))^2)$ with (4.13) shows (4.6). Using again $\mathbb{E}_1((\hat{y_2}(h) - h)^2) = \hat{\sigma}_1^2(h) + \mathbb{E}_1((\hat{y_2}(h) - \hat{y}_1(h))^2)$, together with (4.13) and (4.10), shows (4.7).

Proof of lemma 4.15

$$\begin{aligned} \sup_{t \in \mathbb{R}} \left(\Phi_{m_1, \sigma_1^2}(t) - \Phi_{m_2, \sigma_2^2}(t) \right)^2 &= \sup_{t \in \mathbb{R}} \left(\Phi_{0, \sigma_1^2}(t) - \Phi_{m_2 - m_1, \sigma_2^2}(t) \right)^2 \\ &= \sup_{t \in \mathbb{R}} \left(\Phi_{0,1}(t) - \Phi_{\frac{m_2 - m_1}{\sigma_1}, \frac{\sigma_2^2}{\sigma_1^2}}(t) \right)^2 \end{aligned}$$

Let $m = \frac{m_2 - m_1}{\sigma_1}$ and $\sigma^2 = \frac{\sigma_2^2}{\sigma_1^2}$. Then,

$$\sup_{t \in \mathbb{R}} \left(\Phi_{0,1}(t) - \Phi_{m, \sigma^2}(t) \right)^2 \geq \frac{1}{2} \int_{-1}^{1} \left(\Phi_{0,1}(t) - \Phi_{m, \sigma^2}(t) \right)^2 dt \tag{4.14}$$

Now, from the dominated convergence theorem, the bivariate function $(m, \sigma) \to f(m, \sigma) := \int_{-1}^{1} \left(\Phi_{0,1}(t) - \Phi_{m, \sigma^2}(t) \right)^2 dt$ is twice differentiable. Since it is non-negative, it has a zero gradient at $(m = 0, \sigma = 1)$. From the identifiability of the Gaussian model on $\mathbb{R}$, parameterized by (m, σ), $f(m, \sigma)$ has a positive Hessian matrix at $(m = 0, \sigma = 1)$. Hence, with K being $\frac{1}{2}$ times the smallest eigenvalue of this Hessian, we have, for a positive ϵ

$$f(m, \sigma) \geq K(m^2 + (\sigma - 1)^2) \text{ for } m^2 + (\sigma - 1)^2 \leq \epsilon.$$

From the identifiability of the Gaussian model, there exists $\alpha > 0$ so that, for $m^2 + (\sigma - 1)^2 \geq \epsilon$, $f(m, \sigma) \geq \alpha$.

Hence, finally, we have, from (4.14),

$$\sup_{t \in \mathbb{R}} \left(\Phi_{0,1}(t) - \Phi_{m, \sigma^2}(t) \right)^2 \geq min(\frac{\alpha}{2}, \frac{K}{2}(m^2 + (\sigma - 1)^2)),$$

which completes the proof of the lemma.

Proof of lemma 4.16 Assume that there exist $\epsilon > 0$ and h_n so that for all n, $\mathbb{E}(X_{h_n, n}^2) \geq \epsilon$. Then $(X_{h_n, n})_{n \in \mathbb{N}^*}$ is a sequence of Gaussian variables that goes to zero in probability but that does not go to zero in the mean square sense. This is in contradiction with lemma 1 of [IR78].

4.3 Asymptotic results for Maximum Likelihood

The goal of this section 4.3 is to give some existing results regarding the consistency and asymptotic distribution of the ML estimator of subsection 3.2.2. We consider the two asymptotic frameworks of subsection 4.1: fixed-domain and increasing-domain asymptotics. Roughly speaking, we will see that ML is generally consistent with asymptotic normality in the increasing-domain asymptotic framework. In the fixed-domain asymptotic framework, we will see that some hyper-parameters, defining the covariance function within a parametric family (subsection 3.2.1), can be consistently estimated, while it is proved that others can not be consistently estimated. For the hyper-parameters that can be consistently estimated we will consider the cases of ML and also of other estimators.

4.3.1 Expansion-domain asymptotic results

Consistency and asymptotic normality for ML in Kriging have been proved by [MM84]. The proof in [MM84] is based on [Swe80], which gives a general sufficient condition for the consistency and asymptotic normality of ML, based on continuity, growth and convergence conditions on the random Fisher information matrix. Notably, it is not assumed in [Swe80] that the observations are independent, which explains why the results are applicable in the Kriging framework.

Theorem 2 of [MM84] gives general conditions that imply the conditions in [Swe80], and are therefore sufficient conditions for consistency and asymptotic normality for ML in Kriging. In theorem 4.17, we state these sufficient conditions.

Theorem 4.17. *Consider ML in an universal Kriging case (subsection 3.2.2). Let $\boldsymbol{\beta} \in \mathbb{R}^m$ be the mean parameter and $\boldsymbol{\psi} \in \Psi \subset \mathbb{R}^p$ be the covariance hyper-parameter. Let $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ be the correct mean and covariance parameters. Consider a sequence of observation points $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$. For each n, let $\mathbf{K}_{\boldsymbol{\psi}}$ be the covariance matrix under covariance function $K_{\boldsymbol{\psi}}$ and $\mathbf{H}$ be the regression matrix. Assume that $\boldsymbol{\psi} \to K_{\boldsymbol{\psi}}$ is twice differentiable.*

The parameters to be estimated are $\boldsymbol{\beta}, \boldsymbol{\psi}$ and the $(m+p)\times(m+p)$ Fisher information matrix (proposition 3.15), denoted by $\mathcal{I}_n$, is thus defined by, with $l(\boldsymbol{\beta}, \boldsymbol{\psi})$ the Gaussian likelihood function at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$,

$$\begin{pmatrix} -\mathbb{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}\ln\left(l(\boldsymbol{\psi}^{(0)},\boldsymbol{\beta}^{(0)})\right)\right\}, & -\mathbb{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\beta}}\ln\left(l(\boldsymbol{\psi}^{(0)},\boldsymbol{\beta}^{(0)})\right)\right\} \\ -\mathbb{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\psi}}\ln\left(l(\boldsymbol{\psi}^{(0)},\boldsymbol{\beta}^{(0)})\right)\right\}, & -\mathbb{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}}\ln\left(l(\boldsymbol{\psi}^{(0)},\boldsymbol{\beta}^{(0)})\right)\right\} \end{pmatrix} := \begin{pmatrix} \mathcal{I}_{\boldsymbol{\beta}} & \mathcal{I}_{\boldsymbol{\beta},\boldsymbol{\psi}} \\ \mathcal{I}_{\boldsymbol{\psi},\boldsymbol{\beta}} & \mathcal{I}_{\boldsymbol{\psi}} \end{pmatrix}.$$

Then, the $m\times p$ matrix $\mathcal{I}_{\boldsymbol{\beta},\boldsymbol{\psi}}$ is the zero matrix, the $p\times m$ matrix $\mathcal{I}_{\boldsymbol{\psi},\boldsymbol{\beta}}$ is the zero matrix, the $m\times m$ matrix $\mathcal{I}_{\boldsymbol{\beta}}$ is $\mathbf{H}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{H}$ and the $p\times p$ matrix $\mathcal{I}_{\boldsymbol{\psi}}$ has (i,j)-th term equal to $\frac{1}{2}Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\right)$.

Assume the following, as $n\to+\infty$ and for all $\boldsymbol{\psi}\in\Psi$.

i) For $1\le i,j\le p$, the largest (in absolute value) eigenvalues of the matrices $\mathbf{K}_{\boldsymbol{\psi}}$, $\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}$ and $\frac{\partial^2\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i\partial\psi_j}$ converge to finite constants.

ii) There exist $\delta>0$ and $A>0$ so that $\sum_{k,l=1}^n\left(\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\right)_{k,l}^2 \ge An^{\frac{1}{2}+\delta}$.

iii) The $p\times p$ matrix with term i,j equal to $\frac{\{\mathcal{I}_{\boldsymbol{\psi}}\}_{i,j}}{\sqrt{\{\mathcal{I}_{\boldsymbol{\psi}}\}_{i,i}\{\mathcal{I}_{\boldsymbol{\psi}}\}_{j,j}}}$ converges to a positive matrix.

iv) $(\mathbf{H}^t\mathbf{H})^{-1}$ *goes to the zero matrix.*

Then, $\hat{\boldsymbol{\beta}}_{ML}$ *and* $\hat{\boldsymbol{\psi}}_{ML}$ *go in probability to* $\boldsymbol{\beta}^{(0)}$ *and* $\boldsymbol{\psi}^{(0)}$*. Furthermore*

$$\mathcal{I}_n^{\frac{1}{2}} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}^{(0)} \\ \hat{\boldsymbol{\psi}}_{ML} - \boldsymbol{\psi}^{(0)} \end{pmatrix} \rightarrow_{\mathcal{L}} \mathcal{N}(0, \mathbf{I}_{m+p})$$

Theorem 4.17 is not meant yet to be applied directly in the increasing-domain asymptotic framework. Theorem 4.18, obtained from theorem 3 in [MM84], is. Nevertheless, we can already comment the four conditions $i), ..., iv)$ in theorem 4.17. Condition $i)$ means that the observations at the points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ are not too correlated, so that the information they give is not redundant. In the fixed-domain asymptotic framework, for instance, condition $i)$ would fail. Indeed, for sample for the Matérn model, for any $\boldsymbol{x}, \boldsymbol{x}'$ in a compact set, $K_{\boldsymbol{\psi}}(\boldsymbol{x} - \boldsymbol{x}')$ is larger in absolute value than a positive term depending only on the diameter of the compact set. As a result, all the elements of the matrix $\mathbf{K}_{\boldsymbol{\psi}}$ would be larger in absolute value than a positive constant, so that its largest eigenvalue would go to infinity. On the contrary, because classical covariance models verify $\left|\frac{\partial}{\partial \psi_i} K_{\boldsymbol{\psi}}(\boldsymbol{x} - \boldsymbol{x}')\right| \rightarrow_{|\boldsymbol{x}-\boldsymbol{x}'|\rightarrow+\infty} 0$, condition $ii)$ means that the observation points are not too far away from one another, so that they can still give information on the correlation structure. Condition $iii)$ and $iv)$ are identifiability assumptions for the covariance model, and for the regression model. In particular, when a minimum distance exists between two different observation points, as is classical in increasing-domain asymptotics, condition $iv)$ requires that the functions $h_j(\boldsymbol{x})$ of the regression model have unbounded supports.

We now present, in theorem 4.18, the theorem 3 in [MM84], which is dedicated to the most classical increasing-domain asymptotic framework: the case where the observation points form a regular lattice on $\mathbb{R}^d$.

Theorem 4.18. *Consider the framework of theorem 4.17.*

Assume that the covariance function family $\{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\}$ *is stationary.*

Assume that the observation point sequence $(\boldsymbol{x}^{(i)})_{i\in\mathbb{N}^*}$ *is so that, for all* $N \in \mathbb{N}^*$, $\{\boldsymbol{x}^{(i)}, 1 \leq i \leq N^d\} = \{i_1\boldsymbol{v}^{(1)} + ... + i_d\boldsymbol{v}^{(d)}, 1 \leq i_1, ..., i_d \leq N\}$*, for* d *linearly independent vectors* $\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(d)}$*.*

Let $\boldsymbol{\psi}^{(0)}$ *and* $\boldsymbol{\beta}^{(0)}$ *be the correct hyper-parameters. Assume that, for all* $\boldsymbol{\psi} \in \Psi$, $1 \leq i, j \leq d$*, the functions* $K_{\boldsymbol{\psi}}$, $\frac{\partial K_{\boldsymbol{\psi}}}{\partial \psi_i}$ *and* $\frac{\partial^2 K_{\boldsymbol{\psi}}}{\partial \psi_i \partial \psi_j}$ *are summable on the infinite regular lattice of the observation point sequence.*

Then, if conditions iii) and iv) of theorem 4.17 hold, the conclusion of theorem 4.17 holds.

Theorem 4.18 gives sufficient conditions, for consistency and asymptotic normality, for observation points forming an infinite regular lattice, that are verifiable in practice. Notably, the summability of $K_{\boldsymbol{\psi}}$, $\frac{\partial K_{\boldsymbol{\psi}}}{\partial \psi_i}$ and $\frac{\partial^2 K_{\boldsymbol{\psi}}}{\partial \psi_i \partial \psi_j}$ holds for the Matérn model of subsection 2.1.2. The only difficulty we find in practice is the identifiability condition $iii)$ of theorem 4.17 for the covariance function family. Indeed, this condition is defined in terms of the limit of a rather complex matrix.

In chapter 5, we give ourselves a consistency and asymptotic normality result for ML, in an increasing-domain asymptotic framework, where we study a regular lattice, and randomly perturbed regular lattices. We find that our main improvement, relatively to theorem 4.18, is to give an identifiability condition in terms of only the covariance function family, without requiring

to study the limit of matrix sequences. Note, also, that we show that the Fisher information matrix behaves asymptotically like n times a constant matrix, thus we explicit that we have a classical $\sqrt{n}$ rate of convergence for estimation.

Finally, note that the consistency and asymptotic normality for REML has been proved, in a framework similar to [MM84], in [CL93].

4.3.2 Fixed-domain asymptotic results

In the whole subsection 4.3.2, consider the fixed-domain asymptotic framework where the Gaussian process Y is considered on the compact set $\mathcal{D} \subset \mathbb{R}^d$.

To clarify the content, we consider that Y is centered, as it is generally done in [Ste99] chapter 6 on covariance function estimation, and in the references presented in this subsection 4.3.2.

Like in definition 3.18, we consider a covariance function model

$$\{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\},$$

where $K_{\boldsymbol{\psi}}$ is not necessarily stationary, unless specified otherwise.

Microergodic and non-microergodic covariance hyper-parameters

We have seen in subsection 4.2.2 that a fruitful way to compare two covariance functions, in the fixed-domain asymptotic framework, is to study the equivalence or orthogonality of the two Gaussian measures they yield.

The point of view of this equivalence or orthogonality has also a strong impact on estimation. Hence, it is useful to distinguish two kinds of hyper-parameters. Those that, when varying, yield orthogonal Gaussian measures are called microergodic hyper-parameters and those that, when varying, yield equivalent Gaussian measures are called non-microergodic hyper-parameters. These two notions (presented in [Ste99], p163) are detailed in the following definition.

Definition 4.19. *Consider a hyper-parameter $h(\boldsymbol{\psi})$ for $h : \Psi \to \mathbb{R}^{p'}$. This hyper-parameter is microergodic if, for all $\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)} \in \Psi$, $h(\boldsymbol{\psi}^{(1)}) \neq h(\boldsymbol{\psi}^{(2)})$ implies that the two Gaussian measures $P_{\boldsymbol{\psi}^{(1)}}$ and $P_{\boldsymbol{\psi}^{(2)}}$ (definition 4.7) are orthogonal. A hyper-parameter $h(\boldsymbol{\psi})$ is called non-microergodic if it is not microergodic.*

Remark 4.20. *We make a slight extension of the definition of a hyper-parameter in the context of definition 3.18. Indeed, we name hyper-parameter not only the $\psi_1, ..., \psi_p$ but also any function of them, such as, say, $(\frac{\psi_1}{\psi_2}, \psi_2 + \psi_3)$.*

Remark 4.21. *Because of proposition 4.10, non-microergodic hyper-parameters yield Gaussian measures that are equivalent to one another.*

Non-microergodic hyper-parameters can not be consistently estimated, as shown by the following proposition.

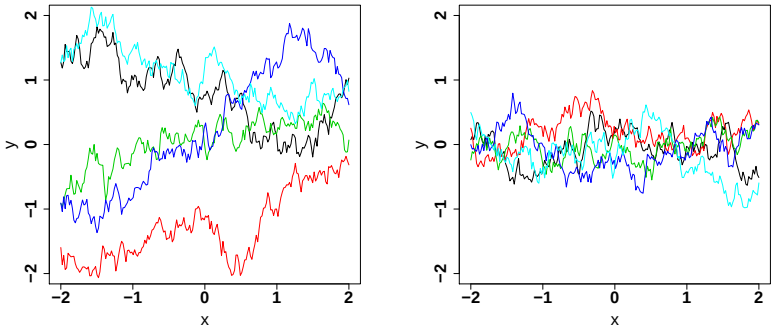

Figure 4.2: Illustration of non-microergodicity. Left: trajectories of a Gaussian process with exponential covariance function with $\sigma^2 = 1$ and $\ell = 4$. Right: trajectories of a Gaussian process with exponential covariance function with $\sigma^2 = \frac{1}{10}$ and $\ell = \frac{4}{10}$.

Proposition 4.22. *Consider a non-microergodic hyper-parameter $h(\boldsymbol{\psi})$. There does not exist an estimator $h(\hat{\boldsymbol{\psi}}) : \mathbb{R}^n \to \mathbb{R}^{p'}$ so that, for all $\boldsymbol{\psi}^{(0)} \in \Psi$, when $\boldsymbol{\psi}^{(0)}$ is the true covariance hyper-parameter, $h(\hat{\boldsymbol{\psi}})$ goes in probability to $h(\boldsymbol{\psi}^{(0)})$.*

Proof. Assume that such an estimator exists and write it $(h(\hat{\boldsymbol{\psi}})_n)_{n\in\mathbb{N}^*}$. Consider $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$ so that $h(\boldsymbol{\psi}^{(1)}) \neq h(\boldsymbol{\psi}^{(2)})$. Let $P_{\boldsymbol{\psi}^{(1)}}$ and $P_{\boldsymbol{\psi}^{(2)}}$ be the two Gaussian measures associated to $\boldsymbol{\psi}^{(1)}$ and $\boldsymbol{\psi}^{(2)}$. Then, for $i = 1, 2$, in $P_{\boldsymbol{\psi}^{(i)}}$-probability, $h(\hat{\boldsymbol{\psi}})_n$ goes to $h(\boldsymbol{\psi}^{(i)})$. Thus, we can extract a subsequence $(N_n)_{n\in\mathbb{N}^*}$ in $\mathbb{N}$, with $N_n \to_{n\to+\infty} +\infty$, so that, for $i = 1, 2$, $P_{\boldsymbol{\psi}^{(i)}}$-almost surely, $h(\hat{\boldsymbol{\psi}})_{N_n}$ goes to $h(\boldsymbol{\psi}^{(i)})$ when $n \to +\infty$. Thus, the event $\mathcal{A} := \{h(\hat{\boldsymbol{\psi}})_{N_n} \to_{n\to+\infty} h(\boldsymbol{\psi}^{(1)})\}$ verifies $P_{\boldsymbol{\psi}^{(1)}}(\mathcal{A}) = 1$ and $P_{\boldsymbol{\psi}^{(2)}}(\mathcal{A}) = 0$. This is a contradiction. □

In figure 4.2, we illustrate non-microergodicity and proposition 4.22. We plot trajectories of two Gaussian processes in dimension 1. Both have an exponential covariance function, with for the first one, $\sigma^2 = 1$, $\ell = 4$ and for the second one $\sigma^2 = \frac{1}{10}$, $\ell = \frac{4}{10}$. We will see below that the two associated Gaussian measures are equivalent, that is to say, the hyper-parameters σ^2 and ℓ are non-microergodic. We see in figure 4.2 that the trajectories are similar, notably in the sense that their local variations are of the same amplitude. Hence, a trajectory of, say, the first covariance function could have been obtained with the second covariance function. This implies that, even when observing a continuous trajectory on a bounded set (which is an infinite, non-countable, number of observations), it is still not possible to know, with probability one, the values of σ^2 and ℓ separately. Notice, in this context, that [ZZ05] shows that if a continuous trajectory is observed on a bounded set, the ML estimator of ℓ can be defined as a functional of this continuous trajectory. It is thus a random variable with non-degenerate distribution. Thus, as shown in [ZZ05], the (finite-sample) ML estimator of ℓ converges, when the number n of observations goes to infinity, to a non-degenerate random variable, and is hence inconsistent, in agreement with proposition 4.22.

However, we will see below that, for the exponential model, the hyper-parameter $\frac{\sigma^2}{\ell}$ is microergodic. For figure 4.2, this hyper-parameter was indeed the same for the two covariance functions. In figure 4.3, we illustrate this microergodicity. We plot trajectories of two Gaussian processes in dimension 1. Both have an exponential covariance function, with for the first one,

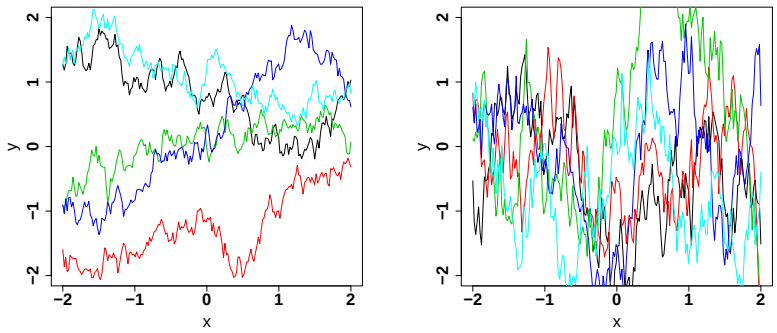

Figure 4.3: Illustration of microergodicity. Left: trajectories of a Gaussian process with exponential covariance function with $\sigma^2 = 1$ and $\ell = 4$. Right: trajectories of a Gaussian process with exponential covariance function with $\sigma^2 = 1$ and $\ell = \frac{4}{10}$.

$\sigma^2 = 1$, $\ell = 4$ and for the second one $\sigma^2 = 1$, $\ell = \frac{4}{10}$. The two associated Gaussian measures are orthogonal. We see in figure 4.3 that the trajectories are of different nature, still in the sense of their local variations: we clearly see that the local variations of the right trajectories are consistently larger than the local variations of the left trajectories. This illustrates that it can be distinguished from the left covariance function to the right one, with probability one, when observing a complete trajectory.

Contrarily to proposition 4.22, and as illustrated in figure 4.3, it is at least possible that microergodic hyper-parameters can be estimated consistently ([Ste99], p163). It is nevertheless difficult to exhibit consistent estimators for microergodic hyper-parameters. No results with the same degree of generality of, say, theorem 4.18 for the increasing-domain asymptotic framework, are yet available in the literature. Instead, consistency of estimators, like the ML estimators, are proved for particular covariance function families.

In the sequel, we will review these results, for the Matérn model of subsection 2.1.2. This review will simultaneously present the existing results on which hyper-parameters are microergodic and which hyper-parameters are non-microergodic. We would like to mention that this kind of review has also been carried out in the PhD thesis [Bet09].

Microergodicity, non-microergodicity and estimation for the Matérn model

We consider the Matérn model of subsection 2.1.2, with either the tensor product version or the isotropic version.

The most important point is that, as shown in proposition 4.23, the smoothness parameter is microergodic. This is not a surprise, since this hyper-parameter governs the regularity of the Gaussian process. Even on a fixed domain, it is conceivable that enough information can be gathered on the regularity of a Gaussian process to make it possible to distinguish between two different smoothness parameters.

Proposition 4.23. *Consider, in any dimension d, the Matérn model, with either the tensorized or isotropic version. Then the smoothness parameter ν is microergodic.*

Proof. Consider first the case $d = 1$. Let K_1 be Matérn $(\sigma_1^2, \ell_1, \nu_1)$ and K_2 be Matérn $(\sigma_2^2, \ell_2, \nu_2)$ and assume $\nu_1 < \nu_2$.

If ν_2 is infinite, the two covariance functions are orthogonal, because they yield Gaussian process trajectories with different a.s. regularities. Hence, we now consider the case where ν_1 and ν_2 are finite.

Then, with, for $i = 1, 2$, $\alpha_i = \frac{2\sqrt{\nu_i}}{\ell_i}$ and $\phi_i = \frac{\sigma_i^2 \Gamma(\nu_i + \frac{1}{2}) \alpha_i^{2\nu_i}}{\pi^{\frac{1}{2}} \Gamma(\nu_i)}$, the Fourier transform of K_1, K_2 are, for $i = 1, 2$,

$$\hat{K}_i = \phi_i \frac{1}{(\alpha_i^2 + \omega^2)^{\frac{1}{2}+\nu_i}}.$$

It is thoroughly discussed in [Ste99] that the behavior at $\omega \to +\infty$ of $\hat{K}$ is the key concept for equivalence and orthogonality of Gaussian measures. This is confirmed in [IR78], p107, where it is shown that, in the present context,

$$\lim_{\omega \to +\infty} \frac{\hat{K}_1(\omega) - \hat{K}_2(\omega)}{\hat{K}_1(\omega)} \sqrt{\omega} = +\infty \tag{4.15}$$

is a sufficient condition for the orthogonality between the Gaussian measures yielded by K_1 and K_2. Since $\nu_1 < \nu_2$, $\lim_{\omega \to +\infty} \frac{\hat{K}_1(\omega) - \hat{K}_2(\omega)}{\hat{K}_1(\omega)} = 1$, so that (4.15) holds.

Consider now the case $d > 1$. Let K_i, $i = 1, 2$, be Matérn $(\sigma_i^2, \ell_{i,1}, ..., \ell_{i,d}, \nu_i)$.

Consider the one-dimensional Gaussian process $t \to \tilde{Y}(t) := Y(\boldsymbol{x} + t\boldsymbol{e}^{(1)})$, on the compact set $\{t, \boldsymbol{x} + t\boldsymbol{e}^{(1)} \in \mathcal{D}\}$. Associate to this Gaussian process the measurable space $(\Omega, \tilde{\mathcal{F}})$, where $\tilde{\mathcal{F}} \subset \mathcal{F}$ with $(\Omega, \mathcal{F})$ the measurable space associated to Y. The Gaussian process $\tilde{Y}$ has the Matérn covariance function with hyper-parameters $(\sigma_i^2, \ell_{i,1}, \nu_i)$, for $i = 1, 2$. Thus, using the proposition for $d = 1$, the two Gaussian measures $\tilde{P}_1$, $\tilde{P}_2$, yielded by $\tilde{Y}$ are orthogonal. These two Gaussian measures are the projections of the two Gaussian measures P_1, P_2 from $\mathcal{F}$ to $\tilde{\mathcal{F}}$. Thus, the two Gaussian measures P_1 and P_2 are also orthogonal (there exists $\mathcal{A} \in \tilde{\mathcal{F}} \subset \mathcal{F}$ so that $\tilde{P}_1(\mathcal{A}) = P_1(\mathcal{A}) = 0$ and $\tilde{P}_2(\mathcal{A}) = P_2(\mathcal{A}) = 1$). □

In view of proposition 4.23, it is at least possible that the smoothness parameter ν can be estimated consistently. However, we have no knowledge of a general consistent estimator of ν in the literature. In the case of the isotropic version of the Matérn model, with $\nu < \frac{d}{2}$, [WLX13] exhibits a consistent estimator of ν, when σ^2 is unknown and the d correlation lengths are equal to an unknown common correlation length. In [WLX13], asymptotic distribution is also shown, when only ν is unknown, and a bound for the estimation error is shown when ν, σ^2 and the correlation length are unknown.

Proposition 4.23 allows us to consider now the case of a fixed and known ν. Indeed, in the case when ν is not fixed, two different ν for two Matérn covariance functions yield orthogonal Gaussian measures regardless of the values of the other hyper-parameters. Hence, if ν is unknown, the priority, with respect to the fixed-domain asymptotic theory, is to estimate it consistently.

We now, first, review the existing results for the isotropic version of the Matérn model.

For fixed and known $\nu < +\infty$, and for $d \leq 3$, the d-dimensional hyper-parameter $(\frac{\sigma^2}{\ell_1^{2\nu}}, ..., \frac{\sigma^2}{\ell_d^{2\nu}})$ is microergodic. A strongly consistent estimator of this hyper-parameter is given in [And10]. In the case where $\ell_1 = ... = \ell_d = \ell$, [Zha04] proves the strong consistency of ML for estimating $\frac{\sigma^2}{\ell^{2\nu}}$. Any hyper-parameter h so that $(h, \frac{\sigma^2}{\ell_1^{2\nu}}, ..., \frac{\sigma^2}{\ell_d^{2\nu}})$ is in one-to-one correspondence with $(\sigma^2, \ell_1, ..., \ell_d)$, such as σ^2, is non-microergodic. This non-microergodicity is proved in [Zha04].

Concerning asymptotic distribution, for $d = 1$, asymptotic normality has been proved for ML, for estimating $\frac{\sigma^2}{\ell^{2\nu}}$ ([DZM09]). This result had been proved before, in the particular case $\nu = \frac{1}{2}$, by [Yin91].

For $\nu < +\infty$, and for $d \geq 5$, all hyper-parameters of the isotropic Matérn covariance function are microergodic, as shown by [And10]. [And10] also presents a strongly consistent estimator.

Finally, the case $d = 4$ remains, to our knowledge open. It is not known whether all hyper-parameters are microergodic or only the d hyper-parameters $(\frac{\sigma^2}{\ell_1^{2\nu}}, ..., \frac{\sigma^2}{\ell_d^{2\nu}})$ are. [And10] also emphasizes that the case $d = 4$ is an open problem.

Consider now the tensor product Matérn model. The case $d = 1$ has actually been discussed above, since the isotropic and tensor product Matérn models are the same in this case. Hence consider now $d > 1$. For $\nu \in (0, \frac{1}{2})$, it is proved in [Daq10] that all the hyper-parameters $\sigma^2, \ell_1, ..., \ell_d$ are microergodic. A consistent estimator is also presented.

For $\nu = \frac{1}{2}$, [Yin93] proves that all hyper-parameters are also microergodic and that they are consistently estimated by ML, with asymptotic normality. For $\nu = \frac{3}{2}$ and $d \geq 3$, all the hyper-parameters are microergodic and are consistently estimated by ML ([Loh05]).

For the Gaussian covariance function ($\nu = +\infty$), all the hyper-parameters $\sigma^2, \ell_1, ..., \ell_d$ are microergodic. This follows for [Ste99] p120, where it is shown that, in dimension 1, two covariance functions with Fourier transforms vanishing exponentially fast are orthogonal whenever they are non identical on $\{t - s, t \in \mathcal{D}, s \in \mathcal{D}\}$. The argument for going from orthogonality in dimension one to orthogonality in dimension larger than one is similar to the proof of proposition 4.23. [LL00] proves that ML is consistent for estimating $(\ell_1, ..., \ell_d)$. To our knowledge, no consistency results are available for the ML estimation of σ^2 for the Gaussian covariance function.

Conclusion on estimation in the fixed-domain asymptotic context

As we have seen, the issue of microergodicity of the hyper-parameters needs to be solved before studying consistent estimators in the fixed-domain asymptotic framework. This first question is already difficult in itself. Indeed, as we have seen, the problem is explicitly unsolved for the isotropic Matérn covariance function for $d = 4$, and the proved results in other dimensions yield two sharply different regimes. For $d = 1, 2, 3$, not all hyper-parameters are microergodic, while for $d \geq 5$, all hyper-parameters are microergodic. For the tensor product version, in dimension $d \geq 2$, we are not aware of any non-microergodic hyper-parameters, but not all hyper-parameters are proved microergodic. The presently available results, as we have seen, depend on the smoothness parameter ν and of the dimension d.

Generally, when a hyper-parameter is proved microergodic, consistent estimators are exhibited for it. In fact, as in [And10], exhibiting consistent estimators can be a way to prove microergodicity. Studying explicitly ML is more difficult than studying an estimator designed for a particular situation. Therefore ML is not proved consistent for all hyper-parameters that are proved to be microergodic. Studying asymptotic distribution, for ML or another estimator, is even more difficult than studying consistency. The number of results available on asymptotic distribution is therefore relatively limited.

Nevertheless, all the particular results discussed above, are in agreement with the following

qualitative statement: the hyper-parameters that do have an asymptotic influence on predictions can be consistently estimated. This qualitative statement leads to the hope for a theory assessing that, in a certain sense, one can, despite using estimated covariance hyper-parameters, obtain asymptotically optimal predictions. This is discussed in the beginning of chapter 6 in [Ste99], and some results supporting this statement are shown in [PY01]. Similarly, the estimation results discussed above do not seem to contradict this kind of theory. Nevertheless, there is still room for a more unified theoretical work in this direction, which explains that hyper-parameter estimation and prediction in fixed-domain asymptotics is an active research area.

Part II

Cross Validation and Maximum Likelihood for covariance hyper-parameter estimation

Chapter 5

Cross Validation and Maximum Likelihood with well-specified family of covariance functions

This chapter is inspired by the manuscript [Bac], submitted to the Journal of Multivariate Analysis.

5.1 Introduction

This chapter 5 addresses an asymptotic investigation of hyper-parameter estimation in Kriging. Indeed, since exact finite-sample results are generally not reachable and can be specific to the situation, asymptotic theory is widely used to give approximations of the estimated hyper-parameter distribution.

We follow a triple objective here. First, we aim at studying asymptotically the CV procedure of (3.13). Indeed, while we have seen in chapter 4 that several results exist for ML, we are not aware of similar results for CV. For CV to be relevant in practice, it is preferable that, in the frameworks where ML is asymptotically consistent, it be asymptotically consistent as well. Furthermore in the cases where a rate of convergence is proved for ML, such as in subsection 4.3.1, it is desirable that CV have the same rate of convergence. Thus, the first objective of this chapter 5 is to study the consistency and asymptotic distribution of CV, in the frameworks where it has been done for ML.

Following this idea, the asymptotic theory for ML in Kriging is essentially done in the well-specified framework, meaning that the true covariance function does belong to the parametric set of covariance functions used for estimation. In this setting, we have seen in subsections 3.1.2 and 4.3.1 that the ML estimator is asymptotically unbiased, with asymptotic variance the Cramér-Rao bound. It is thus expected that ML performs asymptotically better than CV in this well-specified framework. Our second objective in this chapter 5 is to confirm this statement.

Finally, we are interested in studying the impact of the spatial sampling on the covariance function estimation. This question of how the set of experiments should be designed arises in

many areas of science involving measurements or data acquisition [Mon05]. Generally speaking, it is known that in many situations, an irregular, or even random, spatial sampling is preferable to a regular one. Examples of these situations are found in many fields. For numerical integration, Gaussian quadrature rules generally yield irregular grids [PTVF07, ch.4]. The best known low-discrepancy sequences for quasi-Monte Carlo methods (van der Corput, Halton, Sobol, Faure, Hammersley,...) are not regular either [Nie92]. In the compressed-sensing domain, it has been shown that one can recover a signal very efficiently, and at a small cost, by using random measurements [CT06].

The spatial sampling, and particularly its degree of regularity, plays an important role for the covariance function estimation. In chapter 6.9 of [Ste99], it is shown that adding three observation points with small spacing to a one-dimensional regular grid of twenty points dramatically improves the estimation in two ways. First, it enables to detect without ambiguities that a Gaussian covariance model is poorly adapted, when the true covariance function is Matérn $\frac{3}{2}$. Second, when the Matérn model is used for estimation, it subsequently improves the estimation of the smoothness parameter. It is shown in [ZZ06] that the optimal samplings, for maximizing the log of the determinant of the Fisher information matrix, averaged over a Bayesian prior on the true covariance hyper-parameters, contain closely spaced points. Similarly, in the geostatistical community, it is acknowledged that adding sampling crosses, that are small crosses of observation points making the different input quantities vary slightly, enables a better identification of the small scale behavior of the random field, and therefore a better overall estimation of its covariance function [JDLI08]. The common conclusion of the three examples we have given is that irregular samplings, in the sense that they contain at least pairs of observation points with small spacing, compared to the average density of observation points in the domain, work better for covariance function estimation than regular samplings, that is samplings with evenly spaced points. This conclusion has become commonly admitted in the Kriging literature. Our third objective is thus to address this conclusion, in an asymptotic framework.

Given these three objectives, the two main asymptotic frameworks that can be studied are increasing and fixed-domain asymptotics, as discussed in chapter 4. Notice, for the comparison between increasing and fixed-domain asymptotics, that in increasing-domain asymptotics, as shown in subsection 5.5.1, all the hyper-parameters have strong asymptotic influences on predictions. Similarly all the hyper-parameters (satisfying a very general identifiability assumption) can be consistently estimated, see chapter 4. This is the contrary, we recall, in fixed-domain asymptotics where (see chapter 4) non-microergodic hyper-parameters can not be consistently estimated and do not have asymptotic influences on predictions.

We have decided to address increasing-domain asymptotics in this chapter 5. The first reason, related to our two first goals, is that increasing-domain asymptotic results exist for ML (4.3.1) in a fairly general way. This is because increasing-domain asymptotics is a favorable setting for estimation, so that ML can be consistent and with asymptotic normality in a very general setting. On the contrary, ML is not always consistent in fixed-domain asymptotic (subsection 4.3.2) and the existing results are very specific (for instance [Yin93] addresses the case of the tensorized exponential model). Thus, despite the significant insight fixed-domain asymptotics brings on prediction and estimation (see chapter 4), studying increasing-domain asymptotics

before fixed-domain asymptotics may yield more general results. Historically, this is how this happened for ML estimation, increasing-domain asymptotics being treated in essentially the two articles [MM84] and [CL93] in 84 and 93, while fixed-domain asymptotics has been studied in many articles, from 91 ([Yin91]) onward.

The second reason for studying increasing-domain asymptotics is that we would like to compare sampling techniques by inspection of the asymptotic distributions of the hyper-parameter estimators. In fixed-domain asymptotics, when an asymptotic distribution is proved for ML [Yin91, Yin93, DZM09], it turns out that it is independent of the dense sequence of observation points. This makes it impossible to compare the effect of spatial sampling on hyper-parameter estimation using fixed-domain asymptotics techniques. On the contrary, we show in this chapter that, in increasing-domain asymptotics, the asymptotic variances of the hyper-parameter estimators strongly depend on the spatial sampling.

Thus, this chapter 5 aims at studying an increasing-domain asymptotic framework. We propose a sequence of random spatial samplings of size $n \in \mathbb{N}^*$. The regularity of the spatial sampling sequence is characterized by a regularity parameter $\epsilon \in (-\frac{1}{2}, \frac{1}{2})$. $\epsilon = 0$ corresponds to a regular grid, and the irregularity increases with ϵ. We study the ML and CV estimators of chapter 3. For CV, to the best of our knowledge, no asymptotic results are yet available in the literature. For both estimators, we prove an asymptotic normality result for the estimation, with a $\sqrt{n}$ convergence, and an asymptotic covariance matrix which is a deterministic function of ϵ. The asymptotic normality yields, classically, approximate confidence intervals for finite-sample estimation. Then, carrying out an exhaustive analysis of the asymptotic covariance matrix, for the one-dimensional Matérn model, we show that large values of the regularity parameter ϵ always yield an improvement of the ML estimation. We also show that ML has a smaller asymptotic variance than CV, which is expected since we address the well-specified case here, in which the true covariance function does belong to the parametric set used for estimation. Thus, our general conclusion is a confirmation of the aforementioned results in the literature: using a large regularity parameter ϵ yields groups of observation points with small spacing, which improve the ML estimation, which is the preferable method to use.

The rest of chapter 5 is organized as follows. In section 5.2, we introduce the random sequence of observation points, that is parameterized by the regularity parameter ϵ. In subsection 5.3.1, we give the asymptotic normality results. Some explicit expressions, for the asymptotic variances, are given in subsection 5.3.2. In section 5.4, we carry out an exhaustive study of the asymptotic variance. In section 5.5, we analyze the Kriging prediction for the asymptotic framework we consider. In section 5.7, we give the proofs for chapter 5.

5.2 Expansion-domain asymptotic framework with randomly perturbed regular grid

Stationary covariance function family

Let Y be a stationary Gaussian process on $\mathbb{R}^d$. We consider two cases for the parameterization of the covariance function of Y.

In the case of the ML estimation, the full stationary covariance function of Y is parameterized. We denote $\Psi = [\psi_{inf}, \psi_{sup}]^p$. The covariance function of Y is $K_{\boldsymbol{\psi}^{(0)}}$ with $\psi_{inf} < \psi_i^{(0)} < \psi_{sup}$, for $1 \leq i \leq p$. $K_{\boldsymbol{\psi}^{(0)}}$ belongs to a parametric model

$$\{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\}, \tag{5.1}$$

with $K_{\boldsymbol{\psi}}$ a stationary covariance function. We use this parametric model because, except for practical numerical optimization reasons that are not treated in this chapter 5, ML does not make use of the variance/correlation separation of (3.4). This separation corresponds to selecting the covariance function in the set

$$\{\sigma^2 R_{\boldsymbol{\theta}}, \sigma^2 > 0, \boldsymbol{\theta} \in \Theta\}, \tag{5.2}$$

with $R_{\boldsymbol{\theta}}$ a stationary correlation function and Θ a compact subset of $\mathbb{R}^{p-1}$.

In contrast, we have seen in chapter 3 that the CV estimation follows a two-step approach based on (5.2), estimating the correlation hyper-parameter $\boldsymbol{\theta}$ in a fist step and the variance hyper-parameter σ^2 in a second step. Basically, in this chapter 5, when addressing CV, we will focus on the estimation of the correlation hyper-parameter in (5.2). The estimation of the correlation hyper-parameter provides a sufficient insight on the impact of the spatial sampling, and on the comparison of the asymptotic distributions of ML and CV. Furthermore, the asymptotic distribution of the CV estimation of the variance hyper-parameter σ^2 can be obtained more easily, since the estimator of σ^2 is explicit (see chapter 3).

For the correlation-only estimation case of CV, two frameworks are possible. First Y can be assumed to have a known global variance hyper-parameter σ_0^2 equal to 1, which can be restrictive. Second, Y can be considered to have a constant global variance hyper-parameter σ_0^2 (because it is stationary). The value of this variance hyper-parameter does not interest us in this chapter 5, so that we can assume (incorrectly) that it is equal to $\sigma_1^2 \neq \sigma_0^2$. Now, since the distribution of the CV estimator (3.13) does not depend on the assumed variance hyper-parameter σ_1^2, nor on the true one σ_0^2, this second interpretation gives exactly the same development as the first one. Since the first one simplifies the notations, we adopt it in this chapter 5.

Thus, in the correlation-only estimation case of CV, we assume that Y has a known variance 1. We choose to write the correlation function of Y $K_{\boldsymbol{\psi}^{(0)}}$ instead of $R_{\boldsymbol{\theta}^{(0)}}$, as would have been done in the general framework of chapter 3. This is because many of the theoretical developments are common between ML and CV. Using the notation $K_{\boldsymbol{\psi}}$ for ML and $R_{\boldsymbol{\theta}}$ for CV would make it necessary to use two different notations for quantities that have exactly the same meanings. This would complicate the reading of the proofs.

Hence, in the rest of this chapter 5, $K_{\boldsymbol{\psi}}$ denotes a stationary covariance function, and, in the case of CV, we will always mention the additional condition $K_{\boldsymbol{\psi}}(0) = 1$ for all $\boldsymbol{\psi}$, meaning that $K_{\boldsymbol{\psi}}$ is a correlation function.

We denote, for both ML and CV, $\Psi = [\psi_{inf}, \psi_{sup}]^p$. The covariance function of Y is $K_{\boldsymbol{\psi}^{(0)}}$ with $\psi_{inf} < \psi_i^{(0)} < \psi_{sup}$, for $1 \leq i \leq p$. $K_{\boldsymbol{\psi}^{(0)}}$ belongs to a parametric model $\{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\}$, with $K_{\boldsymbol{\psi}}$ a stationary covariance function.

We shall assume the following condition for the parametric model $\{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\}$. This condition is satisfied in all classical cases, and especially for the Matérn model of chapter 2.

Condition 5.1. *i) For all $\boldsymbol{\psi} \in \Psi$, the covariance function $K_{\boldsymbol{\psi}}$ is stationary. The covariance function $K_{\boldsymbol{\psi}}$ is three times differentiable with respect to $\boldsymbol{\psi}$. For all $q \in \{0, ..., 3\}$, $i_1, ..., i_q \in \{1, ..., p\}$, there exists $C_{i_1,...,i_q} < +\infty$ so that for all $\boldsymbol{\psi} \in \Psi$, $\boldsymbol{t} \in \mathbb{R}^d$,*

$$\frac{\partial}{\partial \psi_{i_1}} ... \frac{\partial}{\partial \psi_{i_q}} K_{\boldsymbol{\psi}}(\boldsymbol{t}) \leq \frac{C_{i_1,...,i_q}}{1 + |\boldsymbol{t}|^{d+1}}, \tag{5.3}$$

where $|\boldsymbol{t}|$ is the Euclidian norm of $\boldsymbol{t}$.

We define the Fourier transform of a function $h : \mathbb{R}^d \to \mathbb{R}$ by

$$\hat{h}(\boldsymbol{f}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} h(\boldsymbol{t}) e^{-\mathrm{i}\boldsymbol{f}\cdot\boldsymbol{t}} d\boldsymbol{t},$$

where $\mathrm{i}^2 = -1$. Then, for all $\boldsymbol{\psi} \in \Psi$, the covariance function $K_{\boldsymbol{\psi}}$ has a Fourier transform $\hat{K}_{\boldsymbol{\psi}}$ that is continuous and bounded.

ii) For all $\boldsymbol{\psi} \in \Psi$, $K_{\boldsymbol{\psi}}$ satisfies

$$K_{\boldsymbol{\psi}}(\boldsymbol{t}) = \int_{\mathbb{R}^d} \hat{K}_{\boldsymbol{\psi}}(\boldsymbol{f}) e^{\mathrm{i}\boldsymbol{f}\cdot\boldsymbol{t}} d\boldsymbol{f}.$$

iii) $(\boldsymbol{\psi}, \boldsymbol{f}) \to \hat{K}_{\boldsymbol{\psi}}(\boldsymbol{f})$ is continuous and positive on $\Psi \times \mathbb{R}^d$.

Let us make some comments on condition 5.1. Condition *i)*, as we will see, implies the summability of the stationary covariance function over all the observation points. The same summability assumption was present in theorem 4.18. Basically, it ensures a non-redundancy of the observations between distant observation points. Let us remark also that condition *i)* implies the summability over $\mathbb{R}^d$ of, say, $K_{\boldsymbol{\psi}}$, because, by a multidimensional spherical change of variables, with $2\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$ the area of the the unit sphere in dimension d,

$$\int_{\mathbb{R}^d} \frac{1}{1 + |\boldsymbol{t}|^{d+1}} d\boldsymbol{t} = 2\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \frac{r^{d-1}}{1 + r^{d+1}} dr < +\infty.$$

Condition *ii)* is a separate assumption from condition *i)*, because the latter does not give information on the regularity of $K_{\boldsymbol{\psi}}(\boldsymbol{t})$ w.r.t. $\boldsymbol{t}$, or similarly, on the summability of $\hat{K}_{\boldsymbol{\psi}}(\boldsymbol{f})$ w.r.t. $\boldsymbol{f}$. Condition *ii)* is especially used in the proof of proposition 5.26.

Condition *iii)*, the positivity of the Fourier transform, implies that all the covariance matrices obtained from n different observation points are invertible. To see this, write for n different observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, and n scalar coefficients $\alpha_1, ..., \alpha_n$,

$$0 \leq \sum_{i,j=1}^{n} \alpha_i \alpha_j K_{\boldsymbol{\psi}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \int_{\mathbb{R}^d} \hat{K}(\boldsymbol{\omega}) \left| \sum_{j=1}^{n} \alpha_j e^{\left(\mathrm{i}\boldsymbol{\omega}.\boldsymbol{x}^{(j)}\right)} \right|^2 d\boldsymbol{\omega},$$

and notice that the $\boldsymbol{\omega} \to e^{\left(\mathrm{i}\boldsymbol{\omega}.\boldsymbol{x}^{(j)}\right)}$ are n linearly independent functions for different observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. We speak of non-degenerate covariance functions for the stationary covariance functions verifying *iii)*.

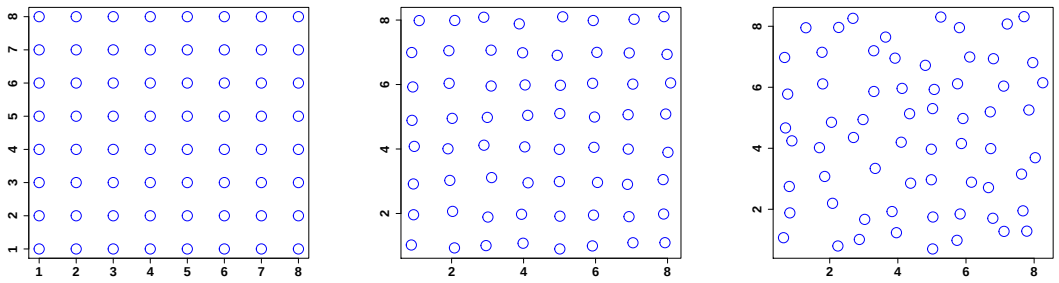

Figure 5.1: Examples of three perturbed grids. The dimension is $d = 2$ and the number of observation points is $n = 8^2$. From left to right, the values of the regularity parameter are 0, $\frac{1}{8}$ and $\frac{3}{8}$. $\epsilon = 0$ corresponds to a regular observation grid, while, when $|\epsilon|$ is close to $\frac{1}{2}$, the observation set is highly irregular.

Randomly perturbed regular grid

We denote by $\left(\boldsymbol{v}^{(i)}\right)_{i\in\mathbb{N}^*}$ a sequence of deterministic points in $\mathbb{N}^d$ so that for all $N \in \mathbb{N}^*$, $\left\{\boldsymbol{v}^{(i)}, 1 \leq i \leq N^d\right\} = \{1, ..., N\}^d$. Basically the $\boldsymbol{v}^{(i)}$ constitute a square regular grid on $(\mathbb{N}^*)^d$. See figure 5.1, left plot, for an example in dimension 2.

Y is observed at the points $\boldsymbol{v}^{(i)} + \epsilon X_i$, $1 \leq i \leq n$, $n \in \mathbb{N}^*$, with $-\frac{1}{2} < \epsilon < \frac{1}{2}$ and $X_i \sim_{iid} \mathcal{L}_X$. $\mathcal{L}_X$ is a symmetric probability distribution with support $S_X \subset [-1, 1]^d$, and with a positive probability density function on S_X. ϵX_i is the random perturbation of the grid at the point $\boldsymbol{v}^{(i)}$. We denote, for $n \in \mathbb{N}^*$, $\boldsymbol{X} = (X_1, ..., X_n)$ as the perturbation vector, where we do not write explicitly the dependence in n for clarity. $\boldsymbol{X}$ is a random variable with distribution $\mathcal{L}_X^{\otimes n}$.

Two remarks can be made on this sequence of observation points:

- This is indeed an increasing-domain asymptotic context. The condition $-\frac{1}{2} < \epsilon < \frac{1}{2}$ ensures a minimal spacing between two distinct observation points.
- The observation sequence we study is random, and the parameter ϵ is a regularity parameter. $\epsilon = 0$ corresponds to a regular observation grid, while, when $|\epsilon|$ is close to $\frac{1}{2}$, the observation set is highly irregular. Examples of observation sets are given in figure 5.1, with $d = 2$, $n = 8^2$, and different values of ϵ.

Maximum Likelihood and Cross Validation

We recall $L(\boldsymbol{\psi}) := \frac{1}{n}\left\{\ln\left(|\mathbf{K}_{\boldsymbol{\psi}}|\right) + \boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y}\right\}$ the modified opposite log-likelihood criterion of chapter 3, where we do not write explicitly the dependence in $\boldsymbol{X}$, Y, n and ϵ. We denote by $\hat{\boldsymbol{\psi}}_{ML}$ the Maximum Likelihood estimator, defined by

$$\hat{\boldsymbol{\psi}}_{ML} \in \underset{\boldsymbol{\psi}\in\Psi}{\operatorname{argmin}}\, L(\boldsymbol{\psi}), \tag{5.4}$$

where we do not write explicitly the dependence of $\hat{\boldsymbol{\psi}}_{ML}$ with respect to $\boldsymbol{X}$, Y, ϵ and n.

Remark 5.2. *The ML estimator in* (5.4) *is actually not entirely defined, since the likelihood function of* (5.4) *can have more than one global minimizer. Nevertheless, the convergence results*

of $\hat{\boldsymbol{\psi}}_{ML}$, as $n \to +\infty$, hold when $\hat{\boldsymbol{\psi}}_{ML}$ is any random vector belonging to the set of the global minimizers of the likelihood of (5.4)*, regardless of the value chosen in this set. Furthermore, it can be shown that, with probability converging to one, as $n \to \infty$ (see remark 5.39 in subsection 5.7.1), the likelihood function has a unique global minimum. To define a measurable function $\hat{\boldsymbol{\psi}}_{ML}$ of Y and $\boldsymbol{X}$, belonging to the set of the minimizers of the likelihood, one possibility is the following. For a given realization of Y and $\boldsymbol{X}$, let $\mathcal{K}$ be the set of the minimizers of the likelihood. Let $\mathcal{K}_0 = \mathcal{K}$ and, for $0 \leq k \leq p-1$, $\mathcal{K}_{k+1}$ is the subset of $\mathcal{K}_k$ whose elements have their $k+1$th coordinates equal to $\min\left\{\tilde{\boldsymbol{\psi}}_{k+1}, \tilde{\boldsymbol{\psi}} \in \mathcal{K}_k\right\}$. Since, $\mathcal{K}$ is compact (because the likelihood function is continuous with respect to $\boldsymbol{\psi}$ and defined on the compact set Ψ), the set $\mathcal{K}_p$ is composed of a unique element, that we define as $\hat{\boldsymbol{\psi}}_{ML}$, which is a measurable function of $\boldsymbol{X}$ and Y. The same remark can be made for the Cross Validation estimator of* (5.5)*.*

When the increasing-domain asymptotics sequence of observation points is deterministic, we have seen in chapter 4 that $\hat{\boldsymbol{\psi}}_{ML}$ converges to a centered Gaussian random vector (under suitable assumptions). The asymptotic covariance matrix is the inverse of the Fisher information matrix. Since the literature has not addressed yet the asymptotic distribution of $\hat{\boldsymbol{\psi}}_{ML}$ in increasing-domain asymptotics with random observation points, we give complete proofs about it in subsection 5.7.1. Our techniques are original and not specifically oriented towards ML contrary to the ones in chapter 4, so that they allow us to address the asymptotic distribution of the CV estimator in the same fashion.

We recall the CV estimation of the correlation hyper-parameter $\boldsymbol{\psi}$,

$$\hat{\boldsymbol{\psi}}_{LOO} \in \underset{\boldsymbol{\psi} \in \Psi}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - \hat{y}_{i,\boldsymbol{\psi}}\}^2, \tag{5.5}$$

where, for $1 \leq i \leq n$, $\hat{y}_{i,\boldsymbol{\psi}} := \mathbb{E}_{\boldsymbol{\psi}|\boldsymbol{X}}\left(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\right)$ is the Kriging Leave-One-Out prediction of y_i with covariance hyper-parameters $\boldsymbol{\psi}$. $\mathbb{E}_{\boldsymbol{\psi}|\boldsymbol{X}}$ denotes the expectation with respect to the distribution of Y with the covariance function $K_{\boldsymbol{\psi}}$, given $\boldsymbol{X}$.

Recall also that the criterion (5.5) can be computed with a single matrix inversion, by means of the virtual LOO formulas, see chapter 3. These virtual LOO formulas yield

$$\sum_{i=1}^{n} \{y_i - \hat{y}_{i,\boldsymbol{\psi}}\}^2 = \boldsymbol{y}^t \mathbf{K}_{\boldsymbol{\psi}}^{-1} Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \boldsymbol{y},$$

which will also be useful in the proofs on CV. We hence define

$$LOO(\boldsymbol{\psi}) := \frac{1}{n} \boldsymbol{y}^t \mathbf{K}_{\boldsymbol{\psi}}^{-1} Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \boldsymbol{y}$$

as the CV criterion, where we do not write explicitly the dependence in $\boldsymbol{X}$, n, Y and ϵ. Hence we have, equivalently to (5.5), $\hat{\boldsymbol{\psi}}_{LOO} \in \operatorname{argmin}_{\boldsymbol{\psi} \in \Psi} LOO(\boldsymbol{\psi})$.

Identifiability

A very important point is that, for a given $\epsilon > 0$, the difference between two different observation points is

$$\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon(X_i - X_j).$$

This difference is thus of the form

$$\boldsymbol{v} + \epsilon \boldsymbol{t},$$

with $\boldsymbol{v} \in (\mathbb{Z})^d \backslash 0$ and $\boldsymbol{t} \in C_{S_X}$, where

$$C_{S_X} := \left\{ \boldsymbol{t}^{(1)} - \boldsymbol{t}^{(2)}, \boldsymbol{t}^{(1)} \in S_X, \boldsymbol{t}^{(2)} \in S_X \right\}$$

is the set of all possible differences between two points in S_X. So, the set

$$D_\epsilon := \cup_{\boldsymbol{v} \in \mathbb{Z}^d \backslash 0} \left(\boldsymbol{v} + \epsilon C_{S_X} \right) \tag{5.6}$$

is the set of all the possible difference vectors between two different observation points. We also call this set the set of inter-point distances covered by the random sampling.

Two covariance functions that differ only for points outside D_ϵ in (5.6) can not be distinguished with the random sampling we study. Thus, the two following identifiability conditions are necessary for the ML and CV estimators to be consistent.

Condition 5.3. *For $\epsilon = 0$, there does not exist $\boldsymbol{\psi} \neq \boldsymbol{\psi}^{(0)}$ so that $K_{\boldsymbol{\psi}}(\boldsymbol{v}) = K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{Z}^d$.*

For $\epsilon \neq 0$, with D_ϵ as in (5.6), there does not exist $\boldsymbol{\psi} \neq \boldsymbol{\psi}^{(0)}$ so that $K_{\boldsymbol{\psi}} = K_{\boldsymbol{\psi}^{(0)}}$ a.s. on D_ϵ, according to the Lebesgue measure on D_ϵ, and $K_{\boldsymbol{\psi}}(0) = K_{\boldsymbol{\psi}^{(0)}}(0)$.

Condition 5.4. *For $\epsilon = 0$, there does not exist $\boldsymbol{\psi} \neq \boldsymbol{\psi}^{(0)}$ so that $K_{\boldsymbol{\psi}}(\boldsymbol{v}) = K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{Z}^d \backslash 0$.*

For $\epsilon \neq 0$, with D_ϵ as in (5.6), there does not exist $\boldsymbol{\psi} \neq \boldsymbol{\psi}^{(0)}$ so that $K_{\boldsymbol{\psi}} = K_{\boldsymbol{\psi}^{(0)}}$ a.s. on D_ϵ, according to the Lebesgue measure on D_ϵ.

Notice the slight difference between condition 5.3 for ML and 5.4 for CV. Since ML also aims at estimating a variance hyper-parameter impacting on $K_{\boldsymbol{\psi}}(0)$ only, its identifiability condition is slightly relaxed compared to that of CV.

We also state the two local identifiability conditions 5.5 and 5.6. We call them local in contrast with the identifiability conditions 5.3 and 5.4 that are global.

Condition 5.5. *For $\epsilon = 0$, there does not exist $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero, so that $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in \mathbb{Z}^d$.*

For $\epsilon \neq 0$, with D_ϵ as in (5.6), there does not exist $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero, so that $\boldsymbol{t} \to \sum_{k=1}^p \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})$ is almost surely zero on D_ϵ, with respect to the Lebesgue measure on D_ϵ, and that $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(0)$ is null.

Condition 5.6. *For $\epsilon = 0$, there does not exist $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero, so that $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in \mathbb{Z}^d \backslash 0$.*

For $\epsilon \neq 0$, with D_ϵ as in (5.6), there does not exist $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero, so that $\boldsymbol{t} \to \sum_{k=1}^p \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})$ is almost surely zero on D_ϵ, with respect to the Lebesgue measure on D_ϵ.

We will see in propositions 5.10 and 5.14 that the conditions 5.5 and 5.6 are necessary for the asymptotic distributions of ML and CV to exist with a "non-degenerate" $\sqrt{n}$ rate of convergence. For an immediate interpretation, for instance for condition 5.5 with $\epsilon = 0$, assume that there exists $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero, so that $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) = 0$ for all

$\boldsymbol{v} \in \mathbb{Z}^d$. Then, for t small, the two hyper-parameters $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\psi}^{(t)} = (\psi_1^{(0)} + t\lambda_1, ..., \psi_d^{(0)} + t\lambda_d)$ verify, for all $\boldsymbol{v} \in \mathbb{Z}^d$,

$$|K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) - K_{\boldsymbol{\psi}^{(t)}}(\boldsymbol{v})| = o(t).$$

Hence, two different hyper-parameters $\boldsymbol{\psi}^{(0)}$ and $\boldsymbol{\psi}^{(t)}$, with a difference of the order t, gives, up to a $o(t)$, the same covariance function. We interpret this as a non-identifiability of the covariance model $K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi$, locally around $\boldsymbol{\psi}^{(0)}$.

Notation

We recall that, for $n \in \mathbb{N}^*$, $\boldsymbol{X} = (X_1, ..., X_n)$ is the perturbation vector, where we do not write explicitly the dependence in n for clarity. $\boldsymbol{X}$ is a random variable with distribution $\mathcal{L}_X^{\otimes n}$. We also denote $\boldsymbol{x} = \left(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}\right)$, an element of $(S_X)^n$, as a realization of $\boldsymbol{X}$.

We define the $n \times n$ random covariance matrix $\mathbf{K}_{\boldsymbol{\psi}}$ by

$$(\mathbf{K}_{\boldsymbol{\psi}})_{i,j} = K_{\boldsymbol{\psi}}\left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(X_i - X_j\right)\right).$$

We do not write explicitly the dependence of $\mathbf{K}_{\boldsymbol{\psi}}$ with respect to $\boldsymbol{X}$, ϵ and n. We shall denote, as a simplification, $\mathbf{K} := \mathbf{K}_{\boldsymbol{\psi}^{(0)}}$.

We define the random vector $\boldsymbol{y}$ of size n by $y_i = Y\left(\boldsymbol{v}^{(i)} + \epsilon X_i\right)$. We do not write explicitly the dependence of $\boldsymbol{y}$ with respect to $\boldsymbol{X}$, ϵ and n.

We denote, for a real $n \times n$ matrix $\mathbf{A}$, $||\mathbf{A}||_2^2 = \frac{1}{n}\sum_{ij,=1}^n A_{i,j}^2$ and $||\mathbf{A}||$ the largest singular value of $\mathbf{A}$. $||.||_2$ and $||.||$ are norms and $||.||$ is a matrix norm. We denote by $\phi_i(\mathbf{M})$, $1 \leq i \leq n$, the eigenvalues of a symmetric matrix $\mathbf{M}$. We denote, for two sequences of square matrices $\mathbf{A}$ and $\mathbf{B}$, depending on $n \in \mathbb{N}^*$, $\mathbf{A} \sim \mathbf{B}$ if $||\mathbf{A} - \mathbf{B}||_2 \to_{n\to+\infty} 0$ and $||\mathbf{A}||$ and $||\mathbf{B}||$ are bounded with respect to n. Finally, for a square matrix $\mathbf{A}$, we denote by $Diag(\mathbf{A})$ the matrix obtained by setting to 0 all non diagonal elements of $\mathbf{A}$.

Finally, for a sequence of real random variables z_n, we denote $z_n \to_p 0$ and $z_n = o_p(1)$ when z_n converges to zero in probability.

5.3 Consistency and asymptotic normality for Maximum Likelihood and Cross Validation

5.3.1 Consistency and asymptotic normality

Maximum Likelihood

Proposition 5.7 addresses the consistency of the ML estimator. The only assumptions on the parametric family of covariance functions are the regularity and summability assumption 5.1 and the identifiability assumption 5.3. This identifiability assumption is necessary.

Proposition 5.7. *Assume that conditions 5.1 and 5.3 are satisfied. Then the ML estimator is consistent.*

In proposition 5.8, we address the asymptotic normality of ML. The convergence rate is $\sqrt{n}$, as in a classical *iid* framework, and we prove the existence of a deterministic asymptotic covariance matrix of $\sqrt{n}\hat{\boldsymbol{\psi}}_{ML}$, which depends only on the regularity parameter ϵ.

Proposition 5.8. *Assume that condition 5.1 is satisfied.*

For all $1 \leq i, j \leq p$, *the random trace* $\frac{1}{n} Tr\left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_j}\right)$ *converges a.s. to the element* $(\mathbf{\Sigma}_{ML})_{i,j}$ *of a* $p \times p$ *deterministic matrix* $\mathbf{\Sigma}_{ML}$ *as* $n \to +\infty$.

If $\hat{\boldsymbol{\psi}}_{ML}$ *is consistent and if* $\mathbf{\Sigma}_{ML}$ *is positive, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}_{ML} - \boldsymbol{\psi}^{(0)}\right) \to_{\mathcal{L}} \mathcal{N}\left(0, 2\mathbf{\Sigma}_{ML}^{-1}\right).$$

Remark 5.9. *In proposition 5.8, we call* $\frac{1}{n} Tr\left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_j}\right)$ *a random trace because, for* $\epsilon \neq 0$, *it is a function of the random perturbation vector* $\boldsymbol{X}$. *When* $\epsilon = 0$, *we still call this quantity a random trace, although it is deterministic. This is still mathematically correct, and it facilitates the discussions by avoiding to distinguish the two cases* $\epsilon = 0$ *and* $\epsilon \neq 0$. *We will follow this principle for CV in proposition 5.12.*

In proposition 5.10, we prove that the asymptotic Fisher information matrix $\mathbf{\Sigma}_{ML}$ is positive, as long as the local identifiability condition 5.5 holds.

Proposition 5.10. *Assume that conditions 5.1 and 5.5 are satisfied. Then* $\mathbf{\Sigma}_{ML}$ *is positive.*

The condition 5.5 is necessary in proposition 5.10. To see this, assume, for instance with $\epsilon = 0$, that there exists $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero, so that $\sum_{k=1}^{p} \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) = 0$ for all $\boldsymbol{v} \in \mathbb{Z}^d$. Then

$$\begin{aligned}
\boldsymbol{\lambda}^t \mathbf{\Sigma}_{ML} \boldsymbol{\lambda} &= \lim_{n \to +\infty} \frac{1}{n} \sum_{i,j=1}^{p} \lambda_i \lambda_j Tr\left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_j}\right) \\
&= \lim_{n \to +\infty} \frac{1}{n} Tr\left(\mathbf{K}^{-1} \left(\sum_{i=1}^{p} \frac{\partial \mathbf{K}}{\partial \psi_i}\right) \mathbf{K}^{-1} \left(\sum_{j=1}^{p} \frac{\partial \mathbf{K}}{\partial \psi_j}\right)\right).
\end{aligned}$$

For all n, the matrix $\sum_{j=1}^{p} \frac{\partial \mathbf{K}}{\partial \psi_j}$ is the zero matrix so that, by taking the limit as $n \to +\infty$, $\boldsymbol{\lambda}^t \mathbf{\Sigma}_{ML} \boldsymbol{\lambda} = 0$, meaning that the matrix $\mathbf{\Sigma}_{ML}$ is singular.

Cross Validation

Proposition 5.11 addresses the consistency of the CV estimator. The identifiability assumption is required, like in the ML case. Notice also that, as discussed above, the CV estimator (5.5) is designed for estimating the correlation hyper-parameter.

Proposition 5.11. *Assume that conditions 5.1 and 5.4 are satisfied and that for all* $\boldsymbol{\psi} \in \Psi$, $K_{\boldsymbol{\psi}}(0) = 1$. *Then the CV estimator is consistent.*

Proposition 5.12 gives the expression of the covariance matrix of the gradient of the CV criterion $LOO(\boldsymbol{\psi})$ and of the mean matrix of its Hessian. As we have seen in chapter 3, these moments are classically used in statistics to prove asymptotic distributions of consistent estimators. We also prove the convergence of these moments to $p \times p$ matrices $\mathbf{\Sigma}_{CV,1}$ and $\mathbf{\Sigma}_{CV,2}$, of which we prove the existence. These matrices are deterministic and depend only on the regularity parameter ϵ.

Proposition 5.12. *Assume that condition 5.1 is satisfied and that for all* $\boldsymbol{\psi} \in \Psi$, $K_{\boldsymbol{\psi}}(0) = 1$.

With, for $1 \leq i \leq p$,

$$\mathbf{M}_{\boldsymbol{\psi}}^{i} = \mathbf{K}_{\boldsymbol{\psi}}^{-1} Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2} \left\{ Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}}^{-1}\right) Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1} - \mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i} \right\} \mathbf{K}_{\boldsymbol{\psi}}^{-1},$$

we have, for all $1 \leq i, j \leq p$,

$$\frac{\partial}{\partial \psi_i} LOO(\boldsymbol{\psi}) = \frac{1}{n} 2 \boldsymbol{y}^t \mathbf{M}_{\boldsymbol{\psi}}^{i} \boldsymbol{y},$$

and

$$Cov\left(\sqrt{n} \frac{\partial}{\partial \psi_i} LOO(\boldsymbol{\psi}^{(0)}), \sqrt{n} \frac{\partial}{\partial \psi_j} LOO(\boldsymbol{\psi}^{(0)}) \middle| \boldsymbol{X} \right) = \tag{5.7}$$
$$2 \frac{1}{n} Tr\left[\left\{ \mathbf{M}_{\boldsymbol{\psi}^{(0)}}^{i} + \left(\mathbf{M}_{\boldsymbol{\psi}^{(0)}}^{i}\right)^t \right\} \mathbf{K}_{\boldsymbol{\psi}^{(0)}} \left\{ \mathbf{M}_{\boldsymbol{\psi}^{(0)}}^{j} + \left(\mathbf{M}_{\boldsymbol{\psi}^{(0)}}^{j}\right)^t \right\} \mathbf{K}_{\boldsymbol{\psi}^{(0)}} \right].$$

Furthermore, the random trace in (5.7) *converges a.s. to the element* $\left(\boldsymbol{\Sigma}_{CV,1}\right)_{i,j}$ *of a* $p \times p$ *deterministic matrix* $\boldsymbol{\Sigma}_{CV,1}$ *as* $n \to +\infty$.

We also have

$$\mathbb{E}\left(\frac{\partial^2}{\partial \psi_i \partial \psi_j} LOO(\boldsymbol{\psi}^{(0)}) | \boldsymbol{X} \right) = \tag{5.8}$$
$$- \frac{8}{n} Tr\left\{ Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)^{-3} Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right) \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi_j} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \right\}$$
$$+ \frac{2}{n} Tr\left\{ Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)^{-2} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi_j} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \right\}$$
$$+ \frac{6}{n} Tr\left\{ Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)^{-4} Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right) Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi_j} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right) \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \right\}.$$

Furthermore, the random trace in (5.8) *converges a.s. to the element* $\left(\boldsymbol{\Sigma}_{CV,2}\right)_{i,j}$ *of a* $p \times p$ *deterministic matrix* $\boldsymbol{\Sigma}_{CV,2}$ *as* $n \to +\infty$.

In proposition 5.13, we address the asymptotic normality of CV. The convergence rate is also $\sqrt{n}$, and we have the expression of the deterministic asymptotic covariance matrix of $\sqrt{n}\hat{\boldsymbol{\psi}}_{LOO}$, depending only on the matrices $\boldsymbol{\Sigma}_{CV,1}$ and $\boldsymbol{\Sigma}_{CV,2}$ of proposition 5.12.

Proposition 5.13. *Assume that condition 5.1 is satisfied and that for all* $\boldsymbol{\psi} \in \Psi$, $K_{\boldsymbol{\psi}}(0) = 1$.

If $\hat{\boldsymbol{\psi}}_{LOO}$ *is consistent and if* $\boldsymbol{\Sigma}_{CV,2}$ *is positive, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}_{LOO} - \boldsymbol{\psi}^{(0)}\right) \to_{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{\Sigma}_{CV,2}^{-1} \boldsymbol{\Sigma}_{CV,1} \boldsymbol{\Sigma}_{CV,2}^{-1}\right) \quad as\ n \to +\infty.$$

In proposition 5.14, we prove that the asymptotic Hessian matrix $\boldsymbol{\Sigma}_{CV,2}$ is positive as long as the local identifiability condition 5.6 holds.

Proposition 5.14. *Assume that conditions 5.1 and 5.6 are satisfied and that for all* $\boldsymbol{\psi} \in \Psi$, $K_{\boldsymbol{\psi}}(0) = 1$. *Then* $\boldsymbol{\Sigma}_{CV,2}$ *is positive.*

The condition 5.6 is necessary in proposition 5.14. This can be seen the same way as for proposition 5.10 for ML: if the condition does not hold, for each fixed n, the Hessian matrix of the CV criterion has a non-empty kernel that is independent of n. Thus, the limit matrix $\boldsymbol{\Sigma}_{CV,2}$ is singular.

The conclusion for ML and CV is that, for all the most classical parametric families of covariance functions, consistency and asymptotic normality hold, with deterministic positive asymptotic covariance matrices depending only on the regularity parameter ϵ. The rate of convergence is $\sqrt{n}$ in both cases. This result was the first objective of this chapter 5.

In section 5.4, we analyze the asymptotic covariance matrices of propositions 5.8 and 5.13. We aim at comparing them to verify that, for $p = 1$, the asymptotic variance is smaller for ML than for CV. We also aim at studying their dependence with respect to the regularity parameter ϵ, to address the influence of the irregularity of the spatial sampling on the ML and CV estimation.

In order to do so, we are interested in the derivatives of the asymptotic covariance matrices with respect to ϵ. We hence now study this point.

Derivatives of the asymptotic covariance matrices

In proposition 5.16 we show that, under the mild conditions 5.15, the asymptotic covariance matrices obtained from $\mathbf{\Sigma}_{ML}$, $\mathbf{\Sigma}_{CV,1}$ and $\mathbf{\Sigma}_{CV,2}$, of propositions 5.8 and 5.12, are twice differentiable with respect to ϵ. This result is useful for the numerical study of the section 5.4.

Condition 5.15.

- *Condition 5.1 is satisfied.*
- $K_{\boldsymbol{\psi}}(\boldsymbol{t})$ *and* $\frac{\partial}{\partial \psi_i} K_{\boldsymbol{\psi}}(\boldsymbol{t})$, *for* $1 \leq i \leq p$, *are three times differentiable in* $\boldsymbol{t}$ *for* $\boldsymbol{t} \neq 0$.
- *For all* $T > 0$, $1 \leq i \leq p$, $k \in \{1,2,3\}$, $i_1, ..., i_k \in \{1, ..., d\}^k$, *there exists* $C_T < +\infty$ *so that for* $|\boldsymbol{t}| \geq T$, $\boldsymbol{\psi} \in \Psi$,

$$\begin{aligned} \frac{\partial}{\partial t_{i_1}} ... \frac{\partial}{\partial t_{i_k}} K_{\boldsymbol{\psi}}(\boldsymbol{t}) &\leq \frac{C_T}{1 + |\boldsymbol{t}|^{d+1}}, \\ \frac{\partial}{\partial t_{i_1}} ... \frac{\partial}{\partial i_k} \frac{\partial}{\partial \psi_i} K_{\boldsymbol{\psi}}(\boldsymbol{t}) &\leq \frac{C_T}{1 + |\boldsymbol{t}|^{d+1}}. \end{aligned} \tag{5.9}$$

Proposition 5.16. *Assume that condition 5.15 is satisfied.*

Let us fix $1 \leq i, j \leq p$. *The elements* $(\mathbf{\Sigma}_{ML})_{i,j}$, $(\mathbf{\Sigma}_{CV,1})_{i,j}$ *and* $(\mathbf{\Sigma}_{CV,2})_{i,j}$ *(as defined in propositions 5.8 and 5.12) are* C^2 *in* ϵ *on* $[0, \frac{1}{2})$. *Furthermore, let us define the matrices* $\mathbf{M}_{ML}^{(i,j)}$, $\mathbf{M}_{CV,1}^{(i,j)}$ *and* $\mathbf{M}_{CV,2}^{(i,j)}$ *by the relations* $\frac{1}{n}\mathbb{E}\left\{Tr\left(\mathbf{M}_{ML}^{(i,j)}\right)\right\} \to (\mathbf{\Sigma}_{ML})_{i,j}$, $\frac{1}{n}\mathbb{E}\left\{Tr\left(\mathbf{M}_{CV,1}^{(i,j)}\right)\right\} \to (\mathbf{\Sigma}_{CV,1})_{i,j}$ *and* $\frac{1}{n}\mathbb{E}\left\{Tr\left(\mathbf{M}_{CV,2}^{(i,j)}\right)\right\} \to (\mathbf{\Sigma}_{CV,2})_{i,j}$ *in propositions 5.8 and 5.12. We then have, for* $(\mathbf{\Sigma})_{i,j}$ *being* $(\mathbf{\Sigma}_{ML})_{i,j}$, $(\mathbf{\Sigma}_{CV,1})_{i,j}$ *or* $(\mathbf{\Sigma}_{CV,2})_{i,j}$ *and* $\mathbf{M}^{(i,j)}$ *being* $\mathbf{M}_{ML}^{(i,j)}$, $\mathbf{M}_{CV,1}^{(i,j)}$ *or* $\mathbf{M}_{CV,2}^{(i,j)}$

$$\frac{\partial^2}{\partial \epsilon^2}(\mathbf{\Sigma})_{i,j} = \lim_{n \to +\infty} \frac{1}{n} \mathbb{E}\left\{ \frac{\partial^2}{\partial \epsilon^2} Tr\left(\mathbf{M}^{(i,j)}\right) \right\}.$$

Proposition 5.16 shows that we can compute numerically the derivatives of $(\mathbf{\Sigma}_{ML})_{i,j}$, $(\mathbf{\Sigma}_{CV,k})_{i,j}$, $k = 1, 2$, with respect to ϵ by computing the derivatives of $\mathbf{M}_{ML}^{(i,j)}$, $\mathbf{M}_{CV,k}^{(i,j)}$, $k = 1, 2$, for n large. The fact that it is possible to exchange the limit in n and the derivative in ϵ was not *a priori* obvious.

5.3.2 Closed form expressions of the asymptotic variances in dimension one

The asymptotic covariance matrices of propositions 5.8 and 5.13 are expressed as functions of a.s. limits of traces of sums, products and inverses of random matrices. In the case $\epsilon = 0$, for $d = 1$, these matrices are deterministic Toeplitz matrices. A $n \times n$ Toeplitz matrix $\mathbf{M}$ is a matrix for which there exist $s_{-(n-1)}, ..., s_{n-1}$ so that

$$M_{i,j} = s_{i-j}. \tag{5.10}$$

For $d = 1$ and $\epsilon = 0$, $(\mathbf{K}_{\boldsymbol{\psi}})_{i,j} = K_{\boldsymbol{\psi}}(i-j)$, so that $\mathbf{K}_{\boldsymbol{\psi}}$ is a Toeplitz matrix.

There exist results for the limits as $n \to +\infty$ of traces of Toeplitz matrices. These limits are based on Fourier transform techniques and we refer to [Gra01] for a further reading on this subject. We will give a short overview about it in subsection 5.7.2.

Furthermore, for $d = 1$, the second derivatives with respect to ϵ, at $\epsilon = 0$, of the asymptotic variance for ML and CV are also expressed as almost sure limits of traces of random matrices (proposition 5.16). In proposition 5.16, for $d = 1$, the random matrices $\mathbf{M}_{ML}^{(i,j)}$, $\mathbf{M}_{CV,1}^{(i,j)}$ and $\mathbf{M}_{CV,2}^{(i,j)}$ conserve some sort of a Toeplitz structure. This makes it possible, in the ML case, for $p = 1$, to obtain an explicit expression of $\frac{\partial^2}{\partial \epsilon^2} (\Sigma_{ML})_{i,j}$, at $\epsilon = 0$, that we present in proposition 5.18. We believe that a similar result for CV may be possible, but the calculations seem much more cumbersome compared to those for ML in subsection 5.7.2.

Hence, in the rest of subsection 5.3.2, we only address the case where $d = 1$, $p = 1$ and where the observation points $v_i + \epsilon X_i$, $1 \leq i \leq n$, $n \in \mathbb{N}^*$, are the $i + \epsilon X_i$, where X_i is uniform on $[-1, 1]$. Since $p = 1$, we have $\Psi = [\psi_{inf}, \psi_{sup}]$.

We define the Fourier transform function $\hat{s}(.)$ of a sequence s_n of $\mathbb{Z}$ by $\hat{s}(f) = \sum_{n \in \mathbb{Z}} s_n e^{inf}$ as in [Gra01]. This function is 2π periodic on $[-\pi, \pi]$.

Then, with t representing the space argument of a stationary covariance function in the notation $\frac{\partial}{\partial t}$,

- The sequence of the $K_{\psi_0}(i)$, $i \in \mathbb{Z}$, has Fourier transform f which is even and non-negative on $[-\pi, \pi]$.
- The sequence of the $\frac{\partial}{\partial \psi} K_{\psi_0}(i)$, $i \in \mathbb{Z}$, has Fourier transform f_ψ which is even on $[-\pi, \pi]$.
- The sequence of the $\frac{\partial}{\partial t} K_{\psi_0}(i) \mathbf{1}_{i \neq 0}$, $i \in \mathbb{Z}$, has Fourier transform $\mathrm{i} f_t$ which is odd and imaginary on $[-\pi, \pi]$.
- The sequence of the $\frac{\partial}{\partial t} \frac{\partial}{\partial \psi} K_{\psi_0}(i) \mathbf{1}_{i \neq 0}$, $i \in \mathbb{Z}$, has Fourier transform $\mathrm{i} f_{t,\psi}$ which is odd and imaginary on $[-\pi, \pi]$.
- The sequence of the $\frac{\partial^2}{\partial t^2} K_{\psi_0}(i) \mathbf{1}_{i \neq 0}$, $i \in \mathbb{Z}$, has Fourier transform $f_{t,t}$ which is even on $[-\pi, \pi]$.
- The sequence of the $\frac{\partial^2}{\partial t^2} \frac{\partial}{\partial \psi} K_{\psi_0}(i) \mathbf{1}_{i \neq 0}$, $i \in \mathbb{Z}$, has Fourier transform $f_{t,t,\psi}$ which is even on $[-\pi, \pi]$.

In this section we assume in condition 5.17 that all these sequences are dominated by a decreasing exponential function, so that the Fourier transforms are C^∞. This condition could be weakened, but it simplifies the proofs, and it is satisfied in our framework.

Condition 5.17. *There exists $C < \infty$ and $a > 0$ so that the sequences of general terms*

$$K_{\psi_0}(i),$$

$$\frac{\partial}{\partial \psi} K_{\psi_0}(i),$$

$$\frac{\partial}{\partial t} K_{\psi_0}(i) \mathbf{1}_{i \neq 0},$$

$$\frac{\partial}{\partial t}\frac{\partial}{\partial \psi} K_{\psi_0}(i) \mathbf{1}_{i \neq 0},$$

$$\frac{\partial^2}{\partial t^2} K_{\psi_0}(i) \mathbf{1}_{i \neq 0}$$

and

$$\frac{\partial^2}{\partial t^2}\frac{\partial}{\partial \psi} K_{\psi_0}(i) \mathbf{1}_{i \neq 0},$$

$i \in \mathbb{Z}$, are bounded by $Ce^{-a|i|}$.

For a 2π-periodic function f on $[-\pi, \pi]$, we denote by $M(f)$ the mean value of f on $[-\pi, \pi]$. Then, proposition 5.18 gives the closed form expressions of Σ_{ML}, $\Sigma_{CV,1}$, $\Sigma_{CV,2}$ and $\frac{\partial^2}{\partial \epsilon^2}\Sigma_{ML}\Big|_{\epsilon=0}$.

Proposition 5.18. *Assume that conditions 5.1 and 5.17 are verified.*

For $\epsilon = 0$,

$$\Sigma_{ML} = M\left(\frac{f_\psi^2}{f^2}\right),$$

$$\begin{aligned}\Sigma_{CV,1} &= 8M\left(\frac{1}{f}\right)^{-6} M\left(\frac{f_\psi}{f^2}\right)^2 M\left(\frac{1}{f^2}\right) \\ &\quad +8M\left(\frac{1}{f}\right)^{-4} M\left(\frac{f_\psi^2}{f^4}\right) \\ &\quad -16M\left(\frac{1}{f}\right)^{-5} M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_\psi}{f^3}\right),\end{aligned}$$

$$\Sigma_{CV,2} = 2M\left(\frac{1}{f}\right)^{-3}\left\{M\left(\frac{f_\psi^2}{f^3}\right) M\left(\frac{1}{f}\right) - M\left(\frac{f_\psi}{f^2}\right)^2\right\},$$

and

$$\begin{aligned}\frac{\partial^2}{\partial\epsilon^2}\Sigma_{ML}\bigg|_{\epsilon=0} &= \frac{4}{3}M\left(\frac{f_\psi}{f^2}\right)M\left(\frac{f_t^2\ f_\psi}{f^2}\right)\\ &\quad -\frac{8}{3}M\left(\frac{1}{f}\right)M\left(\frac{f_{t,\psi}\ f_t\ f_\psi}{f^2}\right)-\frac{8}{3}M\left(\frac{f_\psi}{f^2}\right)M\left(\frac{f_{t,\psi}\ f_t}{f}\right)\\ &\quad +\frac{4}{3}M\left(\frac{1}{f}\right)M\left(\frac{f_t^2\ f_\psi^2}{f^3}\right)+\frac{4}{3}M\left(\frac{f_\psi^2}{f^3}\right)M\left(\frac{f_t^2}{f}\right)\\ &\quad -\frac{4}{3}M\left(\frac{f_{t,t}\ f_\psi^2}{f^3}\right)\\ &\quad +\frac{4}{3}M\left(\frac{1}{f}\right)M\left(\frac{f_{t,\psi}^2}{f}\right)\\ &\quad +\frac{4}{3}M\left(\frac{f_{t,t,\psi}\ f_\psi}{f^2}\right).\end{aligned}$$

Proposition 5.18 is proved in subsection 5.7.2. Notice that there could be prospects for extending this proposition for the regular grid in dimension $d > 1$, based on results similar to those of proposition 5.41, but for multi-level Toeplitz matrices (see e.g. [Tyr96]).

An interesting remark can be made on $\Sigma_{CV,2}$. Using Cauchy-Schwartz inequality, we obtain

$$\Sigma_{CV,2} = 2M\left(\frac{1}{f}\right)^{-3}\left[M\left\{\left(\frac{f_\psi}{f^{\frac{3}{2}}}\right)^2\right\}M\left\{\left(\frac{1}{f^{\frac{1}{2}}}\right)^2\right\}-M\left\{\frac{f_\psi}{f^{\frac{3}{2}}}\frac{1}{f^{\frac{1}{2}}}\right\}^2\right]\geq 0,$$

so that the limit of the second derivative with respect to ψ of the CV criterion at ψ_0 is indeed non-negative. Furthermore, for the limit to be zero, it is necessary that $\frac{f_\psi}{f^{\frac{3}{2}}}$ be proportional to $\frac{1}{f^{\frac{1}{2}}}$, that is to say f_ψ be proportional to f. This is equivalent to $\frac{\partial K_{\psi_0}}{\partial\psi}$ being proportional to K_{ψ_0} on $\mathbb{Z}$, which happens only when around ψ_0, $K_\psi(i) = \frac{\psi}{\psi_0}K_{\psi_0}(i)$, for $i\in\mathbb{Z}$. Hence around ψ_0, ψ would be a global variance hyper-parameter. Therefore, for the regular grid in dimension one, we have shown that the asymptotic variance is positive as long as ψ is not only a global variance hyper-parameter.

5.4 Study of the asymptotic variance

The limit distributions of the ML and CV estimators only depend on the regularity parameter ϵ through the asymptotic covariance matrices in propositions 5.8 and 5.13. The aim of this section is to numerically study the influence of ϵ on these asymptotic covariance matrices. Furthermore, we aim at confirming numerically that the asymptotic variance is larger for CV than for ML.

In the rest of this section 5.4, we specifically address the cases where $d = 1$, $p = 1$ in subsections 5.4.1 and 5.4.2, $p = 2$ in subsection 5.4.3, and the distribution of the X_i, $1\leq i\leq n$, is uniform on $[-1,1]$. Furthermore, in order to compare ML and CV, we will only address the estimation of a correlation hyper-parameter. The variance is thus assumed to be known and equal to 1.

We focus on the case of the Matérn correlation function presented in chapter 2. In dimension one, we recall that this correlation model is parameterized by the correlation length ℓ and the

smoothness parameter ν. The correlation function $K_{\ell,\nu}$ is Matérn (ℓ, ν) where

$$K_{\ell,\nu}(h) = \frac{1}{\Gamma(\nu)\, 2^{\nu-1}} \left(2\sqrt{\nu}\frac{|h|}{\ell}\right)^{\nu} K_\nu \left(2\sqrt{\nu}\frac{|h|}{\ell}\right), \tag{5.11}$$

with Γ the Gamma function and K_ν the modified Bessel function of second order.

5.4.1 Small random perturbations

In our study, the two true hyper-parameters (ℓ_0, ν_0) vary over $0.3 \leq \ell_0 \leq 3$ and $0.5 \leq \nu_0 \leq 5$. We will successively address the two cases where ℓ is estimated and ν is known, and where ν is estimated and ℓ is known. It is shown in subsection 5.3.1 that for both ML and CV, the asymptotic variances are regular functions of ϵ. They are even functions of ϵ, because the distribution of ϵX_i is the same as the distribution of $-\epsilon X_i$. Hence the quantity of interest we study is the ratio of the second derivative with respect to ϵ at $\epsilon = 0$ of the asymptotic variance over its value at $\epsilon = 0$. When this quantity is negative, this means that the asymptotic variance of the hyper-parameter estimator decreases with ϵ, and therefore that an irregular sampling is more favorable for hyper-parameter estimation than a regular one. The second derivative is calculated exactly for ML, using the results of subsection 5.3.2, and is approximated by finite differences for n large for CV. Proposition 5.16 ensures that this approximation is numerically consistent (because the limits in n and the derivatives in ϵ are exchangeable).

In figure 5.2, we show the numerical results for the estimation of ℓ. First we see that the relative improvement of the estimation due to irregularity is maximum when the true correlation length ℓ_0 is small. Indeed, the inter-observation distance being 1, a correlation length of approximately 0.3 means that the observations are almost independent, making the estimation of the covariance very hard. For instance, for $\nu_0 = \frac{3}{2}$, $\ell_0 = 0.3$ and $\epsilon = 0$, the maximum correlation between two different observation points is $(1 + \frac{\sqrt{6}}{0.3}) \exp\left(-\frac{\sqrt{6}}{0.3}\right) \approx 0.0026$. Thus, the vector of n observations looks like an *iid* vector, as we illustrate in figure 5.3, making it difficult to distinguish between $\ell_0 = 0.3$ and, say, $\ell_0 = 0.2$, which would *a fortiori* also make the observation vector look like an *iid* vector.

Hence, for ℓ_0 small the irregularity of the grid creates pairs of observations that are less independent and makes the estimation possible. Indeed, for $i < j$, $|i - j + \epsilon(X_i - X_j)|$ can be smaller than $|i-j|$ when $X_i > X_j$. For instance, for $\epsilon = 0.25$, $\nu_0 = \frac{3}{2}$ and $\ell_0 = 0.3$, the maximum correlation between two different observation points is $(1 + \frac{0.5\sqrt{6}}{0.3}) \exp\left(-\frac{0.5\sqrt{6}}{0.3}\right) \approx 0.08$, to compare to 0.0026 for $\epsilon = 0$. As a conclusion, for ℓ_0 small, the benefit obtained from perturbing the regular grid is large.

For large ℓ_0, it is easier to estimate ℓ when $\epsilon = 0$, because the observation vector does not look like an *iid* vector, as illustrated in figure 5.3. Thus the relative effect of the irregularity is smaller.

Second, we observe in figure 5.2 that for ML the irregularity is always an advantage for estimation. This is not the case for CV, where the asymptotic variance can increase with ϵ. Finally, we can see that the two particular points $(\ell_0 = 0.5, \nu_0 = 5)$ and $(\ell_0 = 2.7, \nu_0 = 1)$ are particularly interesting and representative. Indeed $\ell_0 = 0.5$ and $\nu_0 = 5$ correspond to hyper-parameters for which the irregularity of the sampling has a strong and favorable impact on the estimation for ML and CV, while $\ell_0 = 2.7$ and $\nu_0 = 1$ correspond to hyper-parameters for which

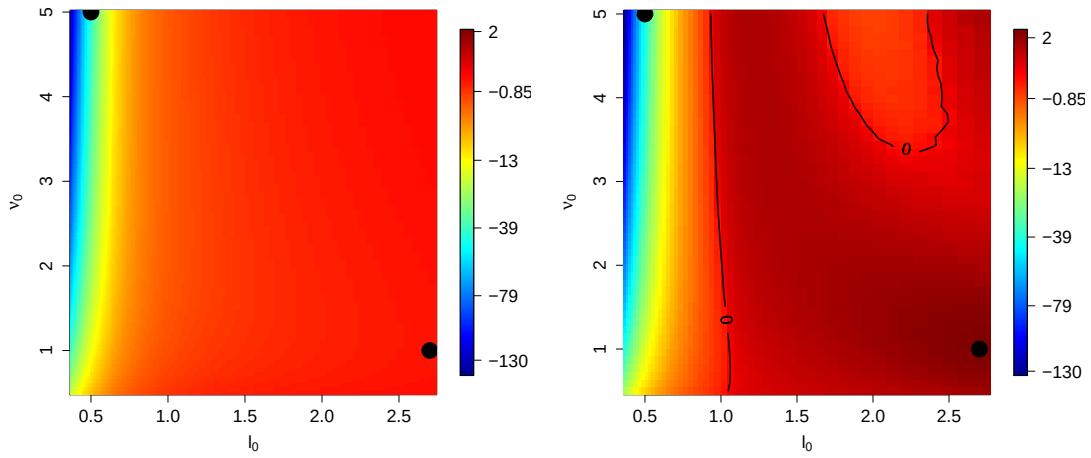

Figure 5.2: Local influence of ϵ for the estimation of the correlation length ℓ. Plot of the ratio of the second derivative of the asymptotic variance over its value at $\epsilon = 0$, for ML (left) and CV (right). The true covariance function is Matérn with varying ℓ_0 and ν_0. The advantage of perturbing the regular grid is maximum when the correlation length ℓ_0 small, i.e. when the observations are almost independent. The asymptotic variance always locally decreases with ϵ for ML (i.e. the second derivative at $\epsilon = 0$ is always negative) but not for CV. We retain the two particular points $(\ell_0 = 0.5, \nu_0 = 5)$ and $(\ell_0 = 2.7, \nu_0 = 1)$ for further investigation in subsection 5.4.2 (these are the black dots).

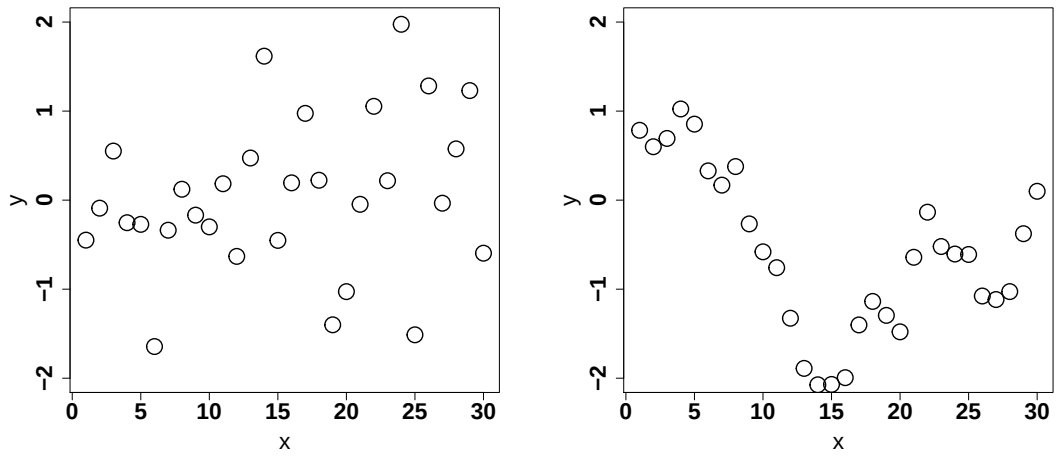

Figure 5.3: Illustration that it is more difficult to estimate ℓ when ℓ_0 is small than when ℓ_0 is large. Plot of two realizations of the Gaussian process Y on the regular grid $\{1, ..., 30\}$ with Matérn covariance function with $\nu_0 = \frac{3}{2}$ and $\ell_0 = 0.3$ (left) and $\ell_0 = 3$ (right). For $\ell_0 = 0.3$ the observation vector seems to follow an *iid* distribution. Thus it is hard to distinguish between $\ell_0 = 0.3$, and, say, $\ell_0 = 0.2$, which would *a fortiori* make the observation vector seem to follow an *iid* distribution. On the contrary, for $\ell_0 = 3$, the observation vector does not seem to follow an *iid* distribution.

the irregularity of the sampling has an unfavorable impact on the estimation for CV. We retain these two points for further global investigation for $0 \leq \epsilon \leq 0.45$ in subsection 5.4.2.

On figure 5.4, we show the numerical results for the estimation of ν. We observe that for ℓ_0 relatively small, the asymptotic variance is an increasing function of ϵ (for small ϵ). This happens approximately in the band $0.4 \leq \ell_0 \leq 0.6$, and for both ML and CV. There is a plausible explanation from this fact, which is not easy to interpret at first sight. It can be seen that for $\ell \approx 0.73$, the value of the one-dimensional Matérn covariance function at $t = 1$ is almost independent of ν for $\nu \in [1, 5]$ (see figure 2.4). As an illustration, for $\nu = 2.5$, the derivative of this value with respect to ν is -3.7×10^{-5} for a value of 0.15. When $0.4 \leq \ell_0 \leq 0.6$, ℓ_0 is small so that most of the information for estimating ν is obtained from the pairs of successive observation points. Perturbing the regular grid creates pairs of successive observation points $i + \epsilon x_i, i + 1 + \epsilon x_{i+1}$ verifying $\frac{|1+\epsilon(x_{i+1}-x_i)|}{\ell_0} \approx \frac{1}{0.73}$, so that the correlation of the two observations becomes almost independent of ν. Thus, due to a specificity of the Matérn covariance function, decreasing the distance between two successive observation points unintuitively removes information on ν.

For $0.6 \leq \ell_0 \leq 0.8$ and $\nu_0 \geq 2$, the relative improvement is maximum. This is explained the same way as above, this time the case $\epsilon = 0$ yields successive observation points for which the correlation is independent of ν, and increasing ϵ changes the distance between two successive observation points, making the correlation of the observations dependent of ν.

In the case $\ell_0 \geq 0.8$, there is no more impact of the specificity of the case $\ell_0 \approx 0.73$ and the improvement of the estimation when ϵ increases remains significant, though smaller. Finally, we see the three particular points $(\ell_0 = 0.5, \nu_0 = 2.5)$, $(\ell_0 = 0.7, \nu_0 = 2.5)$ and $(\ell_0 = 2.7, \nu_0 = 2.5)$ as representative of the discussion above, and we retain them for further global investigation for $0 \leq \epsilon \leq 0.45$ in subsection 5.4.2.

5.4.2 Large random perturbations

On figures 5.5 and 5.6, we plot the ratio of the asymptotic variance for $\epsilon = 0$ over the asymptotic variance for $\epsilon = 0.45$, with varying ℓ_0 and ν_0, for ML and CV and in the two cases where ℓ is estimated and ν known and conversely. We observe that this ratio is always larger than one for ML, that is strong perturbations of the regular grid are always beneficial to ML estimation. This is the most important numerical conclusion of this section 5.4. As ML is the preferable method to use in the well-specified case addressed here, we reformulate this conclusion by saying that, in our experiments, using pairs of closely spaced observation points is always beneficial for covariance hyper-parameter estimation compared to evenly spaced observation points. This is an important practical conclusion, that is in agreement with the references [Ste99] and [ZZ06] discussed in section 5.1.

For CV, on the contrary, we exhibit cases for which strong perturbations of the regular grid decrease the accuracy of the estimation of ℓ. This can be due to the fact that the Leave-One-Out errors in the CV functional (3.12) are unnormalized. Hence, when the regular grid is perturbed, roughly speaking, error terms concerning observation points with close neighbors are small, while error terms concerning observation points without close neighbors are large. Hence, the CV functional mainly depends on the large error terms and hence has a larger variance. This increases the variance of the CV estimator minimizing it.

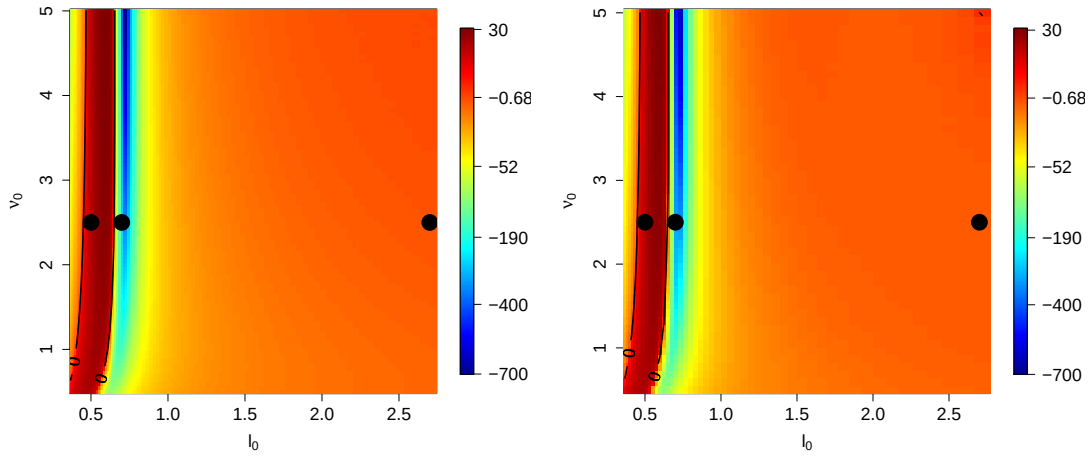

Figure 5.4: Same setting as figure 5.2, but for the estimation of ν. For approximately $0.4 \leq \ell_0 \leq 0.6$, the estimation is damaged by locally perturbing the regular grid. This is because of a particularity of the Matérn covariance function $K_{\ell,\nu}(t)$ at $t = 1$, for $\ell \approx 0.73$. For $0.6 \leq \ell_0 \leq 0.8$, the improvement of the estimation is maximum, and remains positive for larger ℓ_0. We retain the three particular points $(\ell_0 = 0.5, \nu_0 = 2.5)$, $(\ell_0 = 0.7, \nu_0 = 2.5)$ and $(\ell_0 = 2.7, \nu_0 = 2.5)$ for further investigation in subsection 5.4.2.

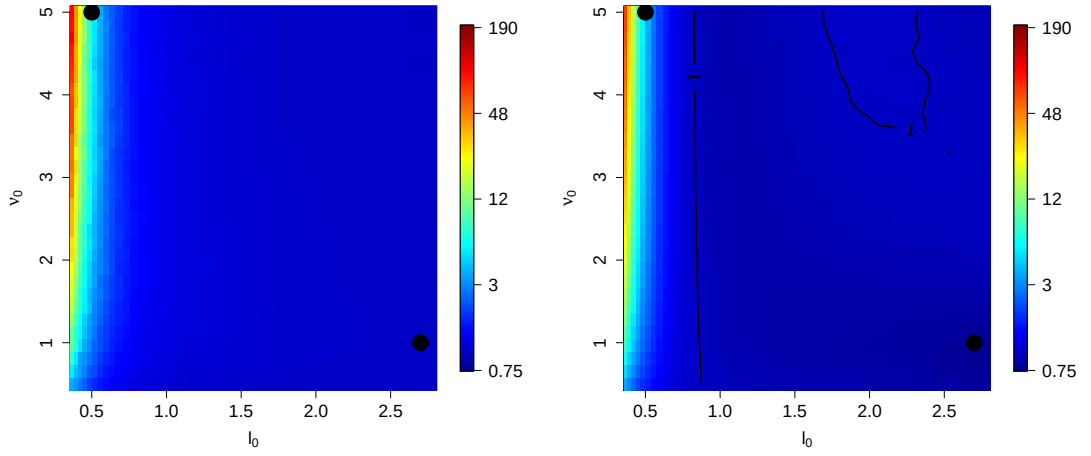

Figure 5.5: Estimation of ℓ. Plot of the ratio of the asymptotic variance for $\epsilon = 0$ over the asymptotic variance for $\epsilon = 0.45$ for ML (left) and CV (right). The true covariance function is Matérn with varying ℓ_0 and ν_0. The ML estimation is always improved by perturbing the regular grid, while the CV estimation can be damaged by perturbing the regular grid. We retain the two particular points $(\ell_0 = 0.5, \nu_0 = 5)$ and $(\ell_0 = 2.7, \nu_0 = 1)$ for further investigation below in this subsection 5.4 (these are the black dots).

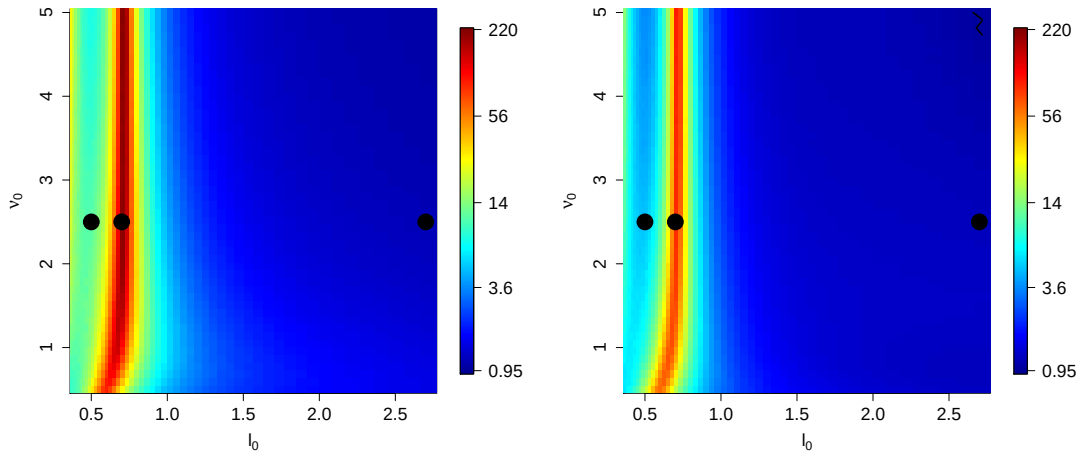

Figure 5.6: Same setting as in figure 5.5, but for the estimation of ν. The ML and CV estimations benefit from an irregular sampling. We retain the three particular points $(\ell_0 = 0.5, \nu_0 = 2.5)$, $(\ell_0 = 0.7, \nu_0 = 2.5)$ and $(\ell_0 = 2.7, \nu_0 = 2.5)$ for further investigation below in this subsection 5.4.

We now consider the five particular points that we have discussed in subsection 5.4.1: $(\ell_0 = 0.5, \nu_0 = 5)$ and $(\ell_0 = 2.7, \nu_0 = 1)$ for the estimation of ℓ and $(\ell_0 = 0.5, \nu_0 = 2.5)$, $(\ell_0 = 0.7, \nu_0 = 2.5)$ and $(\ell_0 = 2.7, \nu_0 = 2.5)$ for the estimation of ν. For these particular points, we plot the asymptotic variances of propositions 5.8 and 5.13 as functions of ϵ for $-0.45 \leq \epsilon \leq 0.45$. The asymptotic variances are even functions of ϵ since $(\epsilon X_i)_{1 \leq i \leq n}$ has the same distribution as $(-\epsilon X_i)_{1 \leq i \leq n}$. Nevertheless, they are approximated by empirical means of *iid* realizations of the random traces in propositions 5.8 and 5.12, for n large enough. Hence, the functions we plot are not exactly even. The fact that they are almost even is a graphical verification that the random fluctuations of the results of the calculations, for finite (but large) n, are very small. We also plot the second-order Taylor-series expansion given by the value at $\epsilon = 0$ and the second derivative at $\epsilon = 0$.

In figure 5.7, we show the numerical results for the estimation of ℓ with $\ell_0 = 0.5$, $\nu_0 = 5$. The first observation is that the asymptotic variance is slightly larger for CV than for ML. This is a confirmation of what we expected: we address a well-specified case, so that the asymptotic variance of ML is the almost sure limit of the Cramér-Rao bound. Therefore, this observation turns out to be true in all the subsection, and we will not comment on it anymore. We see that, for both ML and CV, the improvement of the estimation given by the irregularity of the spatial sampling is true for all values of ϵ. One can indeed gain up to a factor six for the asymptotic variances. This is explained by the reason mentioned in subsection 5.4.1, for ℓ_0 small, increasing ϵ yields pairs of observations that become dependent, and hence give information on the covariance structure.

In figure 5.8, we show the numerical results for the estimation of ℓ with $\ell_0 = 2.7$, $\nu_0 = 1$. For ML, there is a slight improvement of the estimation with the irregularity of the spatial sampling. However, for CV, there is a significant degradation of the estimation. Hence the irregularity of the spatial sampling has more relative influence on CV than on ML. Finally, the advantage of

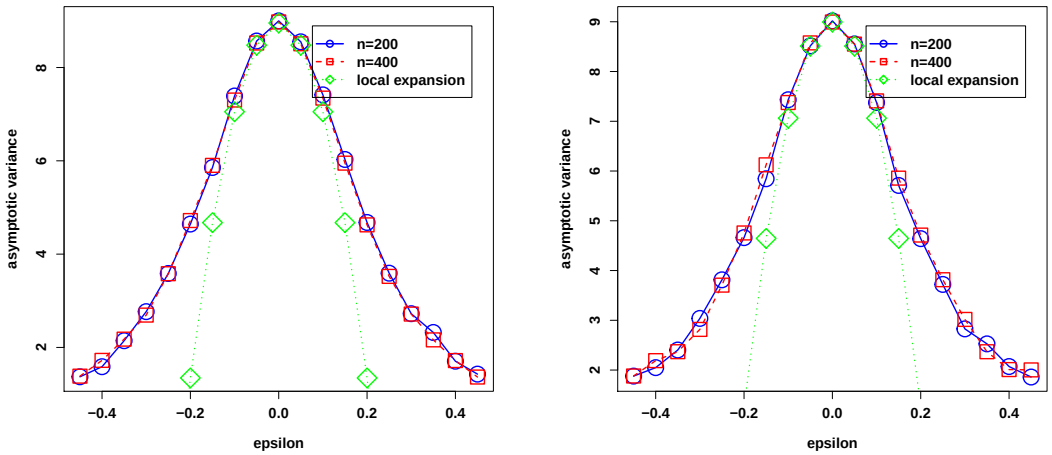

Figure 5.7: Global influence of ϵ for the estimation of the correlation length ℓ. Plot of the asymptotic variance for ML (left) and CV (right), calculated with varying n, and of the second order Taylor series expansion given by the value at $\epsilon = 0$ and the second derivative at $\epsilon = 0$. The true covariance function is Matérn with $\ell_0 = 0.5$ and $\nu_0 = 5$. The asymptotic variance is larger for CV than for ML. The irregularity of the spatial sampling globally improves the estimation for both ML and CV.

ML over CV for the estimation is by a factor seven, contrary to the case $\ell_0 = 0.5$, where this factor was close to one.

On figure 5.9, we show the numerical results for the estimation of ν with $(\ell_0 = 0.5, \nu_0 = 2.5)$. The numerical results are similar for ML and CV. For ϵ small, the asymptotic variance is very large, because, ℓ_0 being small, the observations are almost independent, as the observation points are further apart than the correlation length, making inference on the dependence structure very difficult. We see that, for $\epsilon = 0$, the asymptotic variance is several orders of magnitude larger than for the estimation of ℓ in figure 5.7, where ℓ_0 has the same value. Indeed, in the Matérn model, ν is a smoothness parameter, and its estimation is very sensitive to the absence of observation points with small spacing. We observe, as discussed in figure 5.4, that for $\epsilon \in [0, 0.2]$, the asymptotic variance increases with ϵ because pairs of observation points can reach the state where the covariance of the two observations is almost independent of ν. For $\epsilon \in [0.2, 0.5)$, a threshold is reached where pairs of subsequently dependent observations start to appear, greatly reducing the asymptotic variance for the estimation of ν.

On figure 5.10, we show the numerical results for the estimation of ν with $(\ell_0 = 0.7, \nu_0 = 2.5)$. The numerical results are similar for ML and CV. Similarly to figure 5.9, the asymptotic variance is very large, because the observations are almost independent. For $\epsilon = 0$, it is even larger than in figure 5.7 because we are in the state where the covariance between two successive observations is almost independent of ν. As an illustration, for $\ell = 0.7$ and $\nu = 2.5$, the derivative of this covariance with respect to ν is -1.3×10^{-3} for a value of 0.13 (1% relative variation), while for $\ell = 0.5$ and $\nu = 2.5$, this derivative is -5×10^{-3} for a value of 0.037 (13% relative variation). Hence, the asymptotic variance is globally decreasing with ϵ and the decrease is very strong for

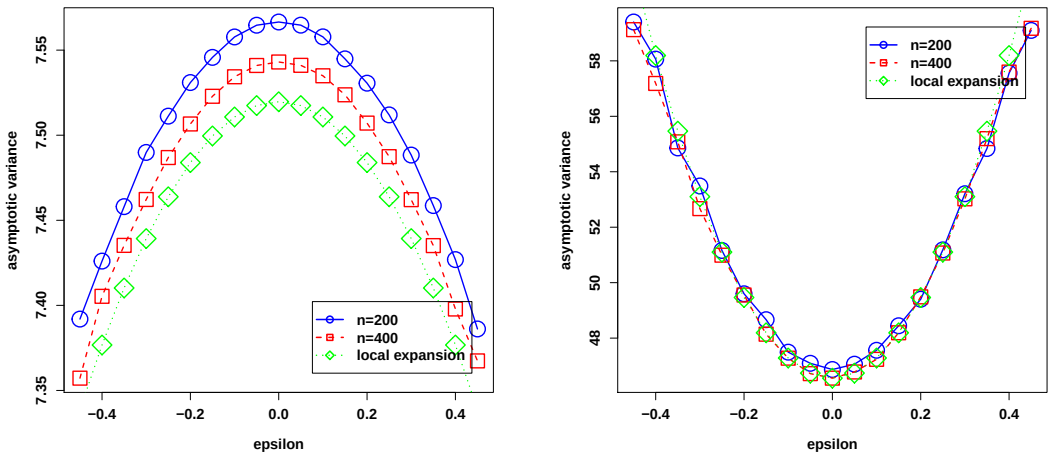

Figure 5.8: Same setting as in figure 5.7 but with $\ell_0 = 2.7$ and $\nu_0 = 1$. The irregularity of the spatial sampling slightly improves ML estimation but degrades CV estimation.

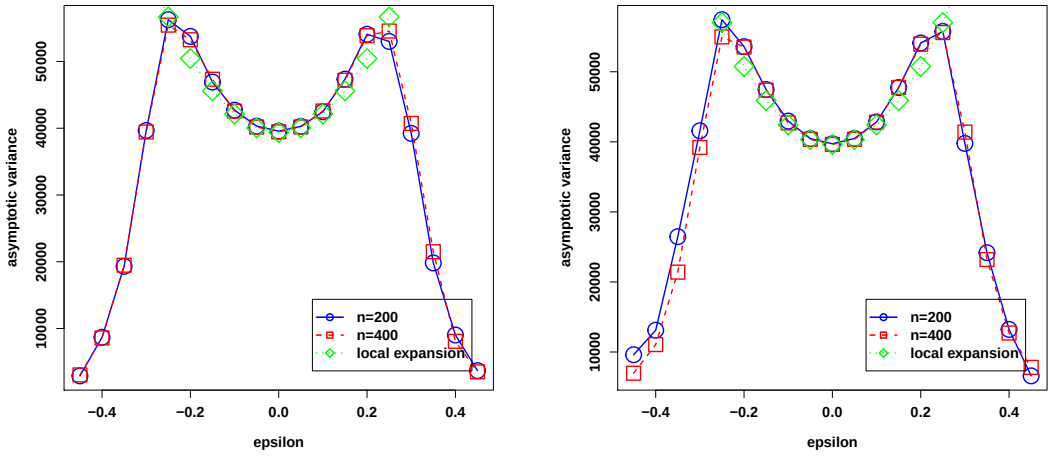

Figure 5.9: Same setting as in figure 5.7 but for the estimation of ν and with $\ell_0 = 0.5$ and $\nu_0 = 2.5$. Results are similar for ML and CV. When $\epsilon = 0$, the estimation is difficult because the observations are almost independent. For $\epsilon \in [0, 0.2]$, because of a specificity of the Matérn covariance model $K_{\ell,\nu}(t)$ at $t = 1$, for $\ell \approx 0.73$, the asymptotic variance increases with ϵ, as we have discussed in figure 5.4. The asymptotic variance decreases with ϵ for $\epsilon \in [0.2, 0.5]$, because pairs of dependent observations start to appear.

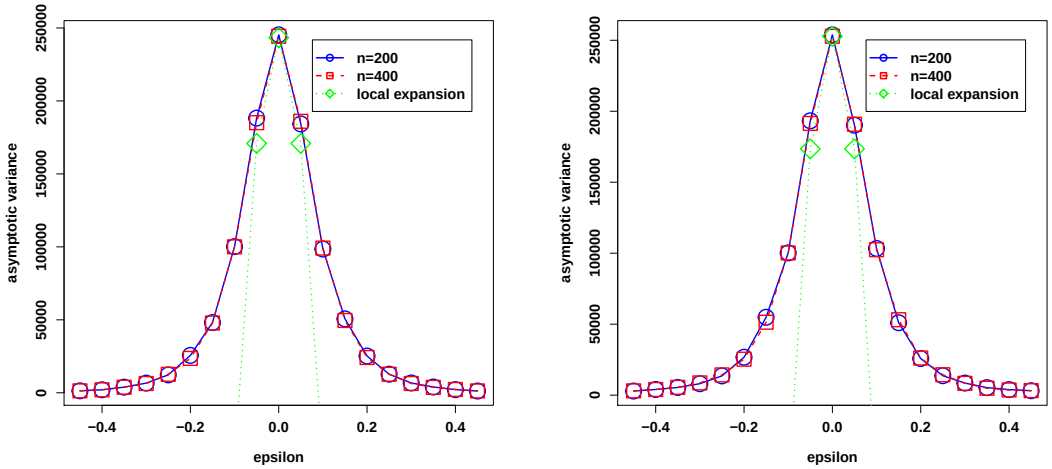

Figure 5.10: Same setting as in figure 5.7 but for the estimation of ν and with $\ell_0 = 0.7$ and $\nu_0 = 2.5$. Results are similar for ML and CV. When $\epsilon = 0$, the estimation is difficult because the observations are almost independent. It is even more difficult than for figure 5.9, although ℓ_0 is smaller in figure 5.9, because of the specificity of the Matérn covariance model $K_{\ell,\nu}(t)$ at $t = 1$, for $\ell \approx 0.73$, that we have discussed in figure 5.4. The estimation is easier for ϵ large, where pairs of dependent observations start to appear.

small ϵ. The variance is several orders of magnitude smaller for large ϵ, where pairs of dependent observations start to appear.

In figure 5.11, we show the numerical results for the estimation of ν with $\ell_0 = 2.7$, $\nu_0 = 2.5$. For both ML and CV, there is a global improvement of the estimation with the irregularity of the spatial sampling. Moreover, the advantage of ML over CV for the estimation is by a factor seven, contrary to figures 5.9 and 5.10 where this factor was close to one.

5.4.3 Estimating both the correlation length and the smoothness parameter

In this subsection 5.4.3, the case of the joint estimation of ℓ and ν is addressed. We denote, for ML and CV, V_ℓ, V_ν and $C_{\ell,\nu}$, the asymptotic variances of $\sqrt{n}\hat{\ell}$ and $\sqrt{n}\hat{\nu}$ and the asymptotic covariance of $\sqrt{n}\hat{\ell}$ and $\sqrt{n}\hat{\nu}$ (propositions 5.8 and 5.13).

Since we here address 2×2 covariance matrices, the impact of the irregularity parameter ϵ on the estimation is now more complex to assess. For instance, increasing ϵ could increase V_ℓ and at the same time decrease V_ν. Thus, it is desirable to build scalar criteria, defined in terms of V_ℓ, V_ν and $C_{\ell,\nu}$, measuring the quality of the estimation. In [ZZ06], the criterion used is the average, over a prior distribution on (ℓ_0, ν_0), of $\log(V_\ell V_\nu - C_{\ell,\nu}^2)$, that is the averaged logarithm of the determinant of the covariance matrix. This criterion corresponds to D-optimality in standard linear regression with uncorrelated errors, as noted in [ZZ06]. In our case, we know the true (ℓ_0, ν_0), so that the Bayesian average is not needed. The first scalar criterion we study is thus $D_{\ell,\nu} := V_\ell V_\nu - C_{\ell,\nu}^2$. This criterion is interpreted as a general objective-free estimation

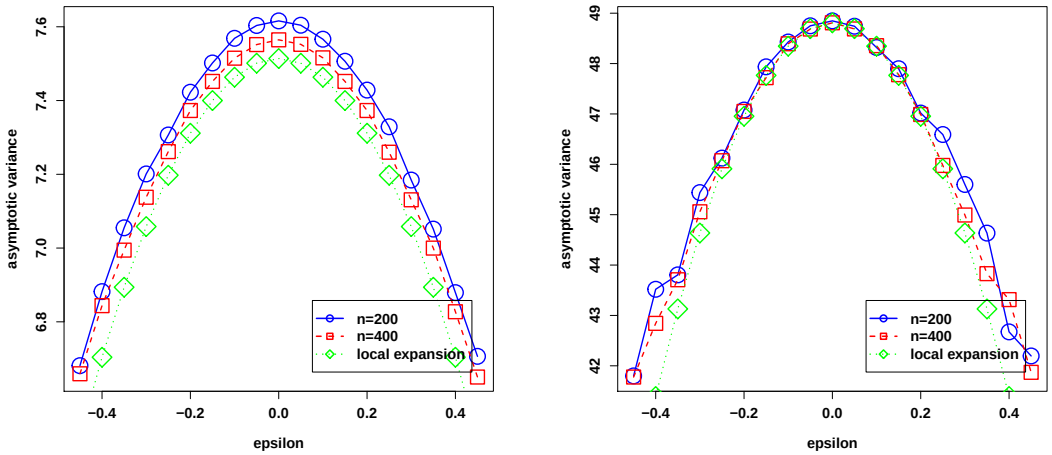

Figure 5.11: Same setting as in figure 5.7 but for the estimation of ν and with $\ell_0 = 2.7$ and $\nu_0 = 2.5$. For both ML and CV, there is a global improvement of the estimation with the irregularity of the spatial sampling. ML has a substantial advantage over CV for the estimation.

criterion, in the sense that the impact of the estimation on Kriging predictions that would be made afterward, on new input points, is not directly addressed in $D_{\ell,\nu}$.

One could build other scalar criteria, explicitly addressing the impact of the covariance function estimation error, on the quality of the Kriging predictions that are made afterward. In [ZZ06], the criterion studied for the prediction error is the integral over the prediction domain of $\mathbb{E}\left[\left(\hat{y}_{\theta_0}(t) - \hat{y}_{\hat{\theta}}(t)\right)^2\right]$, where $\hat{y}_\theta(t)$ is the prediction (2.9) of $Y(t)$, from the observation vector, and under covariance function K_θ. This criterion is the difference of integrated prediction mean square error, between the estimated and true covariance functions, see (3.19). In [Abt99] and in [ZZ06], two different asymptotic approximations of this criterion are studied. In [ZZ06], another criterion, focusing on the accuracy of the Kriging predictive variances built from $\hat{\theta}$, is also treated, together with a corresponding asymptotic approximation. In chapter 6, we will also define a criterion for the accuracy of the Kriging predictive variances, obtained from an estimator of the variance hyper-parameter, when the correlation function is fixed and misspecified. Since we specifically address the case of Kriging prediction in the asymptotic framework addressed here in section 5.5, we refer to [Abt99], [ZZ06] and chapter 6 for details on the aforementioned criteria. In this subsection 5.4.3, we study the estimation criteria V_ℓ, V_ν, $C_{\ell,\nu}$ and $D_{\ell,\nu}$.

In figure 5.12, we consider the ML estimation, with varying (ℓ_0, ν_0). We study the ratio of V_ℓ, V_ν and $D_{\ell,\nu}$, between $\epsilon = 0$ and $\epsilon = 0.45$. We first observe that V_ν is always smaller for $\epsilon = 0.45$ than for $\epsilon = 0$, that is to say there is an improvement of the estimation of ν when using a strongly irregular sampling. For V_ℓ, this is the same, except in a thin band around $\ell_0 \approx 0.73$. Our explanation for this fact is the same as for a similar singularity in figure 5.4. For $\ell_0 = 0.73$ and $\epsilon = 0$, the correlation between two successive points is approximately only a function of ℓ. For instance, the derivative of this correlation with respect to ν at $\ell = 0.73, \nu = 2.5$ is -3.7×10^{-5} for a correlation of 0.15. Thus, the very large uncertainty on ν has no negative impact on the information brought by the pairs of successive observation points on ℓ for $\epsilon = 0$.

These pairs of successive points bring most of the information on the covariance function, since ℓ_0 is small. When $\epsilon = 0.45$, this favorable case is broken by the random perturbations, and the large uncertainty on ν has a negative impact on the estimation of ℓ, even when considering the pairs of successive observation points.

Nevertheless, in the band around $\ell_0 \approx 0.73$, when going from $\epsilon = 0$ to $\epsilon = 0.45$, the improvement of the estimation of ν is much stronger than the degradation of the estimation of ℓ. This is confirmed by the plot of $D_{\ell,\nu}$, which always decreases when going from $\epsilon = 0$ to $\epsilon = 0.45$. Thus, we confirm our global conclusion of subsection 5.4.2: strong perturbations of the regular grid create pairs of observation points with small spacing, which is always beneficial for ML in the cases we address.

Finally, notice that we have discussed a case where the estimation of a covariance hyper-parameter is degraded, while the estimation of the other one is improved. This justifies the use of scalar criteria of the estimation, such as $D_{\ell,\nu}$, or the ones related with prediction discussed above.

We retain the particular point $(\ell_0 = 0.73, \nu_0 = 2.5)$, that corresponds to the case where going from $\epsilon = 0$ to $\epsilon = 0.45$ decreases V_ν and increases V_ℓ, for further global investigation in figure 5.14.

In figure 5.13, we address the same setting as in figure 5.12, but for the CV estimation. We observe that going from $\epsilon = 0$ to $\epsilon = 0.45$ can increase $D_{\ell,\nu}$. This is a confirmation of what was observed in figure 5.5: strong irregularities of the spatial sampling can globally damage the CV estimation. The justification is the same as before: the LOO error variances become heterogeneous when the regular grid is perturbed.

We also observe an hybrid case, in which the estimation of ℓ and ν is improved by the irregularity, but the determinant of their asymptotic covariance matrix increases, because the absolute value of their asymptotic covariance decreases. This case happens for instance around the point $(\ell_0 = 1.7, \nu_0 = 5)$, that we retain for a further global investigation in figure 5.15.

In figure 5.14, for $\ell_0 = 0.73, \nu_0 = 2.5$ and for ML, we plot V_ℓ, V_ν and $D_{\ell,\nu}$ with respect to ϵ, for $\epsilon \in [0, 0.45]$. We confirm that when ϵ increases, the decrease of V_ν is much stronger than the increase of V_ℓ. As a result, there is a strong decrease of $D_{\ell,\nu}$. This is a confirmation of our main conclusion on the impact of the spatial sampling on the estimation: using pairs of closely spaced observation points improves the ML estimation.

In figure 5.15, for $\ell_0 = 1.7, \nu_0 = 5$ and for CV, we plot V_ℓ, V_ν, $C_{\ell,\nu}$ and $D_{\ell,\nu}$ with respect to ϵ, for $\epsilon \in [0, 0.45]$. We observe the particular case mentioned in figure 5.13, in which the estimation of ℓ and ν is improved by the irregularity, but the determinant of their asymptotic covariance matrix increases, because the absolute value of their asymptotic covariance decreases. This particular case is again a confirmation that the criteria V_ℓ and V_ν can be insufficient for evaluating the impact of the irregularity on the estimation, in a case of joint estimation.

5.4.4 Discussion

We have seen that local perturbations of the regular grid can damage both the ML and the CV estimation (figure 5.4). The CV estimation can even be damaged for strong perturbations of the regular grid (figure 5.5). This can be due to the fact that the Leave-One-Out errors

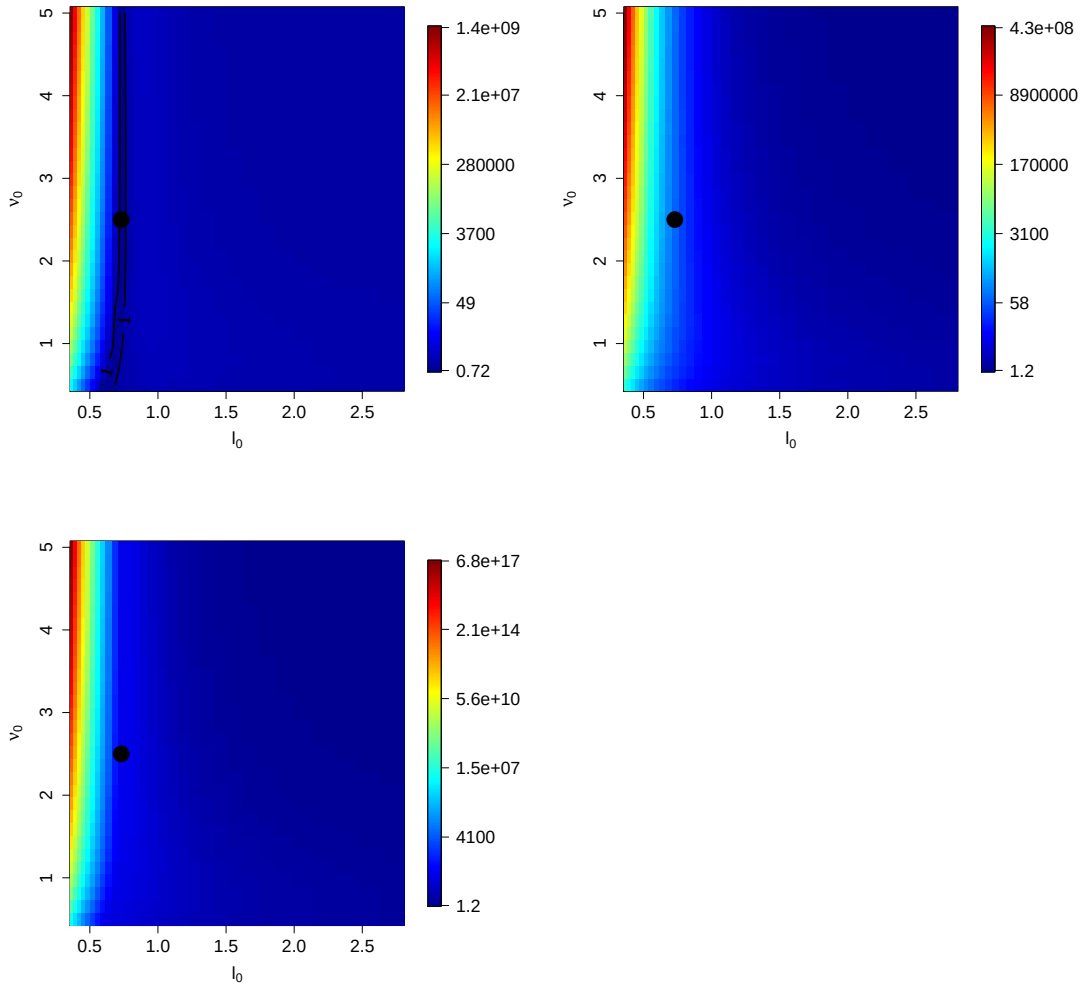

Figure 5.12: For ML, plot of the ratio, between $\epsilon = 0$ and $\epsilon = 0.45$, of V_ℓ (top-left), V_ν (top-right) and $D_{\ell,\nu}$ (bottom). The true covariance function is Matérn with varying ℓ_0 and ν_0. We jointly estimate ℓ and ν. The asymptotic variance V_ℓ increases with ϵ for ℓ_0 in a thin band, because this thin bands correspond to the case when, for $\epsilon = 0$, the large uncertainty on ν has no negative impact for the estimation of ℓ. The asymptotic variance V_ν always decreases with ϵ. Furthermore, when V_ℓ increases with ϵ, V_ν decreases considerably more. As a consequence, $D_{\ell,\nu}$ always decreases with ϵ, meaning that the joint estimation of ℓ and ν always benefits from an irregular sampling. We retain the particular point $(\ell_0 = 0.73, \nu_0 = 2.5)$ for further investigation below in this subsection 5.4.3.

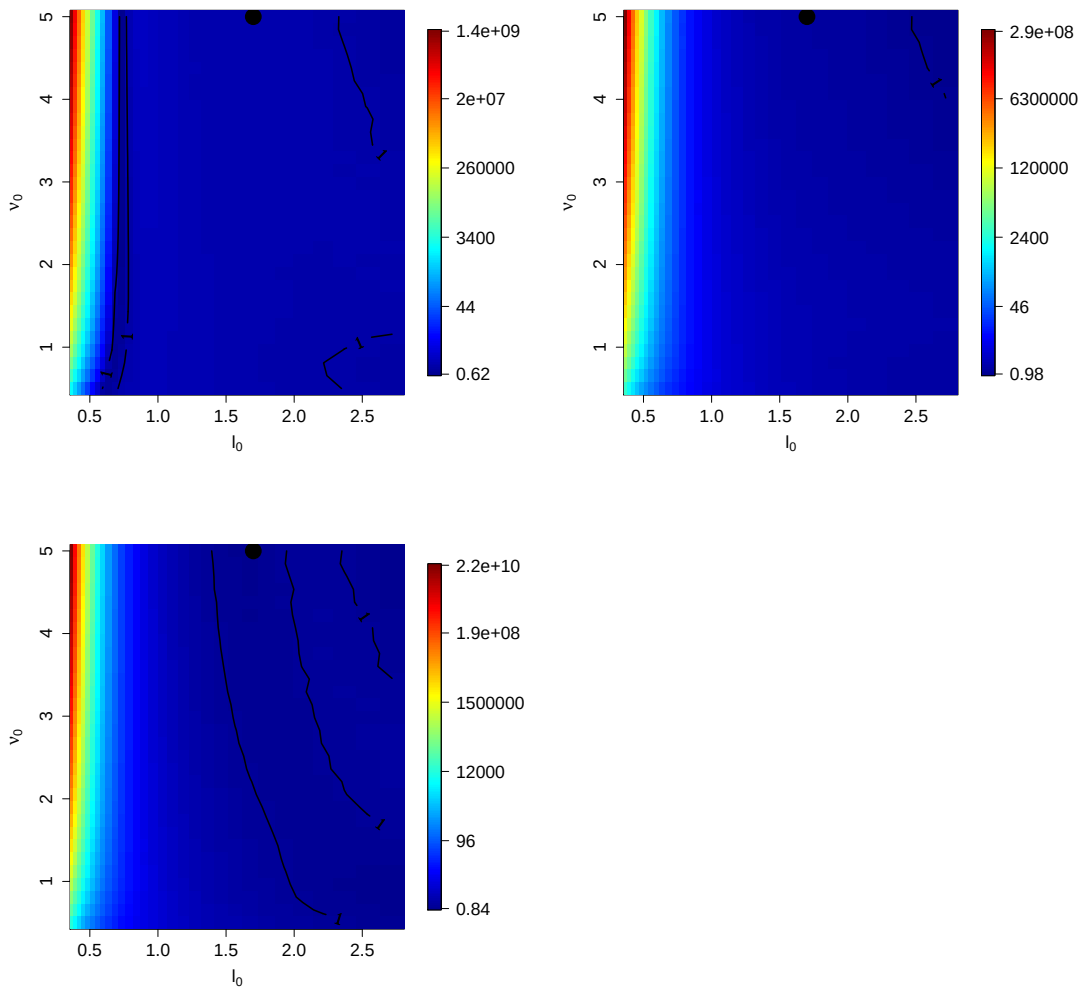

Figure 5.13: Same setting as in figure 5.12 but for CV. The CV estimation can be damaged by the irregularity of the sampling. We retain the particular point $(\ell_0 = 1.7, \nu_0 = 5)$ for further investigation below in this subsection 5.4.3.

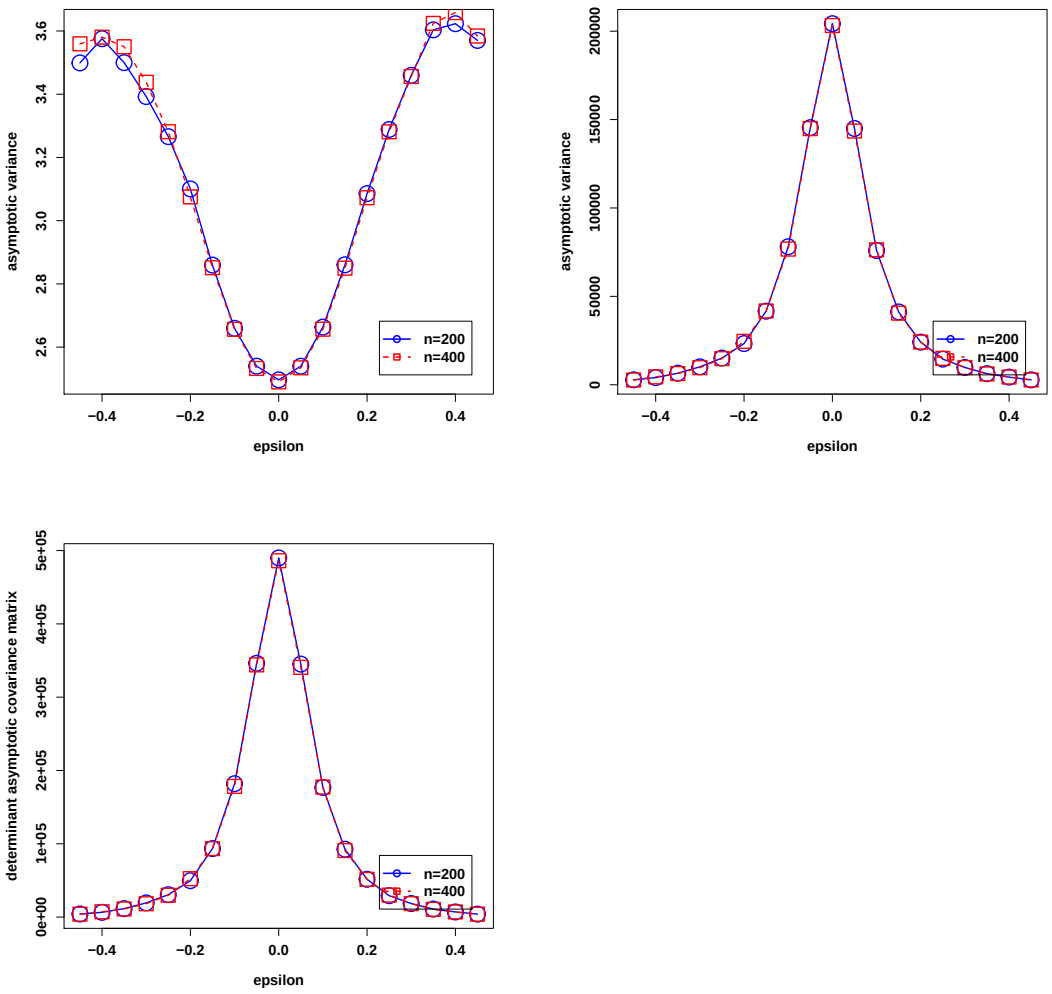

Figure 5.14: Joint estimation of ℓ and ν by ML. $\ell_0 = 0.73$ and $\nu_0 = 2.5$. Plot of V_ℓ (top-left), V_ν (top-right) and $D_{\ell,\nu}$ (bottom) with respect to ϵ. Increasing the irregularity parameter ϵ globally improves the joint ML estimation of ℓ and ν.

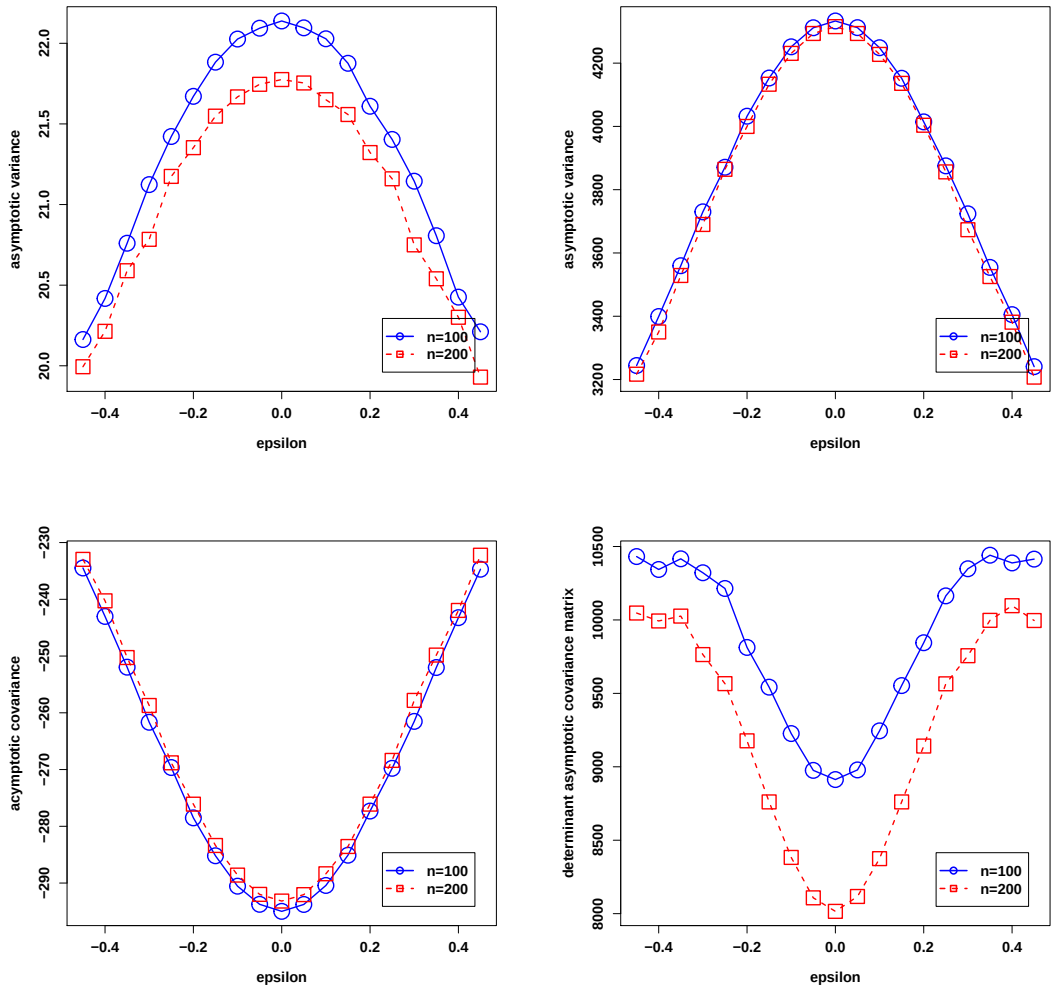

Figure 5.15: Joint estimation of ℓ and ν by CV. $\ell_0 = 1.7$ and $\nu_0 = 5$. Plot of V_ℓ (top-left), V_ν (top-right), $C_{\ell,\nu}$ (bottom-left) and $D_{\ell,\nu}$ (bottom-right) with respect to ϵ. The CV estimation of ℓ and ν is improved by the irregularity, but the determinant of their asymptotic covariance matrix increases, because the absolute value of their asymptotic covariance decreases.

in the CV functional (3.12) are unnormalized. Hence, with $\epsilon \neq 0$, roughly speaking, error terms concerning observation points with close neighbors are small, while error terms concerning observation points without close neighbors are large. Hence, the CV functional mainly depends on the large error terms and hence has a larger variance.

Our main conclusion is that strong perturbations of the regular grid ($\epsilon = 0.45$) are beneficial to the ML estimation in all the cases we have addressed (figures 5.5, 5.6, 5.12). Furthermore, ML is shown to be the preferable estimator in the well-specified case addressed here. This main result is in agreement with the references [Ste99, ZZ06, JDLI08] discussed in section 5.1. The global conclusion is that using groups of observation points with small spacing, compared to the observation point density in the prediction domain, is beneficial for estimation.

5.5 Analysis of the Kriging prediction

The asymptotic analysis of the influence of the spatial sampling on the covariance hyper-parameter estimation being complete, we now address the case of the Kriging prediction error, and its interaction with the covariance function estimation. In short words, we study Kriging prediction with estimated covariance hyper-parameters [Ste99].

In subsection 5.5.1, we show that any fixed, constant, covariance function error has a non-zero asymptotic impact on the prediction error. This fact is interesting in that the conclusion is different in a fixed-domain asymptotic context, for which we have discussed in chapter 4 that there exist non-microergodic covariance hyper-parameters that have no asymptotic influence on prediction.

In subsection 5.5.2, we show that, in the expansion-domain asymptotic context we address, the covariance function estimation procedure has, however, no impact on the prediction error, as long as it is consistent. Thus, the prediction error is a new criterion for the spatial sampling, that is independent of the estimation criteria we address in section 5.4. In subsection 5.5.3, we study numerically, still in the case of the Matérn covariance function, the impact of the regularity parameter ϵ on the mean square prediction error on the prediction domain.

5.5.1 Asymptotic influence of covariance hyper-parameter misspecification on prediction

In proposition 5.19, we show that the misspecification of correlation hyper-parameters has an asymptotic influence on the prediction errors. Indeed, the difference of the asymptotic Leave-One-Out mean square errors, between incorrect and correct covariance hyper-parameters, is lower and upper bounded by finite positive constants times the integrated square difference between the two correlation functions.

Proposition 5.19. *Assume that condition 5.1 is satisfied and that for all $\boldsymbol{\psi} \in \Psi$, $K_{\boldsymbol{\psi}}(0) = 1$.*

Let, for $1 \leq i \leq n$, $\hat{y}_{i,\boldsymbol{\psi}} := \mathbb{E}_{\boldsymbol{\psi}|\boldsymbol{X}}\left(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\right)$ be the Kriging Leave-One-Out prediction of y_i with covariance hyper-parameters $\boldsymbol{\psi}$. We then denote

$$D_p(\boldsymbol{\psi}, \boldsymbol{\psi}^{(0)}) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{y}_{i,\boldsymbol{\psi}}\right\}^2\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{y}_{i,\boldsymbol{\psi}^{(0)}}\right\}^2\right].$$

Then there exist constants $0 < A < B < +\infty$*, independent of* $\boldsymbol{\psi}$*, so that, for* $\epsilon = 0$

$$A \sum_{\boldsymbol{v} \in \mathbb{Z}^d} \left\{ K_{\boldsymbol{\psi}}(\boldsymbol{v}) - K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) \right\}^2 \leq \varliminf_{n \to +\infty} D_p(\boldsymbol{\psi}, \boldsymbol{\psi}^{(0)})$$

and

$$\varlimsup_{n \to +\infty} D_p(\boldsymbol{\psi}, \boldsymbol{\psi}^{(0)}) \leq B \sum_{\boldsymbol{v} \in \mathbb{Z}^d} \left\{ K_{\boldsymbol{\psi}}(\boldsymbol{v}) - K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) \right\}^2 .$$

For $\epsilon \neq 0$*, with* D_ϵ *as in* (5.6)*,*

$$A \int_{D_\epsilon} \left\{ K_{\boldsymbol{\psi}}(t) - K_{\boldsymbol{\psi}^{(0)}}(t) \right\}^2 dt \leq \varliminf_{n \to +\infty} D_p(\boldsymbol{\psi}, \boldsymbol{\psi}^{(0)})$$

and

$$\varlimsup_{n \to +\infty} D_p(\boldsymbol{\psi}, \boldsymbol{\psi}^{(0)}) \leq B \int_{D_\epsilon} \left\{ K_{\boldsymbol{\psi}}(t) - K_{\boldsymbol{\psi}^{(0)}}(t) \right\}^2 dt.$$

Proof. The lower-bound is showed in the proof of proposition 5.11. The upper-bound is obtained with similar techniques. □

In proposition 5.20, we study the case of N^d points with uniform distribution on $[0, N]^d$, with the constraint that there exists a minimum distance between two different observation points. We show that the asymptotic difference of integrated prediction MSE, between the incorrect and true hyper-parameters, is lower-bounded by a finite constant times the integrated square difference between the two associated correlation functions.

Proposition 5.20. *Consider the parameterization* (5.1) *of the Gaussian process* Y*. Assume that for all* $\boldsymbol{\psi} \in \Psi$*,* $K_{\boldsymbol{\psi}}(0) = 1$*.*

Let $\delta > 0$*. Assume that for* $n \in \mathbb{N}$*, the observation points are the* $X_1, ..., X_{N^d}$ *and follow an iid uniform distribution on* $[0, N]^d$*, conditionally to the constraint that, for each* $i \neq j$*,* $|X_i - X_j|_\infty \geq \delta$*.*

Let

$$MSE_{\boldsymbol{\psi}} := \frac{1}{N^d} \int_{[0,N]^d} \left(Y(\boldsymbol{x}) - \hat{y}_{\boldsymbol{\psi}}(\boldsymbol{x}) \right)^2 d\boldsymbol{x},$$

with $\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{x})$ *the Kriging prediction of* $Y(\boldsymbol{x})$ *according to the covariance function* $K_{\boldsymbol{\psi}}(\boldsymbol{x})$ *and the observation points* $Y(X_1), ..., Y(X_{N^d})$*.*

Then, there exists a constant $A > 0$ *so that for any* $\boldsymbol{\psi} \in \Psi$*,*

$$A \int_{\mathbb{R}^d \setminus [-\delta, \delta]^d} \left(K_{\boldsymbol{\psi}}(\boldsymbol{x}) - K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{x}) \right)^2 d\boldsymbol{x} \leq \varliminf_{N \to +\infty} \mathbb{E}(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}).$$

Proposition 5.20 is proved in subsection 5.7.3.

5.5.2 Influence of covariance hyper-parameter estimation on prediction

In proposition 5.21, proved in subsection 5.7.3, we address the influence of covariance hyper-parameter estimation on prediction.

Proposition 5.21. *Assume that condition 5.1 is satisfied and that the Gaussian process Y, with covariance function $K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})$, yields almost surely continuous trajectories. Assume also that for every $\boldsymbol{\psi} \in \Psi$, $1 \leq i \leq p$, $\frac{\partial}{\partial \psi_i} K_{\boldsymbol{\psi}}(\boldsymbol{t})$ is continuous with respect to $\boldsymbol{t}$. Let, for $n \in \mathbb{N}$, the observation points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ be obtained from the randomly perturbed regular grid of section 5.2. Let $\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t})$ be the Kriging prediction of the Gaussian process Y at $\boldsymbol{t}$, under correlation function $K_{\boldsymbol{\psi}}$ and given the observations $y_1, ..., y_n$. For any n, let $N_{1,n}$ so that $N_{1,n}^d \leq n < (N_{1,n}+1)^d$. Define*

$$E_{\epsilon,\boldsymbol{\psi}} := \frac{1}{N_{1,n}^d} \int_{[0,N_{1,n}]^d} \left(\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t}) - Y(\boldsymbol{t})\right)^2 d\boldsymbol{t}. \tag{5.12}$$

Consider a consistent estimator $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}^{(0)}$. Then

$$|E_{\epsilon,\boldsymbol{\psi}^{(0)}} - E_{\epsilon,\hat{\boldsymbol{\psi}}}| = o_p(1). \tag{5.13}$$

Furthermore, there exists a constant $A > 0$ so that for all n,

$$\mathbb{E}\left(E_{\epsilon,\boldsymbol{\psi}^{(0)}}\right) \geq A. \tag{5.14}$$

In proposition 5.21, the condition that the Gaussian process Y yields continuous trajectories is not restrictive, as we have seen in proposition 2.23. In proposition 5.21, we show that the mean square prediction error, over the observation domain, with a consistently estimated covariance hyper-parameter, is asymptotically equivalent to the corresponding error when the true covariance hyper-parameter is known. Furthermore, the mean value of this prediction error with the true covariance hyper-parameter does not vanish when $n \to +\infty$. This is intuitive because the density of observation points in the domain is constant.

Hence, expansion-domain asymptotics yields a situation in which the estimation error goes to zero, but the prediction error does not, because the prediction domain increases with the number of observation points. Thus, increasing-domain asymptotic context enables us to address the prediction and estimation problems separately, and the conclusions on the estimation problem are fruitful, as we have seen in sections 5.3 and 5.4. However, this context does not enable us to study theoretically all the practical aspects of the joint problem of prediction with estimated covariance hyper-parameters. For instance, the impact of the estimation method on the prediction error is asymptotically zero under this theoretical framework, and using a constant proportion of the observation points for estimation rather than prediction cannot decrease the asymptotic prediction error with estimated covariance hyper-parameters.

The two aforementioned practical problems would benefit from an asymptotic framework that would fully reproduce them, by giving a stronger impact to the estimation on the prediction. Possible candidates for this framework are the mixed increasing-domain asymptotic framework, presented in chapter 4, and addressed for instance in [Lah03] and [LM04], and fixed-domain asymptotics. In both frameworks, the estimation error, with respect to the number of observation points, is larger and the prediction error is smaller, thus giving hope for more impact of the estimation on the prediction. Nevertheless, even in fixed-domain asymptotics, notice that in [PY01] and referring to [dV96], it is shown that, for the particular case of the tensor product exponential covariance function in two dimensions, the prediction error, under covariance hyper-parameters estimated by ML, is asymptotically equal to the prediction error under the true covariance hyper-parameters. This is a particular case in which estimation has no impact on prediction, even under fixed-domain asymptotics.

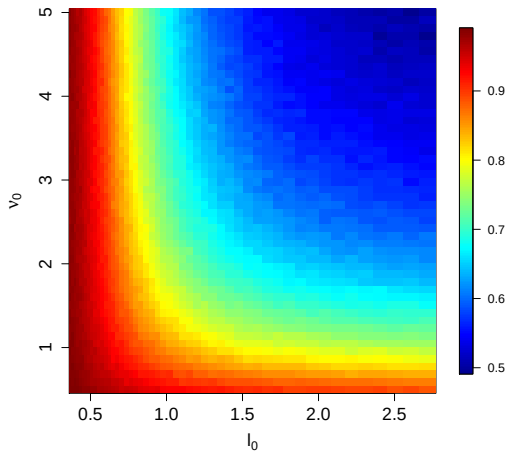

Figure 5.16: Ratio of the mean square prediction error $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$ in proposition 5.21, between $\epsilon = 0$ and $\epsilon = 0.45$, as a function of ℓ_0 and ν_0, for $n = 100$. The mean square prediction error increases with the irregularity of the sampling.

5.5.3 Analysis of the impact of the spatial sampling on the Kriging prediction

In this subsection 5.5.3, we study the prediction mean square error $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$ of proposition 5.21, as a function of ϵ, ℓ_0 and ν_0, for the one-dimensional Matérn model, and for large n. The distribution of the $(X_i)_{1\leq i\leq n}$ is still uniform on $[-1,1]$. The function $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$ is independent of the estimation, as we have seen, so there is now no point in distinguishing between ML and CV. In the following figures, this function is approximated by the average of *iid* realizations of its conditional mean value given $\boldsymbol{X} = \boldsymbol{x}$,

$$\frac{1}{n}\int_0^n \left(1 - \boldsymbol{k}^t_{\ell_0,\nu_0}(t)\mathbf{K}^{-1}_{\ell_0,\nu_0}\boldsymbol{k}_{\ell_0,\nu_0}(t)\right) dt,$$

where $\left(\boldsymbol{k}_{\ell_0,\nu_0}(t)\right)_i = K_{\ell_0,\nu_0}(i + \epsilon x_i - t)$ and $\left(\mathbf{K}_{\ell_0,\nu_0}\right)_{i,j} = K_{\ell_0,\nu_0}(i - j + \epsilon[x_i - x_j])$.

On figure 5.16, we plot the ratio of the mean square prediction error $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$, between $\epsilon = 0$ and $\epsilon = 0.45$, as a function of ℓ_0 and ν_0, for $n = 100$ (we observed the same results for $n = 50$). We see that this ratio is always smaller than one, meaning that strongly perturbing the regular grid always increases the prediction error. This result is in agreement with the common practices of using regular, also called space filling, samplings for optimizing the Kriging predictions with known covariance hyper-parameters, as illustrated in figure 3 of [ZZ06].

In figure 5.17, we fix the true covariance hyper-parameters $\ell_0 = 0.5$, $\nu_0 = 2.5$, and we study the variations with respect to ϵ of the asymptotic variance of the ML estimation of ν, when ℓ_0 is known (figure 5.9), and of the prediction mean square error $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$, for $n = 50$ and $n = 100$. The results are the same for $n = 50$ and $n = 100$. We first observe that $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$ is globally an increasing function of ϵ. In fact, we observe the same global increase of $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$, for $n = 50$ and $n = 100$, with respect to ϵ, for all the values $(0.5, 5)$, $(2.7, 1)$, $(0.5, 2.5)$, $(0.7, 2.5)$, $(2.7, 2.5)$, $(0.73, 2.5)$ and $(1.7, 5)$, for (ℓ_0, ν_0), that we have studied in section 5.4. This is again a

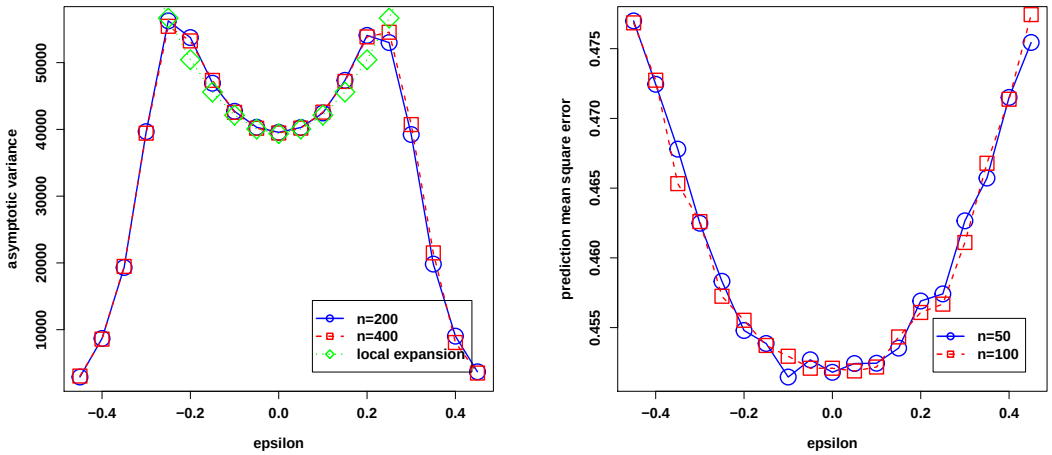

Figure 5.17: $\ell_0 = 0.5$, $\nu_0 = 2.5$. Left: asymptotic variance for the ML estimation of ν, when ℓ is known, as a function of ϵ (same setting as in figure 5.9). Right: prediction mean square error $\mathbb{E}\left(E_{\epsilon,\ell_0,\nu_0}\right)$ in proposition 5.21 as a function of ϵ. There is no simple antagonistic relationship between the impact of the irregularity of the spatial sampling on estimation and on prediction

confirmation that, in the increasing-domain asymptotic framework treated here, evenly spaced observations perform best for prediction.

The second conclusion than can be drawn for figure 5.17 is that there is independence between estimation (ML in this case) and prediction. Indeed, the estimation error first increases and then decreases with respect to ϵ, while the prediction error globally decreases. Hence, in figure 5.17, the regular grid still gives better prediction, although it leads to less asymptotic variance than mildly irregular samplings. Therefore, there is no simple antagonistic relationship between the impact of the irregularity of the spatial sampling on estimation and on prediction.

5.6 Conclusion

We have considered an increasing-domain asymptotic framework to study the consistency and asymptotic normality of the CV estimator, to compare asymptotically CV and ML and to address the influence of the irregularity of the spatial sampling on the estimation of the covariance hyper-parameters. This asymptotic framework is based on a random sequence of observation points, for which the deviation from the regular grid is controlled by a single scalar regularity parameter ϵ.

We have proved consistency and asymptotic normality for the ML and CV estimators, under rather minimal conditions. These results are dedicated to the randomly perturbed regular grid. We believe that, for the proof methods we have used for consistency and asymptotic normality, the most important feature of the randomly perturbed grid is the minimum distance between two different observation points. Hence, it may be possible to extend the proofs of this chapter, for consistency and asymptotic normality, to other samplings verifying this minimum distance assumption, with possibly more technicality.

We have thus shown that CV is consistent and furthermore has the same rate of convergence as ML. The asymptotic covariance matrices are deterministic functions of the regularity parameter only. By numerically investigating them, we point out that ML is more efficient that CV, in the well-specified case of chapter 5. Furthermore the asymptotic covariance matrices are the natural tool to assess the influence of the irregularity of the spatial sampling on the ML and CV estimators.

This is carried out by means of an exhaustive study of the Matérn model. It is shown that mildly perturbing the regular grid can damage both ML and CV estimation, and that CV estimation can also be damaged when strongly perturbing the regular grid. However, we put into evidence that strongly perturbing the regular grid always improves the ML estimation, which is a more efficient estimator than CV, in the well-specified case addressed here. Hence, we confirm the conclusion of [Ste99] and [ZZ06] that using groups of observation points with small spacing, compared to the observation density in the observation domain, improves the covariance function estimation. In geostatistics, such groups of points are also added to regular samplings in practice [JDLI08].

We have also studied the impact of the spatial sampling on the prediction error. Regular samplings were shown to be the most efficient as regards to this criterion. This is in agreement with, e.g. [ZZ06] and with [PM12] where samplings for Kriging prediction with known covariance hyper-parameters are selected by optimizing a space filling criterion. An example of space filling criterion is the maximin criterion (6.21) presented in chapter 6.

The ultimate goal of a Kriging model is prediction with estimated covariance hyper-parameters. Hence, efficient samplings for this criterion must address two criteria that have been shown to be antagonistic. In the literature, there seems to be a commonly admitted practice for solving this issue [ZZ06, PM12]. Roughly speaking, for selecting an efficient sampling for prediction with estimated covariance hyper-parameters, one may select a regular sampling for prediction with known covariance hyper-parameters and augment it with a sampling for estimation (with closely spaced observation points). The proportion of points for the two samplings is optimized in the two aforementioned references by optimizing a criterion for prediction with estimated covariance hyper-parameters. This criterion is more expensive to compute, but is not optimized in a large dimensional space. In [ZZ06, PM12], the majority of the observation points belong to the regular sampling for prediction with known covariance function. This is similar in the geostatistical community [JDLI08], where regular samplings, augmented with few closely spaced observation points, making the inputs vary mildly, are used. In view of our theoretical and practical results of sections 5.4 and 5.5, we are in agreement with this method for building samplings for prediction with estimated covariance hyper-parameters.

An important limitation we see, though, in the expansion-domain asymptotic framework we address in this chapter, is that prediction with estimated covariance hyper-parameters corresponds asymptotically to prediction with known covariance function. Said differently, the proportion of observation points addressing estimation, in the aforementioned trade-off, would go to zero. As we discuss after proposition 5.21, mixed increasing-domain or fixed-domain asymptotics could give more importance to the estimation problem, compared to the problem of predicting with known covariance function.

Finally, in the well-specified-framework addressed in this chapter, ML is more precise than CV to estimate a correlation hyper-parameter. The practical interest of CV arises in the complementary framework, that we call the misspecified framework, where the true correlation function does not belong to the parametric set of correlation functions used for estimation. This is the object of the next chapter 6.

5.7 Proofs

In the proofs, we distinguish three probability spaces.

$(\Omega_X, \mathcal{F}_X, P_X)$ is the probability space associated with the random perturbation of the regular grid. $(X_i)_{i\in\mathbb{N}^*}$ is a sequence of *iid* S_X-valued random variables defined on $(\Omega_X, \mathcal{F}_X, P_X)$, with distribution $\mathcal{L}_X$. We denote by ω_X an element of Ω_X.

$(\Omega_Y, \mathcal{F}_Y, P_Y)$ is the probability space associated with the Gaussian process. Y is a centered Gaussian process with covariance function $K_{\boldsymbol{\psi}^{(0)}}$ defined on $(\Omega_Y, \mathcal{F}_Y, P_Y)$. We denote by ω_Y an element of Ω_Y.

$(\Omega, \mathcal{F}, \mathbb{P})$ is the product space $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y, P_X \times P_Y)$. We denote by ω an element of Ω.

All the random variables in the proofs can be defined relatively to the product space $(\Omega, \mathcal{F}, \mathbb{P})$. Hence, all the probabilistic statements in the proofs hold with respect to this product space, unless it is stated otherwise.

In the proofs, when $(f_n)_{n\in\mathbb{N}^*}$ is a sequence of real functions of $\boldsymbol{X} = (X_i)_{i=1}^n$, f_n is also a sequence of real random variables on $(\Omega_X, \mathcal{F}_X, P_X)$. When we write that f_n is bounded uniformly in n and $\boldsymbol{x}$, we mean that there exists a finite constant T so that $\sup_n \sup_{\boldsymbol{x}\in S_X^n} |f_n(\boldsymbol{x})| \leq T$. We then have that f_n is bounded P_X-a.s., i.e $\sup_n f_n \leq T$ for a.e. $\omega_X \in \Omega_X$. We may also write that f_n is lower-bounded uniformly in n and $\boldsymbol{x}$ when there exists $a > 0$ so that $\inf_n \inf_{\boldsymbol{x}\in S_X^n} f_n(\boldsymbol{x}) \geq a$. When f_n also depends on $\boldsymbol{\psi}$, we say that f_n is bounded uniformly in n, $\boldsymbol{x}$ and $\boldsymbol{\psi}$ when $\sup_{\boldsymbol{\psi}\in\Psi} f_n$ is bounded uniformly in n and $\boldsymbol{x}$. We also say that f_n is lower-bounded uniformly in n, $\boldsymbol{x}$ and $\boldsymbol{\psi}$ when $\inf_{\boldsymbol{\psi}\in\Psi} f_n$ is lower-bounded uniformly in n and $\boldsymbol{x}$.

When we write that f_n converges to zero uniformly in $\boldsymbol{x}$, we mean that

$$\sup_{\boldsymbol{x}\in S_X^n} |f_n(\boldsymbol{x})| \to_{n\to+\infty} 0.$$

In this case, f_n converges to zero P_X-a.s. When f_n also depends on $\boldsymbol{\psi}$, we say that f_n converges to zero uniformly in n, $\boldsymbol{x}$ and $\boldsymbol{\psi}$ when $\sup_{\boldsymbol{\psi}\in\Psi} f_n$ converges to zero uniformly in n and $\boldsymbol{x}$.

When f_n is a sequence of real functions of $\boldsymbol{X}$ and Y, f_n is also a sequence of real random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. When we say that f_n is bounded in probability conditionally to $\boldsymbol{X} = \boldsymbol{x}$ and uniformly in $\boldsymbol{x}$, we mean that, for every $\epsilon > 0$, there exist m, N so that $\sup_{n\geq N} \sup_{\boldsymbol{x}\in S_X^n} \mathbb{P}(|f_n| \geq m|\boldsymbol{X} = \boldsymbol{x}) \leq \epsilon$. In this case, f_n is bounded in probability (defined on the product space).

5.7.1 Proofs for subsection 5.3.1

Some matrix relations

In proposition 5.22, we give some matrix relations that are useful in the proofs below.

Proposition 5.22. *Let* $\mathbf{A}$, $\mathbf{B}$ *be two* $n \times n$ *real symmetric positive matrices and let* $\mathbf{M}, \mathbf{P}, \mathbf{Q}, \mathbf{R}$ *be* $n \times n$ *real matrices. Let* $\phi_1(\mathbf{A}) \geq ... \geq \phi_n(\mathbf{A}) > 0$ *be the eigenvalues of* $\mathbf{A}$.

Then,

$$||\mathbf{MN}||_2 \leq ||\mathbf{M}||.||\mathbf{N}||_2, \tag{5.15}$$

$$\phi_n(\mathbf{A})||\mathbf{M}||_2 \leq ||\mathbf{AM}||_2 \leq \phi_1(\mathbf{A})||\mathbf{M}||_2, \tag{5.16}$$

$$||\mathbf{A}^{-1} - \mathbf{B}^{-1}||_2 \leq ||\mathbf{A}^{-1}||.||\mathbf{B}^{-1}||.||\mathbf{A} - \mathbf{B}||_2, \tag{5.17}$$

$$\frac{1}{n}|Tr(\mathbf{M})| \leq ||\mathbf{M}||_2 \tag{5.18}$$

and

$$||\mathbf{QM} - \mathbf{PR}||_2 \leq ||\mathbf{Q}||.||\mathbf{M} - \mathbf{R}||_2 + ||\mathbf{R}||.||\mathbf{Q} - \mathbf{P}||_2. \tag{5.19}$$

For $\boldsymbol{y} \sim \mathcal{N}(0, \mathbf{A})$,

$$Cov(\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}, \boldsymbol{y}^t\mathbf{P}\boldsymbol{y}) = Tr(\mathbf{AMAP}) + Tr(\mathbf{AMAP}^t). \tag{5.20}$$

Proof. The equation (5.15) is proved by lemma 2.3 in [Gra01].

The equation (5.16) is proved by lemma 2.1 in [Gra01].

The equations (5.17) and (5.19) are proved in the proof of theorem 2.1 in [Gra01].

The equation (5.18) is proved in the proof of lemma 2.4 in [Gra01].

For (5.20), $\mathbb{E}(\boldsymbol{y}^t\mathbf{P}\boldsymbol{y}) = Tr(\mathbf{AP})$ and $\mathbb{E}(\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}) = Tr(\mathbf{AM})$. Let $\boldsymbol{y} = \mathbf{A}^{\frac{1}{2}}\boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I}_n)$.

$$\begin{aligned}
\mathbb{E}(\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}\boldsymbol{y}^t\mathbf{P}\boldsymbol{y}) &= \mathbb{E}(\boldsymbol{z}^t(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})\boldsymbol{z}\boldsymbol{z}^t(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}})\boldsymbol{z}) \\
&= \sum_{i,j,k,l=1}^{n} (\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})_{i,j}(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})_{k,l}\mathbb{E}(z_i z_j z_k z_l).
\end{aligned}$$

From appendix A in [Ste99], $\mathbb{E}(z_i z_j z_k z_l) = \delta_{i,j}\delta_{k,l} + \delta_{i,k}\delta_{j,l} + \delta_{i,l}\delta_{j,k}$. Hence

$$\begin{aligned}
\mathbb{E}(\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}\boldsymbol{y}^t\mathbf{P}\boldsymbol{y}) &= \sum_{i,k=1}^{n} (\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})_{i,i}(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}})_{k,k} + \sum_{i,j=1}^{n} (\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})_{i,j}(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}})_{i,j} \\
&\quad + \sum_{i,j=1}^{n} (\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})_{i,j}(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}})_{j,i} \\
&= Tr\left(\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}}\right) Tr\left(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}}\right) + Tr\left((\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}})^t\right) \\
&\quad + Tr\left((\mathbf{A}^{\frac{1}{2}}\mathbf{M}\mathbf{A}^{\frac{1}{2}})(\mathbf{A}^{\frac{1}{2}}\mathbf{P}\mathbf{A}^{\frac{1}{2}})\right) \\
&= Tr\left(\mathbf{AM}\right) Tr\left(\mathbf{AP}\right) + Tr\left((\mathbf{AMAP}^t\right) + Tr\left((\mathbf{AMAP}\right).
\end{aligned}$$

This ends the proof. □

Eigenvalues control for random matrices

One of the key points of the proofs in this chapter 5 is that the eigenvalues of the matrices $\mathbf{K}_{\boldsymbol{\psi}}$, $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ and $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1},...,\partial\psi_{i_q}}$, $1 \leq q \leq 3$, $1 \leq i_1, ..., i_q \leq p$, are bounded regardless of n and the perturbations $X_1, ..., X_n$.

This would not hold in a fixed-domain asymptotic context, and is therefore essential for the proofs in this chapter 5.

In lemma 5.23, we begin by controlling the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}$ and $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1},...,\partial\psi_{i_q}}$, $1 \leq q \leq 3$, $1 \leq i_1, ..., i_q \leq p$

Lemma 5.23. *Assume that condition 5.1 is satisfied.*

For all $|\epsilon| < \frac{1}{2}$ there exists C_ϵ so that the eigenvalues of $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1},...,\partial\psi_{i_q}}$, $0 \leq q \leq 3$, $1 \leq i_1, ..., i_q \leq p$, are bounded by C_ϵ, uniformly in $n \in \mathbb{N}$, $\boldsymbol{x} \in (S_X)^n$ and $\boldsymbol{\psi} \in \Psi$.

Proof of lemma 5.23. Bounding the eigenvalues of $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1},...,\partial\psi_{i_q}}$, $0 \leq q \leq 3$, $1 \leq i_1, ..., i_q \leq p$ is done by controlling the sums of the row elements of these matrices. Lemma 5.24 enables us to do so, by showing how a summable function on $\mathbb{R}^d$ becomes a summable sequence on the $\left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})\right)_{j \neq i}$, for each fixed i.

Lemma 5.24. *Let $f : \mathbb{R}^d \to \mathbb{R}^+$, so that $f(\boldsymbol{t}) \leq \frac{1}{1+|\boldsymbol{t}|_\infty^{d+1}}$. Then, for all $i \in \mathbb{N}^*$, $\epsilon \in (-\frac{1}{2}, \frac{1}{2})$ and $\left(\boldsymbol{x}^{(i)}\right)_{i \in \mathbb{N}^*} \in S_X^{\mathbb{N}^*}$,*

$$\sum_{j \in \mathbb{N}^*, j \neq i} f\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\} \leq 2^d d \sum_{j \in \mathbb{N}} \frac{\left(j + \frac{3}{2}\right)^{d-1}}{1 + j^{d+1}}.$$

Proof of lemma 5.24.

$$\begin{aligned}
&\sum_{j \in \mathbb{N}, j \neq i} f\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\} \\
&\leq \sum_{\boldsymbol{v} \in \mathbb{Z}^d, \boldsymbol{v} \neq 0} \sup_{\boldsymbol{\delta}_{\boldsymbol{v}} \in [-1,1]^d} f(\boldsymbol{v} + \boldsymbol{\delta}_{\boldsymbol{v}}) \\
&= \sum_{j \in \mathbb{N}} \sum_{\boldsymbol{v} \in \{-j-1, ..., j+1\}^d \backslash \{-j, j\}^d} \sup_{\boldsymbol{\delta}_{\boldsymbol{v}} \in [-1,1]^d} f(\boldsymbol{v} + \boldsymbol{\delta}_{\boldsymbol{v}}) \qquad .
\end{aligned}$$

For $\boldsymbol{v} \in \{-j-1, ..., j+1\}^d \backslash \{-j, j\}^d$, $|\boldsymbol{v} + \boldsymbol{\delta}_v|_\infty \geq j$. The cardinality of the set $\{-j-1, ..., j+1\}^d \backslash \{-j, j\}^d$ is

$$(2j+3)^d - (2j+1)^d = \int_{2j+1}^{2j+3} d.t^{d-1} dt \leq 2d\,(2j+3)^{d-1} = 2^d d \left(j + \frac{3}{2}\right)^{d-1}.$$

Hence

$$\sum_{j \in \mathbb{N}} f\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\} \leq \sum_{j \in \mathbb{N}} 2^d d \left(j + \frac{3}{2}\right)^{d-1} \frac{1}{1 + j^{d+1}}.$$

□

Going from a boundedness of the sums of the row elements of a symmetric matrix to a boundedness of its eigenvalues is done by using the following Gershgorin circle theorem (see e.g. [GL96]).

Theorem 5.25 (Gershgorin circle theorem)**.** *Let* $\mathbf{A}$ *be a* $n \times n$ *symmetric matrix, and* λ *one of its eigenvalues. Then, there exists* i *so that*

$$|A_{i,i} - \lambda| \leq \sum_{j \neq i} |A_{i,j}|.$$

Using Gershgorin theorem 5.25 together with lemma 5.24 and (5.3), we get to, with ϕ_{max} the largest eigenvalue of $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_{i_1}, ..., \partial \psi_{i_q}}$,

$$\left| \frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_{i_1}, ..., \partial \psi_{i_q}}(0) - \phi_{max} \right| \leq C,$$

for a constant $C < +\infty$. Hence $|\phi_{max}| \leq C + |\frac{\partial^q K_{\boldsymbol{\psi}}}{\partial \psi_{i_1}, ..., \partial \psi_{i_q}}(0)| < +\infty$. □

The next proposition 5.26 shows that the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ are upper-bounded, or equivalently that the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}$ are lower-bounded.

Proposition 5.26. *Assume that condition 5.1 is satisfied.*

For all $0 \leq \delta < \frac{1}{2}$*, there exists* $C_\delta > 0$ *so that for all* $|\epsilon| \leq \delta$*, for all* $\boldsymbol{\psi} \in \Psi$*, for all* $n \in \mathbb{N}^*$ *and for all* $\boldsymbol{x} \in (S_X)^n$*, the eigenvalues of* $\mathbf{K}_{\boldsymbol{\psi}}$ *are larger than* C_δ*.*

Proof of proposition 5.26. We begin the proof of proposition 5.26 by stating lemma 5.27, which is quite similar to lemma 5.24 and will allow to show that, roughly speaking, when a covariance function has an arbitrarily small correlation length, the sum of the non-diagonal elements of the rows of the matrix it yields is arbitrarily small, uniformly in n and $\boldsymbol{x}$.

Lemma 5.27. *Let* $f : \mathbb{R}^d \to \mathbb{R}^+$*, so that* $f(\boldsymbol{t}) \leq \frac{1}{1+|\boldsymbol{t}|_\infty^{d+1}}$*. We consider* $\delta < \frac{1}{2}$*. Then, for all* $i \in \mathbb{N}^*$*,* $a > 0$*,* $\epsilon \in [-\delta, \delta]$ *and* $\left(\boldsymbol{x}^{(i)}\right)_{i \in \mathbb{N}^*} \in S_X^{\mathbb{N}^*}$*,*

$$\sum_{j \in \mathbb{N}^*, j \neq i} f\left[a\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}\right] \leq 2^d d \sum_{j \in \mathbb{N}} \frac{\left(j + \frac{3}{2}\right)^{d-1}}{1 + a^{d+1}\left(j + 1 - 2\delta\right)^{d+1}}.$$

Proof of lemma 5.27. Similar to the proof of lemma 5.24. □

Let $h : \mathbb{R}^d \to \mathbb{R}$ so that $\hat{h}(\boldsymbol{f}) = \prod_{i=1}^d \hat{h}_i(f_i)$, with $\hat{h}_i(f_i) = \mathbf{1}_{f_i^2 \in [0,1]} \exp\left(-\frac{1}{1-f_i^2}\right)$. For $1 \leq i \leq d$, $\hat{h}_i : \mathbb{R} \to \mathbb{R}$ is C^∞, with compact support, so there exists $C > 0$ so that $|h_i(t_i)| \leq \frac{C^{\frac{1}{d}}}{1+|t_i|^{d+1}}$. Now, since the inverse Fourier transform of $\prod_{i=1}^d \hat{h}_i(f_i)$ is $\prod_{i=1}^d h_i(t_i)$, we have

$$|h(\boldsymbol{t})| \leq C \prod_{i=1}^d \frac{1}{1 + |t_i|^{d+1}} \leq \frac{C}{1 + |\boldsymbol{t}|_\infty^{d+1}}.$$

Hence, from lemma 5.27, for all $i \in \mathbb{N}$ and $a > 0$,

$$\sum_{j \in \mathbb{N}, j \neq i} \left| h\left[a\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}\right] \right| \leq C 2^d d \sum_{j \in \mathbb{N}} \frac{\left(j + \frac{3}{2}\right)^{d-1}}{1 + a^{d+1}\left(j + 1 - 2\delta\right)^{d+1}}. \tag{5.21}$$

The right-hand term in (5.21) goes to zero when $a \to +\infty$. Also, $h(0)$ is positive, because $\hat{h}$ is non-negative and is not almost surely zero on $\mathbb{R}^d$ with respect to the Lebesgue measure. Thus, there exists $0 < a < \infty$ so that for all $i \in \mathbb{N}$,

$$\sum_{j \in \mathbb{N}, j \neq i} \left| h\left[a\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}\right] \right| \leq \frac{1}{2} h(0). \tag{5.22}$$

Using theorem 5.25, for any $n \in \mathbb{N}^*$, $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in S_X$, the eigenvalues of the symmetric matrices $\left(h\left[a\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}\right]\right)_{1\leq i,j\leq n}$ belong to the balls with center $h(0)$ and radius $\sum_{1\leq j\leq n, j\neq i}\left|h\left[a\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}\right]\right|$. Thus, because of (5.22), these eigenvalues belong to the segment $[h(0) - \frac{1}{2}h(0), h(0) + \frac{1}{2}h(0)]$ and are larger than $\frac{1}{2}h(0)$.

Hence, for all n, $t_1, ..., t_n \in \mathbb{R}$, $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in S_X$,

$$\begin{aligned}
\frac{1}{2}h(0)\sum_{i=1}^{n} t_i^2 &\leq \sum_{i,j=1}^{n} t_i t_j h\left[a\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}\right] \\
&= \sum_{i,j=1}^{n} t_i t_j \frac{1}{a^d}\int_{\mathbb{R}^d} \hat{h}\left(\frac{\boldsymbol{f}}{a}\right) e^{\mathrm{i}\boldsymbol{f}\cdot\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}} d\boldsymbol{f} \\
&= \frac{1}{a^d}\int_{\mathbb{R}^d} \hat{h}\left(\frac{\boldsymbol{f}}{a}\right)\left|\sum_{i=1}^{n} t_i e^{\mathrm{i}\boldsymbol{f}\cdot\left(\boldsymbol{v}^{(i)} + \epsilon\boldsymbol{x}^{(i)}\right)}\right|^2 d\boldsymbol{f}.
\end{aligned}$$

Hence, as $\hat{K}_{\boldsymbol{\psi}}(\boldsymbol{f}) : (\Psi \times \mathbb{R}^d) \to \mathbb{R}$ is continuous and positive, using a compacity argument, there exists $C_2 > 0$ so that for all $\boldsymbol{\psi} \in \Psi$, $\boldsymbol{f} \in [-a, a]^d$, $\hat{K}_{\boldsymbol{\psi}}(\boldsymbol{f}) \geq C_2 \hat{h}\left(\frac{\boldsymbol{f}}{a}\right)$. Hence,

$$\begin{aligned}
\frac{1}{2}h(0)\sum_{i=1}^{n} t_i^2 &\leq \frac{1}{a^d C_2}\int_{\mathbb{R}^d} \hat{K}_{\boldsymbol{\psi}}(\boldsymbol{f})\left|\sum_{i=1}^{n} t_i e^{\mathrm{i}\boldsymbol{f}\cdot\left(\boldsymbol{v}^{(i)} + \epsilon\boldsymbol{x}^{(i)}\right)}\right|^2 d\boldsymbol{f}, \\
&= \frac{1}{a^d C_2}\sum_{i,j=1}^{n} t_i t_j K_{\boldsymbol{\psi}}\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\}.
\end{aligned}$$

□

Now, combining lemmas 5.23 and proposition 5.26, we obtain the following lemma 5.28, summarizing the results on the boundedness of the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}$, $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ and $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1}, ..., \partial\psi_{i_q}}$, $1 \leq q \leq 3$, $1 \leq i_1, ..., i_q \leq p$.

Lemma 5.28. *Assume that condition 5.1 is satisfied.*

For all $|\epsilon| < \frac{1}{2}$ there exists C_ϵ so that the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ and of $\frac{\partial^q \mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1}, ..., \partial\psi_{i_q}}$, $0 \leq q \leq 3$, $1 \leq i_1, ..., i_q \leq p$, are bounded by C_ϵ, uniformly in $n \in \mathbb{N}$, $\boldsymbol{x} \in (S_X)^n$ and $\boldsymbol{\psi} \in \Psi$.

From lemma 5.28, the next proposition enables us to control the singular values of the matrices that can be written using only matrix multiplications, the matrix $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$, the matrices $\frac{\partial^k}{\partial\psi_{i_1}, ..., \partial\psi_{i_k}}\mathbf{K}_{\boldsymbol{\psi}}$, for $i_1, ..., i_k \in \{1, ..., p\}$, the *Diag* operator applied to the symmetric products of matrices $\mathbf{K}_{\boldsymbol{\psi}}$, $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ and $\frac{\partial^k}{\partial\psi_{i_1}, ..., \partial\psi_{i_k}}\mathbf{K}_{\boldsymbol{\psi}}$, and the matrix $Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1}$. Examples of sums of these matrices are the matrices $\boldsymbol{\Sigma}_{ML}$, $\boldsymbol{\Sigma}_{CV,1}$ and $\boldsymbol{\Sigma}_{CV,2}$ of propositions 5.8 and 5.12.

Proposition 5.29. *Assume that condition 5.1 is satisfied.*

Let $\boldsymbol{\psi} \in \Psi$. We denote the set of multi-indexes $S_p := \cup_{k\in\{0,1,2,3\}}\{1, ..., p\}^k$. For $I = (i_1, ..., i_k) \in S_p$, we denote $n(I) = k$. Then, we denote for $I \in S_p \cup \{-1\}$,

$$\mathbf{K}_{\boldsymbol{\psi}}^I := \begin{cases} \frac{\partial^{n(I)}}{\partial\psi_{I_1}, ..., \partial\psi_{I_{n(I)}}}\mathbf{K}_{\boldsymbol{\psi}} & \text{if } I \in S_p \\ \mathbf{K}_{\boldsymbol{\psi}}^{-1} & \text{if } I = -1 \end{cases}.$$

We then denote

- $\mathbf{M}_{nd}^I = \mathbf{K}_{\boldsymbol{\psi}}^I$ *for* $I \in S_{nd} := (S_p \cup \{-1\})$

- $\mathbf{M}_{sd}^1 = Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1}$
- $\mathbf{M}_{bd}^I = Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{I_1}...\mathbf{K}_{\boldsymbol{\psi}}^{I_{n(I)}}\right) \qquad for\ I \in S_{bd} := \cup_{k\in\mathbb{N}^*} S_{nd}^k$

For $I = (I_1, ..., I_k) \in S_{bd}$, we also denote $n(I) = k$. We then define $\mathcal{M}_{\boldsymbol{\psi}}$ as the set of sequences of random matrices (defined on $(\Omega_X, \mathcal{F}_X, P_X)$), indexed by $n \in \mathbb{N}^*$, dependent on $\boldsymbol{X}$, which can be written $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}$ with $\{d_1, I_1\}, ..., \{d_K, I_K\} \in (\{nd\} \times S_{nd}) \cup (\{sd\} \times \{1\}) \cup (\{bd\} \times S_{bd})$, and so that, for the matrices $\mathbf{M}_{d_j}^{I_j}$, so that $d_j = bd$, the matrix $\mathbf{K}_{\boldsymbol{\psi}}^{(I_j)_1}...\mathbf{K}_{\boldsymbol{\psi}}^{(I_j)_{n(I_j)}}$ be symmetric.

Then, for every matrix $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}$ of $\mathcal{M}_{\boldsymbol{\psi}}$, the singular values of $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}$ are bounded uniformly in $\boldsymbol{\psi}$, n and $\boldsymbol{x} \in (S_X)^n$.

Proof of proposition 5.29. Let $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K} \in \mathcal{M}_{\boldsymbol{\psi}}$ be fixed in the proof.

The eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}^I$, $I \in S_{nd}$, are bounded uniformly with respect to n, $\boldsymbol{\psi}$ and $\boldsymbol{x}$ (lemma 5.28).

Next, lemma 5.30 enables us to treat the *Diag* operator.

Lemma 5.30. *For $\mathbf{M}$ symmetric real non-negative matrix, $\inf_i \phi_i(Diag(\mathbf{M})) \geq \inf_i \phi_i(\mathbf{M})$ and $\sup_i \phi_i(Diag(\mathbf{M})) \leq \sup_i \phi_i(\mathbf{M})$. Furthermore, if for two sequences of symmetric matrices $\mathbf{M}_n$ and $\mathbf{N}_n$, $\mathbf{M}_n \sim \mathbf{N}_n$, then $Diag(\mathbf{M}_n) \sim Diag(\mathbf{N}_n)$.*

Proof of lemma 5.30. We use $\mathbf{M}_{i,i} = e_i^t \mathbf{M} e_i$, where $(e_i)_{i=1...n}$ is the standard basis of $\mathbb{R}^n$. Hence $\inf_i \phi_i(\mathbf{M}) \leq \mathbf{M}_{i,i} \leq \sup_i \phi_i(\mathbf{M})$ for a symmetric real non-negative matrix $\mathbf{M}$. We also use $||Diag(\mathbf{M})||_2 \leq ||\mathbf{M}||_2$. □

Then, using lemma 5.30, we show that the eigenvalues of $Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1}$ are bounded uniformly in $\boldsymbol{x}$, n and $\boldsymbol{\psi}$. Then, for $\mathbf{M}_{bd}^I = Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{I_1}...\mathbf{K}_{\boldsymbol{\psi}}^{I_{n(I)}}\right)$, the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}^{I_1}...\mathbf{K}_{\boldsymbol{\psi}}^{I_{n(I)}}$ are bounded by the product of the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}^{I_1}, ..., \mathbf{K}_{\boldsymbol{\psi}}^{I_{n(I)}}$. Hence we use lemma 5.30 to show that the eigenvalues of $\mathbf{M}_{bd}^I$ are bounded uniformly in n, $\boldsymbol{\psi}$ and $\boldsymbol{x}$. Finally we use $||\mathbf{A}_1...\mathbf{A}_K|| \leq ||\mathbf{A}_1||...||\mathbf{A}_K||$ to show that $||\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}||$ is bounded uniformly in n, $\boldsymbol{\psi}$ and $\boldsymbol{x}$. □

Almost sure convergence of traces of random matrices

To show that, say, the matrix $\boldsymbol{\Sigma}_{ML}$ in proposition 5.8, whose element i, j is defined as the almost sure limit of the trace of the random matrices

$$\frac{1}{n}\left(\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \psi_i}\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \psi_j}\right), \tag{5.23}$$

exists, we use the following proposition 5.31 on the almost sure convergence of random traces of matrices similar to (5.23).

Proposition 5.31. *Assume that condition 5.1 is satisfied.*

Consider the set of random matrix sequences $\mathcal{M}_{\boldsymbol{\psi}}$ of proposition 5.29.

Then, for every matrix $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}$ of $\mathcal{M}_{\boldsymbol{\psi}}$, denoting $S_n := \frac{1}{n} Tr\left(\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}\right)$, there exists a deterministic limit S, which only depends on ϵ, $\boldsymbol{\psi}$ and $(d_1, I_1), ..., (d_K, I_K)$, so that $S_n \to S$ P_X-almost surely. Furthermore $S_n \to S$ in quadratic mean and $Var(S_n) \to 0$ as $n \to +\infty$

Proof of proposition 5.31. Let $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K} \in \mathcal{M}_{\boldsymbol{\psi}}$ be fixed in the proof.

Because of proposition 5.29, $||\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}||$ is bounded uniformly in n, $\boldsymbol{\psi}$ and $\boldsymbol{x}$.

We decompose n into $n = N_1^d n_2 + r$ with $N_1, n_2, r \in \mathbb{N}$ and $r < N_1^d$. We define $C\left(\boldsymbol{v}^{(i)}\right)$ as the unique $\boldsymbol{v} \in \mathbb{N}^d$ so that $\boldsymbol{v}^{(i)} \in E_{\boldsymbol{v}} := \prod_{k=1}^d \{N_1 v_k + 1, ..., N_1 (v_k + 1)\}$.

We then define the sequence of matrices $\tilde{\mathbf{K}}_{\boldsymbol{\psi}}$ by $\left(\tilde{\mathbf{K}}_{\boldsymbol{\psi}}\right)_{i,j} = \left(\mathbf{K}_{\boldsymbol{\psi}}\right)_{i,j} \mathbf{1}_{C\left(\boldsymbol{v}^{(i)}\right)=C\left(\boldsymbol{v}^{(j)}\right)}$. Roughly speaking, $\tilde{\mathbf{K}}_{\boldsymbol{\psi}}$ corresponds to the distribution of the Gaussian process $\tilde{Y}$, which has the same distribution as Y, except that its sub-processes over two sets $E_{\boldsymbol{v}^{(1)}}$ and $E_{\boldsymbol{v}^{(2)}}$ are independent for $\boldsymbol{v}^{(1)} \neq \boldsymbol{v}^{(2)}$. In lemma 5.32, we show that this approximation is asymptotically exact when the volume and the number of the sets $E_{\boldsymbol{v}}$ containing observation points goes to $+\infty$. This can be interpreted because, when the volume of the sets $E_{\boldsymbol{v}}$ is large, two observation points in two different $E_{\boldsymbol{v}^{(1)}}$ and $E_{\boldsymbol{v}^{(2)}}$ do yield almost independent observations, except for the rare case when both of the observation points are close to the borders of $E_{\boldsymbol{v}^{(1)}}$ and $E_{\boldsymbol{v}^{(2)}}$.

We denote $\tilde{\mathbf{M}}_{d_1}^{I_1}...\tilde{\mathbf{M}}_{d_K}^{I_K}$ the matrix built by replacing $\mathbf{K}_{\boldsymbol{\psi}}$ by $\tilde{\mathbf{K}}_{\boldsymbol{\psi}}$ in the expression of $\mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}$ (we also make the substitution for the inverse and the partial derivatives).

Lemma 5.32. $\left|\left|\tilde{\mathbf{M}}_{d_1}^{I_1}...\tilde{\mathbf{M}}_{d_K}^{I_K} - \mathbf{M}_{d_1}^{I_1}...\mathbf{M}_{d_K}^{I_K}\right|\right|_2^2 \to 0$, *uniformly in* $\boldsymbol{x} \in (S_X)^n$, *when* $N_1, n_2 \to \infty$.

Proof of lemma 5.32. Let $\delta > 0$ and N so that, with C_0 as in (5.3),

$$T_N := C_0^2 2^{2d} d^2 \sum_{j \in \mathbb{N}, j \geq N-1} \frac{\left(j + \frac{3}{2}\right)^{2(d-1)}}{\left(1 + j^{d+1}\right)^2} \leq \delta.$$

Then:

$$\begin{aligned}
\left|\left|\tilde{\mathbf{K}}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}}\right|\right|_2^2 &= \frac{1}{n} \sum_{i,j=1}^n \left\{ \left(\mathbf{K}_{\boldsymbol{\psi}}\right)_{i,j} - \left(\tilde{\mathbf{K}}_{\boldsymbol{\psi}}\right)_{i,j} \right\}^2 \\
&= \frac{1}{n} \sum_{1 \leq i,j \leq n, C\left(\boldsymbol{v}^{(i)}\right) \neq C\left(\boldsymbol{v}^{(j)}\right)} K_{\boldsymbol{\psi}}^2 \left\{ \boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon \left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right) \right\} \\
&\leq \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathbb{N}^*, C\left(\boldsymbol{v}^{(i)}\right) \neq C\left(\boldsymbol{v}^{(j)}\right)} K_{\boldsymbol{\psi}}^2 \left\{ \boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon \left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right) \right\}.
\end{aligned}$$

There exists a unique a so that $(aN_1)^d \leq n < \{(a+1) N_1\}^d$. Among the n deterministic observation points $\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(n)}$, $(aN_1)^d$ are in the $E_{\boldsymbol{v}}$, for $\boldsymbol{v} \in \{1, ..., a\}^d$. The number of remaining points is less than $\{(a+1) N_1\}^d - \{(a) N_1\}^d \leq d N_1 \{(a+1) N_1\}^{d-1}$ which is a $o((aN_1)^d)$ (because $a \to +\infty$ when $N_1, n_2 \to +\infty$), hence a $o(n)$. Therefore,

$$\begin{aligned}
&\left|\left|\tilde{\mathbf{K}}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}}\right|\right|_2^2 \leq \\
&\frac{1}{n} \sum_{\boldsymbol{v} \in \{1,...,a\}^d} \sum_{1 \leq i \leq n, \boldsymbol{v}^{(i)} \in E_{\boldsymbol{v}}} \sum_{j \in \mathbb{N}^*, C\left(\boldsymbol{v}^{(j)}\right) \neq C\left(\boldsymbol{v}^{(i)}\right)} K_{\boldsymbol{\psi}}^2 \left\{ \boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon \left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right) \right\} \\
&+ \frac{1}{n} d N_1 \{(a+1) N_1\}^{d-1} T_0 \\
&= \frac{1}{n} \sum_{\boldsymbol{v} \in \{1,...,a\}^d} \sum_{1 \leq i \leq n, \boldsymbol{v}^{(i)} \in E_{\boldsymbol{v}}} \sum_{j \in \mathbb{N}^*, C\left(\boldsymbol{v}^{(j)}\right) \neq C\left(\boldsymbol{v}^{(i)}\right)} K_{\boldsymbol{\psi}}^2 \left\{ \boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon \left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)} \right) \right\} \\
&+ o(1). \qquad (5.24)
\end{aligned}$$

Then, for fixed $\boldsymbol{v}$, the cardinality of the set of the integers $i \in \{1, ..., n\}$, so that $\boldsymbol{v}^{(i)} \in E_{\boldsymbol{v}}$ and there exists $j \in \mathbb{N}^*$ so that $C(\boldsymbol{v}^{(j)}) \neq C(\boldsymbol{v}^{(i)})$ and $|\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)}|_\infty \leq N$ is $N_1^d - (N_1 - 2N)^d$ and is less than $2NdN_1^{d-1}$.

Now, for the integers i so that for all $j \in \mathbb{N}^*$ so that $C(\boldsymbol{v}^{(j)}) \neq C(\boldsymbol{v}^{(i)})$, $|\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)}|_\infty \geq N$, we use the following lemma 5.33.

Lemma 5.33. *Let $f : \mathbb{R}^d \to \mathbb{R}^+$, so that $f(\boldsymbol{t}) \leq \frac{1}{1+|\boldsymbol{t}|_\infty^{d+1}}$. Then, for all $i \in \mathbb{N}^*$, $N \in \mathbb{N}^*$ and $\left(\boldsymbol{x}^{(i)}\right)_{i \in \mathbb{N}^*} \in S_X^{\mathbb{N}^*}$,*

$$\sum_{j \in \mathbb{N}^*, |\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)}|_\infty \geq N} f\left\{\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \epsilon\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right)\right\} \leq 2^d d \sum_{j \in \mathbb{N}, j \geq N-1} \frac{\left(j + \frac{3}{2}\right)^{d-1}}{1 + j^{d+1}}.$$

Proof of lemma 5.33. Similar to the proof of lemma 5.24. □

Hence, using (5.3) and lemmas 5.24 and 5.33, we get from (5.24),

$$\begin{aligned} \left|\left|\tilde{\mathbf{K}}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}}\right|\right|_2^2 &\leq \frac{1}{n} \sum_{\boldsymbol{v} \in \{1,...,a\}^d} \left(2NdN_1^{d-1}T_0 + N_1^d T_N\right) + o(1) \\ &\leq \frac{1}{a^d N_1^d} a^d \left\{\left(2NdN_1^{d-1}T_0 + N_1^d T_N\right)\right\} + o(1). \end{aligned}$$

This last term is smaller than 2δ for N_1 and n_2 large enough. Hence we showed $\left|\left|\tilde{\mathbf{K}}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}}\right|\right|_2 \to 0$ uniformly in $\boldsymbol{x}$, when $N_1, n_2 \to \infty$. We can show the same result for $\frac{\partial^k \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_{i_1} ... \partial \psi_{i_k}}$ and $\frac{\partial^k \tilde{\mathbf{K}}_{\boldsymbol{\psi}}}{\partial \psi_{i_1} ... \partial \psi_{i_k}}$. Finally we use (5.17) to show that $\left|\left|\tilde{\mathbf{K}}_{\boldsymbol{\psi}}^{-1} - \mathbf{K}_{\boldsymbol{\psi}}^{-1}\right|\right|_2 \to 0$ uniformly in $\boldsymbol{x}$, when $N_1, n_2 \to \infty$.

Using (5.19), together with lemma 5.30, we obtain that, for $d \in \{nd, sd, bd\}$ and $I \in S_{nd} \cup \{1\} \cup S_{bd}$, $||\tilde{\mathbf{M}}_d^I - \mathbf{M}_d^I||_2 \to 0$ uniformly in $\boldsymbol{x}$ when $N_1, n_2 \to +\infty$. Thus, still using (5.19), we obtain $\left|\left|\tilde{\mathbf{M}}_{d_1}^{I_1} ... \tilde{\mathbf{M}}_{d_K}^{I_K} - \mathbf{M}_{d_1}^{I_1} ... \mathbf{M}_{d_K}^{I_K}\right|\right|_2^2 \to 0$, uniformly in $\boldsymbol{x} \in (S_X)^n$, when $N_1, n_2 \to \infty$. □

We denote, for every N_1, n_2 and r, with $0 \leq r < N_1^d$, $n = N_1^d n_2 + r$ and $S_{N_1,n_2} := \frac{1}{n} Tr\left(\tilde{\mathbf{M}}_{d_1}^{I_1} ... \tilde{\mathbf{M}}_{d_K}^{I_K}\right)$, which is a sequence of real random variables defined on $(\Omega_X, \mathcal{F}_X, P_X)$ and indexed by N_1, n_2 and r. Using (5.18) and lemma 5.32, $|S_n - S_{N_1,n_2}| \to 0$ uniformly in $\boldsymbol{x}$ when $N_1, n_2 \to \infty$ (uniformly in r). As the matrices in the expression of S_{N_1,n_2} are block diagonal, we can write $S_{N_1,n_2} = \frac{1}{n_2} \sum_{l=1}^{n_2} S_{N_1^d}^l + o\left(\frac{1}{n_2}\right)$, where the $S_{N_1^d}^l$ are *iid* random variables defined on $(\Omega_X, \mathcal{F}_X, P_X)$ with the distribution of $S_{N_1^d}$. We denote $\bar{S}_{N_1^d} := E_{\boldsymbol{X}}\left(S_{N_1^d}\right)$. Then, using the strong law of large numbers, for fixed N_1, $S_{N_1,n_2} \to \bar{S}_{N_1^d}$ P_X-almost surely when $n_2 \to \infty$ (uniformly in r).

For every $N_1, p_{N_1}, n_2 \in \mathbb{N}^*$, there exists a unique (n_2', r), $n_2' \in \mathbb{N}^*$, $0 < r \leq N_1^d$ so that $(N_1 + p_{N_1})^d n_2 = N_1^d n_2' + r$. Then we have

$$\begin{aligned} |\bar{S}_{(N_1)^d} - \bar{S}_{\left(N_1+p_{N_1}\right)^d}| &\leq |\bar{S}_{(N_1)^d} - S_{N_1,n_2'}| + |S_{N_1,n_2'} - S_{N_1^d n_2'+r}| \\ &\quad + |S_{N_1^d n_2'+r} - S_{\left(N_1+p_{N_1}\right)^d n_2}| + |S_{\left(N_1+p_{N_1}\right)^d n_2} - S_{N_1+p_{N_1},n_2}| \\ &\quad + |S_{N_1+p_{N_1},n_2} - \bar{S}_{\left(N_1+p_{N_1}\right)^d}| \\ &= A + B + C + D + E. \end{aligned} \tag{5.25}$$

Because n_2' and r depend on N_1, p_{N_1} and n_2, A, B, C, D and E are sequences of random variables defined on $(\Omega_X, \mathcal{F}_X, P_X)$ and indexed by N_1, p_{N_1} and n_2. We have seen that there

exists $\tilde{\Omega}_X \subset \Omega_X$, with $P_X(\tilde{\Omega}_X) = 1$ so that for $\omega_X \in \tilde{\Omega}_X$, when $N_1, n_2 \to +\infty$, we also have $N_1 + p_{N_1}, n_2' \to +\infty$, and so B and D converge to zero.

Now, for every $N_1 \in \mathbb{N}^*$, let Ω_{X,N_1} be so that $P_X(\Omega_{X,N_1}) = 1$ and for all $\omega_X \in \Omega_{X,N_1}$, $S_{N_1,n_2} \to_{n_2 \to +\infty} \bar{S}_{N_1^d}$. Let $\tilde{\tilde{\Omega}}_X = \cap_{N_1 \in \mathbb{N}^*} \Omega_{X,N_1}$. Then $P_X(\tilde{\tilde{\Omega}}_X) = 1$ and for all $\omega_X \in \tilde{\tilde{\Omega}}_X$, for all $N_1 \in \mathbb{N}^*$, $S_{N_1,n_2} \to_{n_2 \to +\infty} \bar{S}_{N_1^d}$.

We will now show that $N_1 \to \bar{S}_{N_1^d}$ is a Cauchy sequence. Let $\delta > 0$. $P_X(\tilde{\tilde{\Omega}} \cap \tilde{\Omega}) = 1$ so this set in non-empty. Let us fix $\omega_X \in \tilde{\tilde{\Omega}} \cap \tilde{\Omega}$. In (5.25), C is null. There exist $\bar{N}_1$ and $\bar{n}_2$ so that for every $N_1 \geq \bar{N}_1$, $n_2 \geq \bar{n}_2$, $p_{N_1} > 0$, B and D are smaller than δ. Let us now fix any $N_1 \geq \bar{N}_1$. Then, for every $p_{N_1} > 0$, with $n_2 \geq \bar{n}_2$ large enough, A and E are smaller than δ.

Hence, we showed that, for the $\omega_X \in \tilde{\tilde{\Omega}} \cap \tilde{\Omega}$ we were considering, $N_1 \to \bar{S}_{(N_1)^d}$ is a Cauchy sequence and we denote its limit by S. Since $N_1 \to \bar{S}_{(N_1)^d}$ is deterministic, S is deterministic and $\bar{S}_{(N_1)^d} \to_{N_1 \to +\infty} S$.

Finally, let $n = N_1^d n_2 + r$ with $N_1, n_2 \to \infty$. Then

$$|S_n - S| \leq |S_n - S_{N_1,n_2}| + |S_{N_1,n_2} - \bar{S}_{N_1^d}| + |\bar{S}_{N_1^d} - S|.$$

Using the same arguments as before, we show that, P_X-a.s., $|S_n - S| \to 0$ as $n \to +\infty$.

Now, because of proposition 5.29, the eigenvalues of $\mathbf{M}_{d_1}^{I_1} ... \mathbf{M}_{d_K}^{I_K}$ are uniformly bounded in n and $\boldsymbol{x} \in S_X^n$. Thus, from the dominated convergence theorem, $S_n \to S$ in the mean square sense. Thus, $Var(S_n) \to 0$. □

Consistency

We now have gathered enough preliminary results to start addressing the consistency of ML and CV. We start by the proof of proposition 5.7 which addresses the consistency of ML.

Proof of proposition 5.7. We show that there exist sequences of random variables, defined on $(\Omega_X, \mathcal{F}_X, P_X)$, $D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ and $D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ (functions of n and $\boldsymbol{X}$), so that

$$\sup_{\boldsymbol{\psi}} \left| \left(L(\boldsymbol{\psi}) - L(\boldsymbol{\psi}^{(0)}) \right) - D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \right| \to_p 0$$

(in probability of the product space) and $D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \geq B D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ P_X-a.s. for a constant $B > 0$. We then show that there exists $D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$, a deterministic function of $\boldsymbol{\psi}, \boldsymbol{\psi}^{(0)}$ only, so that

$$\sup_{\boldsymbol{\psi}} \left| D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} - D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \right| = o_p(1)$$

and for any $t > 0$,

$$\inf_{|\boldsymbol{\psi} - \boldsymbol{\psi}^{(0)}| \geq t} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} > 0. \tag{5.26}$$

This implies consistency. Indeed, assume that there exists a sequence $N_n \to +\infty$, $\alpha, t > 0$ so that $P(|\hat{\boldsymbol{\psi}}_{ML,N_n} - \boldsymbol{\psi}^{(0)}| \geq t) \geq \alpha$. Let us write the Likelihood at step n L_n to emphasize the dependence on n. Then, since, with probability larger than α, $L_{N_n}(\hat{\boldsymbol{\psi}}_{ML,N_n}) \geq \inf_{|\boldsymbol{\psi} - \boldsymbol{\psi}^{(0)}| \geq t} L_{N_n}(\boldsymbol{\psi})$ and since $L_{N_n}(\hat{\boldsymbol{\psi}}_{ML,N_n}) \leq L_{N_n}(\boldsymbol{\psi}^{(0)})$, we get, with probability larger than α,

$$\inf_{|\boldsymbol{\psi} - \boldsymbol{\psi}^{(0)}| \geq t} L_{N_n}(\boldsymbol{\psi}) \leq L_{N_n}(\boldsymbol{\psi}^{(0)}).$$

Hence, with probability larger than α,

$$\begin{aligned}
0 &\geq \inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t}\left(L_{N_n}(\boldsymbol{\psi})-L_{N_n}(\boldsymbol{\psi}^{(0)})\right)\\
&\geq \inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} - \sup_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t}\left|L_{N_n}(\boldsymbol{\psi})-L_{N_n}(\boldsymbol{\psi}^{(0)})-D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}\right|\\
&= \inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} + o_p(1)\\
&\geq B\inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t} D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} + o_p(1)\\
&\geq B\inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} - B\sup_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t}\left|D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}-D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}\right| + o_p(1)\\
&= B\inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} + o_p(1).
\end{aligned}$$

Since $\inf_{|\boldsymbol{\psi}-\boldsymbol{\psi}^{(0)}|\geq t} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} > 0$, this is a contradiction.

We have $L(\boldsymbol{\psi}) = \frac{1}{n}\ln\{|\mathbf{K}_{\boldsymbol{\psi}}|\} + \frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y}$. The eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}$ and $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ are bounded uniformly in n and $\boldsymbol{x}$ (lemma 5.28) and from (5.20), $Var\left(L(\boldsymbol{\psi})|\boldsymbol{X}=\boldsymbol{x}\right) = \frac{2}{n^2}Tr\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)$. Thus $Var\left(L(\boldsymbol{\psi})|\boldsymbol{X}=\boldsymbol{x}\right)$ converges to 0 uniformly in $\boldsymbol{x}$, and so $L(\boldsymbol{\psi}) - \mathbb{E}\left(L(\boldsymbol{\psi})|\boldsymbol{X}\right)$ converges in probability $\mathbb{P}$ to zero.

Then, with $\boldsymbol{z} = \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-\frac{1}{2}}\boldsymbol{y}$,

$$\begin{aligned}
&\sup_{k\in\{1,\ldots,p\},\boldsymbol{\psi}\in\Psi}\left|\frac{\partial L(\boldsymbol{\psi})}{\partial\psi_k}\right| =\\
&\sup_{k\in\{1,\ldots,p\},\boldsymbol{\psi}\in\Psi}\frac{1}{n}\left\{Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_k}\right) - \boldsymbol{z}^t\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_k}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}}\boldsymbol{z}\right\}\\
&\leq \sup_{k\in\{1,\ldots,p\},\boldsymbol{\psi}}\left\{max\left(\left|\left|\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right|\right|\left|\left|\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_k}\right|\right|, ||\mathbf{K}_{\boldsymbol{\psi}^{(0)}}||\left|\left|\mathbf{K}_{\boldsymbol{\psi}}^{-2}\right|\right|\left|\left|\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_k}\right|\right|\right)\right\}\left(1+\frac{1}{n}|\boldsymbol{z}|^2\right),
\end{aligned}$$

and is hence bounded in probability conditionally to $\boldsymbol{X}=\boldsymbol{x}$, uniformly in $\boldsymbol{x}$, because of lemma 5.28 and the fact that $\boldsymbol{z}\sim\mathcal{N}(0,\mathbf{I}_n)$ given $\boldsymbol{X}=\boldsymbol{x}$ (so $\frac{1}{n}|\boldsymbol{z}|^2$ is bounded in probability given $\boldsymbol{X}=\boldsymbol{x}$).

Because of the simple convergence and the boundedness of the derivatives,

$$\sup_{\boldsymbol{\psi}}|L(\boldsymbol{\psi}) - \mathbb{E}\left(L(\boldsymbol{\psi})|\boldsymbol{X}\right)| \to_p 0.$$

We then denote

$$D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} := \mathbb{E}\left(L(\boldsymbol{\psi})|\boldsymbol{X}\right) - \mathbb{E}\left(L(\boldsymbol{\psi}^{(0)})|\boldsymbol{X}\right).$$

We then have $\sup_{\boldsymbol{\psi}}\left|\left(L(\boldsymbol{\psi})-L(\boldsymbol{\psi}^{(0)})\right) - D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}\right| \to_p 0$.

We have $\mathbb{E}\left(L(\boldsymbol{\psi})|\boldsymbol{X}\right) = \frac{1}{n}\ln\{|\mathbf{K}_{\boldsymbol{\psi}}|\} + \frac{1}{n}Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right)$ and hence, P_X-a.s.

$$\begin{aligned}
D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} &= \frac{1}{n}\ln\{|\mathbf{K}_{\boldsymbol{\psi}}|\} + \frac{1}{n}Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right) - \frac{1}{n}\ln\left\{\left|\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right|\right\} - 1\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[-\ln\left\{\phi_i\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}}\right)\right\} + \phi_i\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}}\right) - 1\right].
\end{aligned}$$

Using lemma 5.28, there exists $0 < a < b < +\infty$ so that for all $\boldsymbol{x}$, n, $\boldsymbol{\psi}$, $a < \phi_i \left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}} \right) < b$. We denote $f(t) = -\ln(t) + t - 1$. As f is minimal in 1, $f'(1) = 0$ and $f''(1) = 1$, there exists $A > 0$ so that, for $t \in [a, b]$, $f(t)$ is larger than $A(t-1)^2$. Then,

$$\begin{aligned} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} &\geq A \frac{1}{n} \sum_{i=1}^{n} \left\{ 1 - \phi_i \left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}} \right) \right\}^2 \\ &= A \frac{1}{n} Tr \left\{ \left(\mathbf{I} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{\frac{1}{2}} \right)^2 \right\} \\ &= A \frac{1}{n} \left|\left| \mathbf{K}_{\boldsymbol{\psi}}^{-\frac{1}{2}} \left(\mathbf{K}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}} \right) \mathbf{K}_{\boldsymbol{\psi}}^{-\frac{1}{2}} \right|\right|_2^2 . \end{aligned}$$

Then, as the eigenvalues of $\mathbf{K}_{\boldsymbol{\psi}}^{-\frac{1}{2}}$ are larger than $c > 0$, uniformly in n, $\boldsymbol{x}$ and $\boldsymbol{\psi}$, and with (5.16), we obtain, for some $B > 0$, and uniformly in n, $\boldsymbol{x}$ and $\boldsymbol{\psi}$,

$$D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \geq B ||\mathbf{K}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}||_2^2 := B D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}.$$

For $\epsilon = 0$, $D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ is deterministic and converges to

$$D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} := \sum_{\boldsymbol{v} \in \mathbb{Z}^d} \left\{ K_{\boldsymbol{\psi}}(\boldsymbol{v}) - K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v}) \right\}^2 . \tag{5.27}$$

$D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ is continuous in $\boldsymbol{\psi}$ because the series of term $\sup_{\boldsymbol{\psi}} |K_{\boldsymbol{\psi}}(\boldsymbol{v})|^2$, $\boldsymbol{v} \in \mathbb{Z}^d$ is summable using (5.3) and lemma 5.24. Hence, if there exists $\alpha > 0$ so that $\inf_{|\boldsymbol{\psi} - \boldsymbol{\psi}^{(0)}| \geq \alpha} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} = 0$, we can, using a compacity and continuity argument, have $\boldsymbol{\psi}_\infty \neq \boldsymbol{\psi}^{(0)}$ so that (5.27) is null. Hence we showed (5.26) by contradiction, which shows the proposition for $\epsilon = 0$.

For $\epsilon \neq 0$, $D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} = \frac{1}{n} Tr \left\{ \left(\mathbf{K}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}} \right)^2 \right\}$. With fixed $\boldsymbol{\psi}$, using proposition 5.31, $D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ converges in P_X-probability to $D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} := \lim_{n \to \infty} E_{\boldsymbol{X}} \left(D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \right)$. The eigenvalues of the $\frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i}$, $1 \leq i \leq n$, being bounded uniformly in n, $\boldsymbol{\psi}$, $\boldsymbol{x}$, the partial derivatives with respect to $\boldsymbol{\psi}$ of $D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ are uniformly bounded in n, $\boldsymbol{\psi}$ and $\boldsymbol{x}$. Hence $\sup_{\boldsymbol{\psi}} |D_{2,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} - D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}| = o_p(1)$. Then

$$\begin{aligned} &D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} = \\ &\lim_{n \to +\infty} \frac{1}{n} \sum_{1 \leq i,j \leq n, i \neq j} \left[\int_{\epsilon C_{S_X}} \left\{ K_{\boldsymbol{\psi}} \left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \boldsymbol{t} \right) - K_{\boldsymbol{\psi}^{(0)}} \left(\boldsymbol{v}^{(i)} - \boldsymbol{v}^{(j)} + \boldsymbol{t} \right) \right\}^2 f_T(\boldsymbol{t}) \, d\boldsymbol{t} \right] \\ &+ \left\{ K_{\boldsymbol{\psi}}(0) - K_{\boldsymbol{\psi}^{(0)}}(0) \right\}^2 , \end{aligned}$$

with $f_T(\boldsymbol{t})$ the probability density function of $\epsilon(X_i - X_j)$, $i \neq j$. We then show,

$$\begin{aligned} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} &= \sum_{\boldsymbol{v} \in \mathbb{Z}^d \setminus 0} \left[\int_{\epsilon C_{S_X}} \left\{ K_{\boldsymbol{\psi}}(\boldsymbol{v} + \boldsymbol{t}) - K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{v} + \boldsymbol{t}) \right\}^2 f_T(\boldsymbol{t}) \, d\boldsymbol{t} \right] \\ &\quad + \left\{ K_{\boldsymbol{\psi}}(0) - K_{\boldsymbol{\psi}^{(0)}}(0) \right\}^2 \qquad (5.28) \\ &= \int_{D_\epsilon} \left\{ K_{\boldsymbol{\psi}}(\boldsymbol{t}) - K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t}) \right\}^2 \tilde{f}_T(\boldsymbol{t}) \, d\boldsymbol{t} + \left\{ K_{\boldsymbol{\psi}}(0) - K_{\boldsymbol{\psi}^{(0)}}(0) \right\}^2 , \end{aligned}$$

where $\tilde{f}_T(\boldsymbol{t})$ is a positive-valued function, almost surely with respect to the Lebesgue measure on D_ϵ. As $\sup_{\boldsymbol{\psi}} |K_{\boldsymbol{\psi}}(t)|^2$ is summable on D_ϵ, using (5.3), $D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$ is continuous. Hence, if there exists $\alpha > 0$ so that $\inf_{|\boldsymbol{\psi} - \boldsymbol{\psi}^{(0)}| \geq \alpha} D_{\infty,\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} = 0$, we can, using a compacity and continuity argument, show that there exists $\boldsymbol{\psi}_\infty \neq \boldsymbol{\psi}^{(0)}$ so that (5.28) is null. Hence we proved (5.26) by contradiction which proves the proposition for $\epsilon \neq 0$. □

We now prove proposition 5.11, addressing the consistency of CV.

Proof of proposition 5.11. We will show that there exists a sequence of random variables $D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}}$, defined on $(\Omega_X, \mathcal{F}_X, P_X)$, so that

$$\sup_{\boldsymbol{\psi}} \left| \left(LOO(\boldsymbol{\psi}) - LOO(\boldsymbol{\psi}^{(0)}) \right) - D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \right| \rightarrow_p 0$$

and $C > 0$ so that P_X-a.s.

$$D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \geq C ||\mathbf{K}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}||_2^2. \tag{5.29}$$

The proof of the proposition is then carried out similarly to the proof of proposition 5.7.

We firstly show, similarly to the proof of proposition 5.7 that $\sup_{\boldsymbol{\psi}} |LOO(\boldsymbol{\psi}) - \mathbb{E}\left(LOO(\boldsymbol{\psi})|\boldsymbol{X}\right)| \rightarrow_p 0$. We then denote $D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} = \mathbb{E}\left(LOO(\boldsymbol{\psi})|\boldsymbol{X}\right) - \mathbb{E}\left(LOO(\boldsymbol{\psi}^{(0)})|\boldsymbol{X}\right)$. We decompose, for all $i \in \{1, ..., n\}$, with $\mathbf{P}_i$ the matrix that exchanges lines 1 and i of a matrix,

$$\mathbf{P}_i \mathbf{K}_{\boldsymbol{\psi}} \mathbf{P}_i^t = \begin{pmatrix} 1 & \boldsymbol{k}_{i,\boldsymbol{\psi}}^t \\ \boldsymbol{k}_{i,\boldsymbol{\psi}} & \mathbf{K}_{-i,\boldsymbol{\psi}} \end{pmatrix}.$$

The conditional distributions being independent on the numbering of the observations, we have, from the Kriging equations (2.9) and (2.10), denoting

$$\boldsymbol{y}_{-i} = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)^t$$

and using $\mathbb{E}((\hat{y}_{i,\boldsymbol{\psi}} - y_i)^2|\boldsymbol{X}) = \mathbb{E}((\hat{y}_{i,\boldsymbol{\psi}} - \hat{y}_{i,\boldsymbol{\psi}^{(0)}})^2|\boldsymbol{X}) + \mathbb{E}((\hat{y}_{i,\boldsymbol{\psi}^{(0)}} - y_i)^2|\boldsymbol{X})$,

$$\begin{aligned} & D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ \left(\boldsymbol{k}_{i,\boldsymbol{\psi}}^t \mathbf{K}_{-i,\boldsymbol{\psi}}^{-1} \boldsymbol{y}_{-i} - \boldsymbol{k}_{i,\boldsymbol{\psi}^{(0)}}^t \mathbf{K}_{-i,\boldsymbol{\psi}^{(0)}}^{-1} \boldsymbol{y}_{-i} \right)^2 |\boldsymbol{X} \right\} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{k}_{i,\boldsymbol{\psi}}^t \mathbf{K}_{-i,\boldsymbol{\psi}}^{-1} - \boldsymbol{k}_{i,\boldsymbol{\psi}^{(0)}}^t \mathbf{K}_{-i,\boldsymbol{\psi}^{(0)}}^{-1} \right) \mathbf{K}_{-i,\boldsymbol{\psi}^{(0)}} \left(\mathbf{K}_{-i,\boldsymbol{\psi}}^{-1} \boldsymbol{k}_{i,\boldsymbol{\psi}} - \mathbf{K}_{-i,\boldsymbol{\psi}^{(0)}}^{-1} \boldsymbol{k}_{i,\boldsymbol{\psi}^{(0)}} \right). \end{aligned}$$

Similarly to lemma 5.28, it can be shown that the eigenvalues of $\mathbf{K}_{-i,\boldsymbol{\psi}^{(0)}}$ are larger than a constant $A > 0$, uniformly in n and $\boldsymbol{x}$. Then

$$D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \quad \geq \quad A \frac{1}{n} \sum_{i=1}^{n} \left| \left(\boldsymbol{k}_{i,\boldsymbol{\psi}}^t \mathbf{K}_{-i,\boldsymbol{\psi}}^{-1} - \boldsymbol{k}_{i,\boldsymbol{\psi}^{(0)}}^t \mathbf{K}_{-i,\boldsymbol{\psi}^{(0)}}^{-1} \right) \right|^2.$$

Using the virtual Cross Validation equations of proposition 2.35, the vector $\mathbf{K}_{-i,\boldsymbol{\psi}}^{-1} \boldsymbol{k}_{i,\boldsymbol{\psi}}$ is the vector of the $\frac{\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)_{i,j}}{\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)_{i,i}}$ for $1 \leq j \leq n$, $j \neq i$. Hence P_X-a.s.

$$\begin{aligned} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} \quad &\geq \quad A \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \left\{ \frac{\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)_{i,j}}{\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)_{i,i}} - \frac{\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)_{i,j}}{\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)_{i,i}} \right\}^2 \\ &= \quad A \frac{1}{n} \left|\left| Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1} \mathbf{K}_{\boldsymbol{\psi}}^{-1} - Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)^{-1} \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \right|\right|_2^2 \\ &\geq \quad AB \frac{1}{n} \left|\left| Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right) Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1} \mathbf{K}_{\boldsymbol{\psi}}^{-1} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \right|\right|_2^2, \end{aligned}$$

with $B = \inf_{i,n,x} \phi_i^2 \left\{ Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right)^{-1} \right\}$, $B > 0$. The eigenvalues of $Diag\left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1}\right) Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-1}$ are bounded between $a > 0$ and $b < \infty$ uniformly in n and $\boldsymbol{x}$. Hence we have, with $\mathbf{D}_{\boldsymbol{\lambda}}$, the diagonal matrix with values $\lambda_1, ..., \lambda_n$,

$$\begin{aligned} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} &\geq AB \inf_{a \leq \lambda_1,...,\lambda_n \leq b} \left|\left| \mathbf{D}_{\boldsymbol{\lambda}} \mathbf{K}_{\boldsymbol{\psi}}^{-1} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-1} \right|\right|_2^2 \\ &\geq ABC \inf_{a \leq \lambda_1,...,\lambda_n \leq b} ||\mathbf{D}_{\boldsymbol{\lambda}}^{-1} \mathbf{K}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}||_2^2, \quad \text{using (5.17)}, \\ &\geq ABC \inf_{\lambda_1,...,\lambda_n} ||\mathbf{D}_{\lambda} \mathbf{K}_{\boldsymbol{\psi}} - \mathbf{K}_{\boldsymbol{\psi}^{(0)}}||_2^2, \end{aligned}$$

with $C = \frac{1}{b} \inf_{n,x,\boldsymbol{\psi}} \frac{1}{||\mathbf{K}_{\boldsymbol{\psi}}||^2} \frac{1}{||\mathbf{K}_{\boldsymbol{\psi}^{(0)}}||^2}$, $C > 0$. Then

$$\begin{aligned} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} &\geq ABC \frac{1}{n} \inf_{\lambda_1,...,\lambda_n} \sum_{i,j=1}^{n} \left(\lambda_i K_{\boldsymbol{\psi},i,j} - K_{\boldsymbol{\psi}^{(0)},i,j}\right)^2 \\ &= ABC \frac{1}{n} \sum_{i=1}^{n} \inf_{\lambda} \sum_{j=1}^{n} \left(\lambda K_{\boldsymbol{\psi},i,j} - K_{\boldsymbol{\psi}^{(0)},i,j}\right)^2 \\ &= ABC \frac{1}{n} \sum_{i=1}^{n} \inf_{\lambda} \left\{ (\lambda - 1)^2 + \sum_{j \neq i} \left(\lambda K_{\boldsymbol{\psi},i,j} - K_{\boldsymbol{\psi}^{(0)},i,j}\right)^2 \right\}. \end{aligned}$$

We show how to treat the infimum over λ in the following lemma.

Lemma 5.34. *For any $a_1, ..., a_n$ and $b_1, ..., b_n \in \mathbb{R}$,*

$$\inf_{\lambda} \left\{ (\lambda - 1)^2 + \sum_{i=1}^{n} (a_i - \lambda b_i)^2 \right\} \geq \frac{\sum_{i=1}^{n} (a_i - b_i)^2}{1 + \sum_{i=1}^{n} b_i^2}.$$

Proof.

$$(\lambda - 1)^2 + \sum_{i=1}^{n} (a_i - \lambda b_i)^2 = \lambda^2 \left(1 + \sum_{i=1}^{n} b_i^2\right) - 2\lambda \left(1 + \sum_{i=1}^{n} a_i b_i\right) + \left(1 + \sum_{i=1}^{n} a_i^2\right).$$

The minimum in t of $at^2 - 2bt + c$, is $-\frac{b^2}{a} + c$, hence

$$\begin{aligned} (\lambda - 1)^2 + \sum_{i=1}^{n} (a_i - \lambda b_i)^2 &\geq \left(1 + \sum_{i=1}^{n} a_i^2\right) - \frac{\left(1 + \sum_{i=1}^{n} a_i b_i\right)^2}{\left(1 + \sum_{i=1}^{n} b_i^2\right)} \\ &= \frac{\sum_{i=1}^{n} (a_i - b_i)^2 - \left(\sum_{i=1}^{n} a_i b_i\right)^2 + \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)}{1 + \sum_{i=1}^{n} b_i^2} \\ &\geq \frac{\sum_{i=1}^{n} (a_i - b_i)^2}{1 + \sum_{i=1}^{n} b_i^2}, \quad \text{using Cauchy-Schwartz inequality.} \end{aligned}$$

□

Using lemma 5.34, together with (5.3) and lemma 5.24 which ensure that

$$\sum_{j \neq i} (K_{\boldsymbol{\psi},i,j})^2 \leq c < +\infty$$

uniformly in i, $\boldsymbol{\psi}$ and $\boldsymbol{x}$, we obtain

$$\begin{aligned} D_{\boldsymbol{\psi},\boldsymbol{\psi}^{(0)}} &\geq ABC\frac{1}{1+c}\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\left(K_{\boldsymbol{\psi},i,j}-K_{\boldsymbol{\psi}^{(0)},i,j}\right)^2 \\ &= ABC\frac{1}{1+c}||\mathbf{K}_{\boldsymbol{\psi}}-\mathbf{K}_{\boldsymbol{\psi}^{(0)}}||_2^2, \quad \text{because } K_{\boldsymbol{\psi},i,i}=1=K_{\boldsymbol{\psi}^{(0)},i,i}, \end{aligned}$$

which proves (5.29) and ends the proof. □

Convergence of gradients and Hessians for ML and CV

Now, based notably on the preliminary results on the almost sure convergence of random traces, we study the convergence in distribution of the gradients of ML and CV to normal distribution, and the convergence in probability of the Hessian matrices to constant matrices.

We start by proving proposition 5.12 which gives the expressions and some convergence results for the gradient and Hessian matrix for CV.

Proof of proposition 5.12. It is shown in proposition 3.33 that

$$\frac{\partial}{\partial\psi_i}LOO(\boldsymbol{\psi})=\frac{2}{n}\boldsymbol{y}^t\mathbf{M}_{\boldsymbol{\psi}}^i\boldsymbol{y}=\frac{1}{n}\boldsymbol{y}^t\left\{\mathbf{M}_{\boldsymbol{\psi}}^i+\left(\mathbf{M}_{\boldsymbol{\psi}}^i\right)^t\right\}\boldsymbol{y}.$$

From (5.20) we show (5.7). A straightforward but relatively long calculation then shows

$$\begin{aligned} &\frac{\partial^2}{\partial\psi_i\partial\psi_j}LOO(\boldsymbol{\psi})= \\ &-4\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-3}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &-4\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-3}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &+2\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &+6\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-4}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &-4\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-3}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &+2\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-3}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial^2\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &+2\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &+2\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y} \\ &-2\frac{1}{n}\boldsymbol{y}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}Diag\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)^{-2}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial^2\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i\partial\psi_j}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y}. \end{aligned}$$

We then have, using $\mathbb{E}\left(\boldsymbol{y}^t\mathbf{A}\boldsymbol{y}|\boldsymbol{X}\right)=Tr\left(\mathbf{A}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right)$ and for matrices $\mathbf{D}$, $\mathbf{M}_1$ and $\mathbf{M}_2$, with $\mathbf{D}$ diagonal, $Tr\left\{\mathbf{M}_1\mathbf{D}Diag\left(\mathbf{M}_2\right)\right\}=Tr\left\{\mathbf{M}_2\mathbf{D}Diag\left(\mathbf{M}_1\right)\right\}$ and $Tr\left(\mathbf{D}\mathbf{M}_1\right)=Tr\left(\mathbf{D}\mathbf{M}_1^t\right)$,

$$
\begin{aligned}
&\mathbb{E}\left(\frac{\partial^2}{\partial\psi_i\partial\psi_j}LOO(\boldsymbol{\psi}^{(0)})|\boldsymbol{X}\right) = \qquad (5.30)\\
&-8\frac{1}{n}Tr\left\{\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-3}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
&+2\frac{1}{n}Tr\left\{\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-2}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
&+\\
&\frac{6}{n}Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-4}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
&-4\frac{1}{n}Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-3}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
&+2\frac{1}{n}Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-3}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial^2\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
&+4\frac{1}{n}Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-2}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
&-2\frac{1}{n}Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-2}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial^2\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}.
\end{aligned}
$$

The fourth and sixth terms of (5.30) are opposite and hence cancel each other. Indeed,

$$
\begin{aligned}
&Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-3}Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}\\
=&\sum_{i=1}^{n}\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-3}_{i,i}\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)_{i,i}\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)_{i,i}\\
=&\sum_{i=1}^{n}\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-2}_{i,i}\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)_{i,i}\\
=&Tr\left\{Diag\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right)^{-2}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_i}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi_j}\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\right\}.
\end{aligned}
$$

Similarly the fifth and seventh terms of (5.30) cancel each other.

Hence, we show the expression of $\mathbb{E}\left(\frac{\partial^2}{\partial\psi_i\partial\psi_j}LOO(\boldsymbol{\psi}^{(0)})|\boldsymbol{X}\right)$ of the proposition.

We use proposition 5.31 to show the existence of $\boldsymbol{\Sigma}_{CV,1}$ and $\boldsymbol{\Sigma}_{CV,2}$. □

The next proposition 5.35 enables us to prove the convergence in probability of quadratic forms with random matrices.

Proposition 5.35. *Assume that condition 5.1 is satisfied.*

Let $\mathbf{M}\in\mathcal{M}_{\boldsymbol{\psi}}$ *(proposition 5.29). Then,* $\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}$ *converges to*

$$\Sigma := \lim_{n\to+\infty}\frac{1}{n}Tr\left(\mathbf{M}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right),$$

in the mean square sense (on the product space).

Proof of proposition 5.35.

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}\right) &= \mathbb{E}\left\{\mathbb{E}\left(\tfrac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}|\boldsymbol{X}\right)\right\}\\ &= \mathbb{E}\left\{\tfrac{1}{n}Tr\left(\mathbf{M}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right)\right\}\\ &\to \Sigma \quad \text{(proposition 5.31).}\end{aligned}$$

Furthermore

$$Var\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}\right) = \mathbb{E}\left\{Var\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}|\boldsymbol{X}\right)\right\} + Var\left\{\mathbb{E}\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}|\boldsymbol{X}\right)\right\}.$$

The quantity $Var\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x}\right)$ is a $O\left(\frac{1}{n}\right)$, uniformly in $\boldsymbol{x}$, using proposition 5.29 and $||\mathbf{A}+\mathbf{B}|| \leq ||\mathbf{A}||+||\mathbf{B}||$. Therefore $Var\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}|\boldsymbol{X}\right)$ is bounded by $O(\frac{1}{n})$ P_X-a.s. Furthermore, $Var\left\{\mathbb{E}\left(\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}|\boldsymbol{X}\right)\right\} = Var\left\{\frac{1}{n}Tr\left(\mathbf{M}\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\right)\right\} \to 0$, using proposition 5.31. Hence $\frac{1}{n}\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}$ converges to Σ in the mean square sense. □

Proposition 5.36 enables us to quantify the small deviations of random quadratic forms from their limits by giving a convergence in distribution result.

Proposition 5.36. *Assume that condition 5.1 is satisfied.*

We recall $\boldsymbol{X} \sim \mathcal{L}_X^{\otimes n}$ and $y_i = Y\left(\boldsymbol{v}^{(i)} + \epsilon X_i\right)$, $1 \leq i \leq n$. We consider symmetric matrix sequences $\mathbf{M}_1, ..., \mathbf{M}_p$ and $\mathbf{N}_1, ..., \mathbf{N}_p$ (defined on $(\Omega_X, \mathcal{F}_X, P_X)$), functions of $\boldsymbol{X}$, so that the eigenvalues of $\mathbf{N}_1, ..., \mathbf{N}_p$ are bounded uniformly in n and $\boldsymbol{x} \in (S_X)^n$, $Tr\left(\mathbf{M}_i + \mathbf{N}_i\mathbf{K}\right) = 0$ for $1 \leq i \leq p$ and there exists a $p \times p$ matrix $\boldsymbol{\Sigma}$ so that $\frac{1}{n}Tr\left(\mathbf{N}_i\mathbf{K}\mathbf{N}_j\mathbf{K}\right) \to (\boldsymbol{\Sigma})_{i,j}$ P_X-almost surely. Then the sequence of p-dimensional random vectors (defined on the product space) $\left(\frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_i\right) + \boldsymbol{y}^t\mathbf{N}_i\boldsymbol{y}\right\}\right)_{i=1...p}$ converges in distribution to a Gaussian random vector with mean zero and covariance matrix $2\boldsymbol{\Sigma}$.

Proof of proposition 5.36. Let $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$.

$$\begin{aligned}&\mathbb{E}\left(\exp\left[\mathrm{i}\sum_{k=1}^p \lambda_k \frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right) + \boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}\right]\right)\\ &= \mathbb{E}\left\{\mathbb{E}\left(\exp\left[\mathrm{i}\sum_{k=1}^p \lambda_k \frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right) + \boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}\right]\middle|\boldsymbol{X}\right)\right\}.\end{aligned}$$

For fixed $\boldsymbol{x} = \left(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}\right) \in (S_X)^n$, denoting $\sum_{k=1}^p \lambda_k \mathbf{K}^{\frac{1}{2}}\mathbf{N}_k\mathbf{K}^{\frac{1}{2}} = \mathbf{P}^t\mathbf{D}\mathbf{P}$, with $\mathbf{P}^t\mathbf{P} = \mathbf{I}_n$ and $\mathbf{D}$ diagonal, $\boldsymbol{z}_{\boldsymbol{x}} = \mathbf{P}\mathbf{K}^{-\frac{1}{2}}\boldsymbol{y}$ (which is a vector of *iid* standard Gaussian variables, conditionally to $\boldsymbol{X} = \boldsymbol{x}$), we have

$$\begin{aligned}&\sum_{k=1}^p \lambda_k \frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right) + \boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}\\ &= \frac{1}{\sqrt{n}}\left[Tr\left(\sum_{k=1}^p \lambda_k\mathbf{M}_k\right) + \sum_{i=1}^n \phi_i\left(\sum_{k=1}^p \lambda_k\mathbf{K}^{\frac{1}{2}}\mathbf{N}_k\mathbf{K}^{\frac{1}{2}}\right)z_{\boldsymbol{x},i}^2\right]\\ &= \frac{1}{\sqrt{n}}\left[\sum_{i=1}^n \phi_i\left(\sum_{k=1}^p \lambda_k\mathbf{K}^{\frac{1}{2}}\mathbf{N}_k\mathbf{K}^{\frac{1}{2}}\right)\left\{z_{\boldsymbol{x},i}^2 - 1\right\}\right].\end{aligned}$$

Hence

$$\begin{aligned} Var\left[\sum_{k=1}^{p}\lambda_k\frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right)+\boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}\left|\boldsymbol{X}\right.\right] &= \frac{2}{n}\sum_{i=1}^{n}\phi_i^2\left(\sum_{k=1}^{p}\lambda_k\mathbf{K}^{\frac{1}{2}}\mathbf{N}_k\mathbf{K}^{\frac{1}{2}}\right) \\ &= \frac{2}{n}\sum_{k=1}^{p}\sum_{l=1}^{p}\lambda_k\lambda_l Tr\left(\mathbf{K}\mathbf{N}_k\mathbf{K}\mathbf{N}_l\right) \\ &\underset{n\to+\infty}{\to} \boldsymbol{\lambda}^t\left(2\boldsymbol{\Sigma}\right)\boldsymbol{\lambda} \quad \text{for a.e. } \omega_X\ . \end{aligned}$$

Hence, for almost every ω_X, we can apply Lindeberg-Feller criterion (theorem 5.37 below) to the Ω_Y-measurable variables $\frac{1}{\sqrt{n}}\phi_i\left(\sum_{k=1}^{p}\lambda_k\mathbf{K}^{\frac{1}{2}}\mathbf{N}_k\mathbf{K}^{\frac{1}{2}}\right)\left\{z_{\boldsymbol{x},i}^2-1\right\}$, $1\leq i\leq n$, to show that $\sum_{k=1}^{p}\lambda_k\frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right)+\boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}$ converges in distribution to $\mathcal{N}\left(0,\boldsymbol{\lambda}^t\left(2\boldsymbol{\Sigma}\right)\boldsymbol{\lambda}\right)$. Hence, for almost every ω_X,

$$\mathbb{E}\left(\exp\left[\mathrm{i}\sum_{k=1}^{p}\lambda_k\frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right)+\boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}\right]\left|\boldsymbol{X}\right.\right)\to_{n\to+\infty}\exp\left(-\frac{1}{2}\boldsymbol{\lambda}^t\left(2\boldsymbol{\Sigma}\right)\boldsymbol{\lambda}\right).$$

Using the dominated convergence theorem on $(\Omega_X,\mathcal{F}_X,P_X)$,

$$\mathbb{E}\left(\exp\left[\mathrm{i}\sum_{k=1}^{p}\lambda_k\frac{1}{\sqrt{n}}\left\{Tr\left(\mathbf{M}_k\right)+\boldsymbol{y}^t\mathbf{N}_k\boldsymbol{y}\right\}\right]\right)\to_{n\to+\infty}\exp\left\{-\frac{1}{2}\boldsymbol{\lambda}^t\left(2\boldsymbol{\Sigma}\right)\boldsymbol{\lambda}\right\}.$$

Theorem 5.37 (Lindeberg-Feller: see e.g. proposition 2.27 in [Van98])**.** *Let, for all $n\in\mathbb{N}^*$, $y_{n,1},...,y_{n,n}$ be centered independent random variables with zero mean and finite variances $\sigma_{n,1}^2,...,\sigma_{n,n}^2$. Assume that for any $\epsilon>0$, $\sum_{i=1}^{n}\mathbb{E}\left(y_{i,n}^2\mathbf{1}_{|y_{i,n}|>\epsilon}\right)$ goes to zero as $n\to+\infty$. Assume also that $\sum_{i=1}^{n}\sigma_{i,n}^2$ goes to a constant σ^2 as $n\to+\infty$. Then*

$$\sum_{i=1}^{n}y_{i,n}\to_{\mathcal{L}}\mathcal{N}(0,\sigma^2).$$

□

Asymptotic normality

We now show the asymptotic normality for ML and CV, based on the preliminary convergence results for their gradients and Hessians.

Proposition 5.38 enables us, from such convergence results for the gradient and Hessian of an estimator, to prove its asymptotic normality.

Proposition 5.38. *We recall $\boldsymbol{X}\sim\mathcal{L}_X^{\otimes n}$ and $y_i=Y\left(\boldsymbol{v}^{(i)}+\epsilon X_i\right)$, $1\leq i\leq n$. We consider a consistent estimator $\hat{\boldsymbol{\psi}}\in\mathbb{R}^p$ so that $\mathbb{P}\left(c\left(\hat{\boldsymbol{\psi}}\right)=0\right)\to 1$, for a function $c:\Psi\to\mathbb{R}^p$, dependent on $\boldsymbol{X}$ and Y, and twice differentiable in $\boldsymbol{\psi}$. We assume that $\sqrt{n}c\left(\boldsymbol{\psi}^{(0)}\right)\to_{\mathcal{L}}\mathcal{N}\left(0,\boldsymbol{\Sigma}_1\right)$, for a $p\times p$ matrix $\boldsymbol{\Sigma}_1$ and that the matrix $\frac{\partial c\left(\boldsymbol{\psi}^{(0)}\right)}{\partial\boldsymbol{\psi}}$ converges in probability to a $p\times p$ positive matrix $\boldsymbol{\Sigma}_2$ (convergences are defined on the product space). Finally we assume that $\sup_{\tilde{\boldsymbol{\psi}},i,j,k}\left|\frac{\partial^2}{\partial\psi_i\partial\psi_j}c_k\left(\tilde{\boldsymbol{\psi}}\right)\right|$ is bounded in probability.*

Then

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}^{(0)}\right)\to_{\mathcal{L}}\mathcal{N}\left(0,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right).$$

Proof of proposition 5.38. It is enough to consider the case $c\left(\hat{\boldsymbol{\psi}}\right) = 0$, the case $P\left\{c\left(\hat{\boldsymbol{\psi}}\right) = 0\right\} \to 1$ being deduced from it by modifying c on a set with vanishing probability measure, which does not affect the convergence in distribution. Indeed, if a random vector $\boldsymbol{v}$ converges in distribution to a distribution $\mathcal{L}$, the random vector $\boldsymbol{v}\mathbf{1}_{A_n} + \tilde{\boldsymbol{v}}\mathbf{1}_{\overline{A}_n}$, where $P(A_n) \to 1$ and $\tilde{\boldsymbol{v}}$ is an arbitrary random vector, also converges in distribution to $\mathcal{L}$.

For all $1 \leq k \leq p$,

$$0 = c_k\left(\hat{\boldsymbol{\psi}}\right) = c_k\left(\boldsymbol{\psi}^{(0)}\right) + \left\{\frac{\partial}{\partial \boldsymbol{\psi}} c_k\left(\boldsymbol{\psi}^{(0)}\right)\right\}^t \left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{(0)}\right) + r,$$

with random r, so that $|r| \leq \sup_{\tilde{\boldsymbol{\psi}},i,j,k} \left|\frac{\partial^2}{\partial\psi_i \partial\psi_j} c_k\left(\tilde{\boldsymbol{\psi}}\right)\right| \times |\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{(0)}|^2$. Hence $r = o_p\left(|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{(0)}|\right)$. We then have

$$-c_k\left(\boldsymbol{\psi}_0\right) = \left[\left\{\frac{\partial}{\partial \boldsymbol{\psi}} c_k\left(\boldsymbol{\psi}^{(0)}\right)\right\}^t + o_p\left(1\right)\right]\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{(0)}\right),$$

and so

$$\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^{(0)}\right) = -\left\{\frac{\partial}{\partial \boldsymbol{\psi}} c\left(\boldsymbol{\psi}^{(0)}\right) + o_p\left(1\right)\right\}^{-1} c\left(\boldsymbol{\psi}^{(0)}\right). \tag{5.31}$$

We conclude using Slutsky lemma, with $\left(\frac{\partial}{\partial \boldsymbol{\psi}} c(\boldsymbol{\psi}^{(0)}) + o_p(1)\right)^{-1}$ converges in probability to $\boldsymbol{\Sigma}_2^{-1}$ and $\sqrt{n} c(\boldsymbol{\psi}^{(0)})$ converges in distribution to a $\mathcal{N}(0, \boldsymbol{\Sigma}_1)$ distribution.

Remark 5.39. *One can show that, with probability going to one as $n \to +\infty$, the likelihood has a unique global minimizer. Indeed, we first notice that the set of the minimizers is a subset of any open ball of center $\boldsymbol{\psi}^{(0)}$ with probability going to one. For a small enough open ball, the probability that the likelihood function is strictly convex on this open ball converges to one. This is because of the third-order regularity of the likelihood with respect to $\boldsymbol{\psi}$, and because the limit of the second derivative matrix of the Likelihood at $\boldsymbol{\psi}^{(0)}$ is positive.*

□

We now prove proposition 5.8, on the asymptotic normality of ML.

Proof of proposition 5.8. For $1 \leq i, j \leq p$, we use proposition 5.31 to show that

$$\frac{1}{n} Tr\left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_j}\right)$$

has a P_X-almost sure limit as $n \to +\infty$.

We calculate $\frac{\partial}{\partial \psi_i} L(\boldsymbol{\psi}) = \frac{1}{n}\left\{Tr\left(\mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i}\right) - \boldsymbol{y}^t \mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \boldsymbol{y}\right\}$. We use proposition 5.36 with $\mathbf{M}_i = \mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i}$ and $\mathbf{N}_i = -\mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \psi_i} \mathbf{K}_{\boldsymbol{\psi}}^{-1}$, together with proposition 5.31, to show that

$$\sqrt{n} \frac{\partial}{\partial \boldsymbol{\psi}} L(\boldsymbol{\psi}^{(0)}) \to_{\mathcal{L}} \mathcal{N}\left(0, 2\boldsymbol{\Sigma}_{ML}\right).$$

We calculate

$$\begin{aligned} \frac{\partial^2}{\partial \psi_i \partial \psi_j} L(\boldsymbol{\psi}^{(0)}) &= \frac{1}{n} Tr\left(-\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_j} + \mathbf{K}^{-1} \frac{\partial^2 \mathbf{K}}{\partial \psi_i \partial \psi_j}\right) \\ &+ \frac{1}{n} \boldsymbol{y}^t \left(2\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \psi_j} \mathbf{K}^{-1} - \mathbf{K}^{-1} \frac{\partial^2 \mathbf{K}}{\partial \psi_i \partial \psi_j} \mathbf{K}^{-1}\right) \boldsymbol{y}. \end{aligned}$$

Hence, using proposition 5.35, $\frac{\partial^2}{\partial \boldsymbol{\psi}^2} L(\boldsymbol{\psi}^{(0)})$ converges to $\boldsymbol{\Sigma}_{ML}$ in the mean square sense (on the product space).

Finally, $\frac{\partial^3}{\partial\psi_i \partial\psi_j \partial\psi_k} L(\tilde{\boldsymbol{\psi}})$ can be written as $\frac{1}{n}\left\{Tr\left(\mathbf{M}_{\tilde{\boldsymbol{\psi}}}\right) + \boldsymbol{z}^t \mathbf{N}_{\tilde{\boldsymbol{\psi}}} \boldsymbol{z}\right\}$, where $\mathbf{M}_{\tilde{\boldsymbol{\psi}}}$ and $\mathbf{N}_{\tilde{\boldsymbol{\psi}}}$ are sums of matrices of $\mathcal{M}_{\tilde{\boldsymbol{\psi}}}$ (proposition 5.29) and where $\boldsymbol{z}$ depends on $\boldsymbol{X}$ and Y and $\mathcal{L}(\boldsymbol{z}|\boldsymbol{X}) = \mathbf{N}(0, \mathbf{I}_n)$. Hence, the singular values of $\mathbf{M}_{\tilde{\boldsymbol{\psi}}}$ and $\mathbf{N}_{\tilde{\boldsymbol{\psi}}}$ are bounded uniformly in $\tilde{\boldsymbol{\psi}}$, n and $\boldsymbol{x}$, and so $\sup_{i,j,k,\tilde{\boldsymbol{\psi}}} \frac{\partial^3}{\partial\psi_i \partial\psi_j \partial\psi_k} L(\tilde{\boldsymbol{\psi}})$ is bounded by $a + b\frac{1}{n}|\boldsymbol{z}|^2$, with constant $a, b < +\infty$ and is hence bounded in probability. Hence we apply proposition 5.38 to conclude. □

Asymptotic normality for CV is now addressed by proving proposition 5.13.

Proof of proposition 5.13. We use proposition 5.36, with

$$\mathbf{N}_i = -\left\{\mathbf{M}^i + \left(\mathbf{M}^i\right)^t\right\},$$

where $\mathbf{M}^i$ is the notation of proposition 5.12, together with propositions 5.29, 5.31 and 5.12 to show that

$$\sqrt{n}\frac{\partial}{\partial \boldsymbol{\psi}} LOO(\boldsymbol{\psi}^{(0)}) \to_{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma}_{CV,1}).$$

We have seen in the proof of proposition 5.12 that there exist matrices $\mathbf{P}_{i,j}$ in $\mathcal{M}_{\boldsymbol{\psi}^{(0)}}$ (proposition 5.29), so that $\frac{\partial^2}{\partial\psi_i \partial\psi_j} LOO(\boldsymbol{\psi}^{(0)}) = \frac{1}{n}\boldsymbol{y}^t \mathbf{P}_{i,j} \boldsymbol{y}$, with $\frac{1}{n} Tr(\mathbf{P}_{i,j}\mathbf{K}) \to (\boldsymbol{\Sigma}_{CV,2})_{i,j}$ P_X-almost surely. Hence, using proposition 5.35, $\frac{\partial^2}{\partial \boldsymbol{\psi}^2} L(\boldsymbol{\psi}^{(0)})$ converges to $\boldsymbol{\Sigma}_{CV,2}$ in the mean square sense (on the product space).

Finally, $\frac{\partial^3}{\partial\psi_i \partial\psi_j \partial\psi_k} LOO(\tilde{\boldsymbol{\psi}})$ can be written as $\frac{1}{n}\left(\boldsymbol{z}^t \mathbf{N}^{i,j,k}_{\tilde{\boldsymbol{\psi}}} \boldsymbol{z}\right)$, where the $\mathbf{N}^{i,j,k}_{\tilde{\boldsymbol{\psi}}}$ are sums of matrices of $\mathcal{M}_{\tilde{\boldsymbol{\psi}}}$ (proposition 5.29) and $\boldsymbol{z}$ depending on $\boldsymbol{X}$ and Y with $\mathcal{L}(\boldsymbol{z}|\boldsymbol{X}) = \mathcal{N}(0, \mathbf{I}_n)$. The singular values of $\mathbf{N}^{i,j,k}_{\tilde{\boldsymbol{\psi}}}$ are bounded uniformly in $\tilde{\boldsymbol{\psi}}$, n and $\boldsymbol{x}$ and so $\sup_{i,j,k,\tilde{\boldsymbol{\psi}}}\left(\frac{\partial^3}{\partial\psi_i \partial\psi_j \partial\psi_k} LOO(\tilde{\boldsymbol{\psi}})\right)$ is bounded by $b\frac{1}{n}\boldsymbol{z}^t\boldsymbol{z}$, $b < +\infty$, and is hence bounded in probability. We apply proposition 5.38 to conclude. □

Positivity of the Hessians for ML and CV

We conclude the proofs for subsection 5.7.1 by showing that the asymptotic Hessian matrices of propositions 5.8 and 5.13 for ML and CV are positive matrices.

Proposition 5.10 addresses ML.

Proof of proposition 5.10. We firstly prove the proposition in the case $p = 1$, when Σ_{ML} is a scalar. We then show how to generalize the proposition to the case $p > 1$.

For $p = 1$ we have seen that $\frac{1}{n} Tr\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi} \mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi}\right) \to_{P_X} \Sigma_{ML}$. Then

$$\begin{aligned}
& \frac{1}{n} Tr\left(\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi} \mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi}\right) \\
= & \frac{1}{n} Tr\left(\mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi} \mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}} \mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi} \mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}}\right) \\
= & \left|\left|\mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi} \mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}}\right|\right|_2^2 \\
\geq & \inf_{i,n,x} \phi_i\left(\mathbf{K}^{-\frac{1}{2}}_{\boldsymbol{\psi}^{(0)}}\right)^4 \left|\left|\frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial\psi}\right|\right|_2^2.
\end{aligned}$$

By lemma 5.28, there exists $a > 0$ so that $\inf_{i,n,x} \phi_i \left(\mathbf{K}_{\boldsymbol{\psi}^{(0)}}^{-\frac{1}{2}} \right)^4 \geq a$. We then show, similarly to the proof of proposition 5.7, that the limit of $\left|\left| \frac{\partial \mathbf{K}_{\boldsymbol{\psi}^{(0)}}}{\partial \psi} \right|\right|_2^2$ exists and is positive.

We now address the case $p > 1$. Let $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^p$, $\boldsymbol{\lambda}$ different from zero. We define the model $\{K_\delta, \delta \in [\delta_{inf}, \delta_{sup}]\}$, with $\delta_{inf} < 0 < \delta_{sup}$ by $K_\delta = K_{\left(\boldsymbol{\psi}^{(0)}\right)_1 + \delta\lambda_1, ..., \left(\boldsymbol{\psi}^{(0)}\right)_p + \delta\lambda_p}$. Then $K_{\delta=0} = K_{\boldsymbol{\psi}^{(0)}}$. We have

$$\frac{\partial}{\partial \delta} K_{\delta=0}(\boldsymbol{t}) = \sum_{k=1}^{p} \lambda_k \frac{\partial}{\partial \psi_k} K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t}),$$

so the model $\{K_\delta, \delta \in [\delta_{inf}, \delta_{sup}]\}$ verifies the hypotheses of proposition 5.10 for $p = 1$. Hence, the $\mathbb{P}$-mean square limit of $\frac{\partial^2}{\partial \delta^2} L(\delta = 0)$ is positive. We conclude with $\frac{\partial^2}{\partial \delta^2} L(\delta = 0) = \boldsymbol{\lambda}^t \left(\frac{\partial^2}{\partial \boldsymbol{\psi}^2} L(\boldsymbol{\psi}^{(0)}) \right) \boldsymbol{\lambda}$. □

Proposition 5.14 now addresses the positivity of the Hessian for CV.

Proof of proposition 5.14. We show the proposition in the case $p = 1$, the generalization to the case $p > 1$ being the same as in proposition 5.10.

Similarly to the proof of proposition 5.7, we show that

$$\left| \frac{\partial^2}{\partial \psi^2} LOO(\psi_0) - \mathbb{E}\left(\frac{\partial^2}{\partial \psi^2} LOO(\psi_0) \middle| \boldsymbol{X} \right) \right| \to_p 0.$$

We will then show that there exists $C > 0$ so that P_X-a.s.,

$$\mathbb{E}\left(\frac{\partial^2}{\partial \psi^2} LOO(\psi_0) \middle| \boldsymbol{X} \right) \geq C \left|\left| \frac{\partial \mathbf{K}_\psi}{\partial \psi} \right|\right|_2^2. \tag{5.32}$$

The proof of the proposition will hence be carried out similarly as in the proof of proposition 5.7.

$\frac{\partial^2}{\partial \psi^2} LOO(\psi_0)$ can be written as $\boldsymbol{z}^t \mathbf{M} \boldsymbol{z}$ with $\boldsymbol{z}$ depending on $\boldsymbol{X}$ and Y and $\mathcal{L}(\boldsymbol{z}|\boldsymbol{X}) = \mathcal{N}(0, \mathbf{I}_n)$, and $\mathbf{M}$ a sum of matrices of $\mathcal{M}_{\psi_0}$ (proposition 5.29). Hence, using proposition 5.29, uniformly in n, $\sup_\psi \left| \frac{\partial^2}{\partial \psi^2} LOO(\psi) \right| \leq a \frac{1}{n} \boldsymbol{z}^t \boldsymbol{z}$ with $a < +\infty$. Hence, for fixed n, we can exchange derivatives and means conditionally to $\boldsymbol{X}$ and so

$$\mathbb{E}\left(\frac{\partial^2}{\partial \psi^2} LOO(\psi_0) | \boldsymbol{X} \right) = \frac{\partial^2}{\partial \psi^2} \mathbb{E}\left(LOO(\psi_0) | \boldsymbol{X} \right).$$

Then, with $\boldsymbol{k}_{i,\psi}$, $\mathbf{K}_{-i,\psi}$ and $\boldsymbol{y}_{-i}$ the notation of the proof of proposition 5.11,

$$\begin{aligned}
&\mathbb{E}\left(LOO(\psi) | \boldsymbol{X} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[1 - \boldsymbol{k}_{i,\psi_0}^t \mathbf{K}_{-i,\psi_0}^{-1} \boldsymbol{k}_{i,\psi_0} + \mathbb{E}\left\{ \left(\boldsymbol{k}_{i,\psi_0}^t \mathbf{K}_{-i,\psi_0}^{-1} \boldsymbol{y}_{-i} - \boldsymbol{k}_{i,\psi}^t \mathbf{K}_{-i,\psi}^{-1} \boldsymbol{y}_{-i} \right)^2 | \boldsymbol{X} \right\} \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(1 - \boldsymbol{k}_{i,\psi_0}^t \mathbf{K}_{-i,\psi_0}^{-1} \boldsymbol{k}_{i,\psi_0} \right) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{k}_{i,\psi}^t \mathbf{K}_{-i,\psi}^{-1} - \boldsymbol{k}_{i,\psi_0}^t \mathbf{K}_{-i,\psi_0}^{-1} \right) \mathbf{K}_{-i,\psi_0} \left(\mathbf{K}_{-i,\psi}^{-1} \boldsymbol{k}_{i,\psi} - \mathbf{K}_{-i,\psi_0}^{-1} \boldsymbol{k}_{i,\psi_0} \right).
\end{aligned}$$

By differentiating twice with respect to ψ and taking the value at ψ_0 we obtain

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2}{\partial\psi^2}LOO(\psi_0)|\boldsymbol{X}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\psi}\left(\mathbf{K}_{-i,\psi_0}^{-1}\boldsymbol{k}_{i,\psi_0}^{t}\right)\right\}^{t}\mathbf{K}_{-i,\psi_0}\left\{\frac{\partial}{\partial\psi}\left(\mathbf{K}_{-i,\psi_0}^{-1}\boldsymbol{k}_{i,\psi_0}^{t}\right)\right\}\\ &\geq A\frac{1}{n}\sum_{i=1}^{n}\left|\left\{\frac{\partial}{\partial\psi}\left(\mathbf{K}_{-i,\psi_0}^{-1}\boldsymbol{k}_{i,\psi_0}^{t}\right)\right\}\right|^2,\end{aligned}$$

with $A=\inf_{n,i,x}\phi_i^2\left(\mathbf{K}_{-i,\psi_0}\right)$, $A>0$. Then, using the virtual CV formulas of proposition 2.35,

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2}{\partial\psi^2}LOO(\psi_0)|\boldsymbol{X}\right) &\geq A\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{j\neq i}\left[\frac{\partial}{\partial\psi}\left\{\frac{\left(\mathbf{K}_{\psi_0}^{-1}\right)_{i,j}}{\left(\mathbf{K}_{\psi_0}^{-1}\right)_{i,i}}\right\}\right]^2\\ &= A\left|\left|\tfrac{\partial}{\partial\psi}\left\{Diag\left(\mathbf{K}_{\psi_0}^{-1}\right)^{-1}\mathbf{K}_{\psi_0}^{-1}\right\}\right|\right|_2^2.\end{aligned} \tag{5.33}$$

Now,

$$\begin{aligned}&\frac{\partial}{\partial\psi}\left\{Diag\left(\mathbf{K}_{\psi_0}^{-1}\right)^{-1}\mathbf{K}_{\psi_0}^{-1}\right\}\\ &= Diag\left(\mathbf{K}_{\psi_0}^{-1}\right)^{-1}Diag\left(\mathbf{K}_{\psi_0}^{-1}\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\mathbf{K}_{\psi_0}^{-1}\right)Diag\left(\mathbf{K}_{\psi_0}^{-1}\right)^{-1}\mathbf{K}_{\psi_0}^{-1}\\ &- Diag\left(\mathbf{K}_{\psi_0}^{-1}\right)^{-1}\left(\mathbf{K}_{\psi_0}^{-1}\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\mathbf{K}_{\psi_0}^{-1}\right).\end{aligned} \tag{5.34}$$

Hence, from (5.33) and (5.34), and with $B=\inf_{i,n,x}\phi_i\left(\mathbf{K}_{\psi_0}^{-1}\right)$, $B>0$,

$$\begin{aligned}&\mathbb{E}\left(\frac{\partial^2}{\partial\psi^2}LOO(\psi_0)|\boldsymbol{X}\right)\\ &\geq A^2B\left|\left|Diag\left(\mathbf{K}_{\psi_0}^{-1}\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\mathbf{K}_{\psi_0}^{-1}\right)Diag\left(\mathbf{K}_{\psi_0}^{-1}\right)^{-1}-\mathbf{K}_{\psi_0}^{-1}\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\right|\right|_2^2\\ &\geq A^2B\inf_{\lambda_1,\ldots,\lambda_n}\left|\left|\mathbf{D}_{\boldsymbol{\lambda}}-\mathbf{K}_{\psi_0}^{-1}\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\right|\right|_2^2\\ &\geq A^2B^2\inf_{\lambda_1,\ldots,\lambda_n}\left|\left|\mathbf{K}_{\psi_0}\mathbf{D}_{\boldsymbol{\lambda}}-\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\right|\right|_2^2.\end{aligned}$$

Then, as $K_\psi(0)=1$ for all ψ, and hence $\frac{\partial}{\partial\psi}K_{\psi_0}(0)=0$,

$$\begin{aligned}&\mathbb{E}\left(\frac{\partial^2}{\partial\psi^2}LOO(\psi_0)|\boldsymbol{X}\right)\\ &\geq A^2B^2\inf_{\lambda_1,\ldots,\lambda_n}\frac{1}{n}\sum_{i=1}^{n}\left[\lambda_i^2+\sum_{j\neq i}\left\{\lambda_i\left(\mathbf{K}_{\psi_0}\right)_{i,j}-\left(\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\right)_{i,j}\right\}^2\right]\\ &= A^2B^2\frac{1}{n}\sum_{i=1}^{n}\inf_{\lambda}\left[\lambda^2+\sum_{j\neq i}\left\{\lambda\left(\mathbf{K}_{\psi_0}\right)_{i,j}-\left(\frac{\partial\mathbf{K}_{\psi_0}}{\partial\psi}\right)_{i,j}\right\}^2\right].\end{aligned}$$

We then show, similarly to lemma 5.34, that

$$\lambda^2+\sum_{i=1}^{n}\left(a_i-\lambda b_i\right)^2\geq\frac{\sum_{i=1}^{n}a_i^2}{1+\sum_{i=1}^{n}b_i^2}. \tag{5.35}$$

Hence, with $C \in [1, +\infty)$, by using (5.3) and lemma 5.24,

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2}{\partial\psi^2}LOO(\psi_0)|\boldsymbol{X}\right) &\geq \frac{A^2B^2}{C}\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\left\{\left(\frac{\partial \mathbf{K}_{\psi_0}}{\partial\psi}\right)_{i,j}\right\}^2 \\ &= \frac{A^2B^2}{C}\left|\left|\frac{\partial \mathbf{K}_{\psi_0}}{\partial\psi}\right|\right|_2^2 \quad \text{because } \tfrac{\partial}{\partial\psi}K_{\psi_0(0)} = 0.\end{aligned}$$

We then showed (5.32), which concludes the proof in the case $p = 1$. □

Proof for the second derivatives of the asymptotic covariance matrices

Proof of proposition 5.16. It is enough to show the proposition for $\epsilon \in [0, \alpha]$ for all $\alpha < \frac{1}{2}$. We use the following lemma.

Lemma 5.40. *Let f_n be a sequence of C^2 functions on a segment of $\mathbb{R}$. We assume $f_n \to_{unif} f$, $f_n' \to_{unif} g$, $f_n'' \to_{unif} h$. Then, f is C^2, $f' = g$, and $f'' = h$.*

We denote $f_n(\epsilon) = \frac{1}{n}\mathbb{E}\left\{Tr\left(\mathbf{M}^{(i,j)}\right)\right\}$ where $(\mathbf{M}^{(i,j)})_{n\in\mathbb{N}^*}$ is a random matrix sequence defined on $(\Omega_X, \mathcal{F}_X, P_X)$ which belongs to $\mathcal{M}_\theta$ (proposition 5.29). We showed in proposition 5.31 that f_n converges simply to $\boldsymbol{\Sigma}_{i,j}$ on $[0, \alpha]$. We firstly use the dominated convergence theorem to show that f_n is C^2 and that f_n' and f_n'' are of the form

$$\mathbb{E}\left\{\frac{1}{n}Tr\left(\mathbf{N}^{(i,j)}\right)\right\}, \tag{5.36}$$

with $\mathbf{N}^{(i,j)}$ a sum of random matrix sequences of $\tilde{\mathcal{M}}_{\theta_0}$. $\tilde{\mathcal{M}}_{\theta_0}$ is similar to $\mathcal{M}_{\theta_0}$ (proposition 5.29), with the addition of the derivative matrices with respect to ϵ. We can then, using (5.9), adapt proposition 5.31 to show that f_n' and f_n'' converge simply to some functions g and h on $[0, \alpha]$.

Finally, adapting proposition 5.29, the singular values of $\mathbf{N}^{(i,j)}$ are bounded uniformly in x and n. Hence, using $Tr(\mathbf{A}) \leq n||\mathbf{A}||$, for a symmetric matrix $\mathbf{A}$, the derivatives of f_n, f_n' and f_n'' are bounded uniformly in n, so that the simple convergence implies the uniform convergence. The conditions of lemma 5.40 are hence fulfilled.

□

5.7.2 Proofs for subsection 5.3.2

We denote $\partial_\psi\mathbf{K} = \frac{\partial}{\partial\psi}\mathbf{K}$, $\partial_\epsilon\mathbf{K} = \frac{\partial}{\partial\epsilon}\mathbf{K}$, $\partial_{\epsilon,\psi}\mathbf{K} = \frac{\partial}{\partial\epsilon}\frac{\partial}{\partial\psi}\mathbf{K}$, $\partial_{\epsilon,\epsilon}\mathbf{K} = \frac{\partial^2}{\partial\epsilon^2}\mathbf{K}$ and $\partial_{\epsilon,\epsilon,\psi}\mathbf{K} = \frac{\partial^2}{\partial\epsilon^2}\frac{\partial}{\partial\psi}\mathbf{K}$.

These matrices have some sort of the Toeplitz structure of (5.10). For instance, $\partial_\psi\mathbf{K}$ is Toeplitz because

$$(\partial_\psi\mathbf{K})_{i,j} = \frac{\partial}{\partial\psi}K_{\psi_0}(i-j)$$

and $\partial_{\epsilon,\psi}\mathbf{K}$ is the element-wise product of a Toeplitz matrix with a zero-mean matrix because

$$(\partial_{\epsilon,\psi}K)_{i,j} = (X_i - X_j)\frac{\partial}{\partial t}\frac{\partial}{\partial\psi}K_{\psi_0}(i-j).$$

Thus, we will make use of the following classical results on the convergence of traces, products and inverses of Toeplitz matrices.

Proposition 5.41. *Let $k \in \mathbb{N}^*$ and $f_1, ..., f_k$ be C^∞ 2π-periodic complex functions on $[-\pi, \pi]$ so that $f(-t) = \overline{f(t)}$, where $\overline{f(t)}$ is the conjugate of $f(t)$.*

We define their associated Fourier transform sequences by the unique sequences $s^{(1)}, ..., s^{(k)}$ on $\mathbb{R}^{\mathbb{Z}}$ so that $f_i(t) = \sum_{a \in \mathbb{Z}} s_a^{(i)} e^{\mathrm{i}at}$ for all $t \in [-\pi, \pi]$.

Define the sequences of Toeplitz matrices $\mathbf{T}(f_1), ..., \mathbf{T}(f_k)$, by, at step n, $\mathbf{T}(f_i)$ is defined by

$$(\mathbf{T}(f_i))_{k,l} = s^{(i)}_{k-l}.$$

Let $I_1, ..., I_k \in \{-1, 1\}$, and assume that for $I_j = -1$, f_j is positive-valued.

Then, $\frac{1}{n} Tr(\mathbf{T}(f_1)^{I_1} ... \mathbf{T}(f_k)^{I_k})$ converges to $M(f_1^{I_1}, ..., f_k^{I_k})$, with $M(f)$ the mean value of f on $[-\pi, \pi]$.

Proof of proposition 5.41. The proposition naturally follows for the results given in [Gra01], where the proofs are pedagogic. □

As an example of the utilization of proposition 5.41, the trace

$$\frac{1}{n} Tr((\partial_\psi \mathbf{K}) \mathbf{K}^{-1})$$

converges, with the notations of subsection 5.3.2, to $M(\frac{f_\psi}{f})$.

Proof of proposition 5.18

Proof of proposition 5.18. We only give the proof of the expression of $\frac{\partial^2}{\partial \epsilon^2} \Sigma_{ML} \Big|_{\epsilon=0}$, since the proofs of the expressions of Σ_{ML}, $\Sigma_{CV,1}$ and $\Sigma_{CV,2}$ are simpler and essentially follow from proposition 5.41.

Using proposition 5.42 below,

$$\begin{aligned}
& \frac{1}{n} \left\{ \frac{\partial^2}{\partial \epsilon^2} Tr \left(\mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \right) \right\} \\
= \; & 2 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_\epsilon \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\epsilon \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \right) \\
& - 4 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_{\epsilon,\psi} \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\epsilon \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \right) \\
& + 4 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\epsilon \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\epsilon \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \right) \\
& - 2 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \, \mathbf{K}^{-1} \, \partial_{\epsilon,\epsilon} \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \right) \\
& + 2 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_{\epsilon,\psi} \mathbf{K} \, \mathbf{K}^{-1} \, \partial_{\epsilon,\psi} \mathbf{K} \right) \\
& - 4 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \, \mathbf{K}^{-1} \, \partial_\epsilon \mathbf{K} \, \mathbf{K}^{-1} \, \partial_{\epsilon,\psi} \mathbf{K} \right) \\
& + 2 \frac{1}{n} Tr \left(\mathbf{K}^{-1} \, \partial_\psi \mathbf{K} \, \mathbf{K}^{-1} \, \partial_{\epsilon,\epsilon,\psi} \mathbf{K} \right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\frac{1}{n}\left\{\frac{\partial^2}{\partial\epsilon^2}Tr\left(\mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\right)\right\} \qquad (5.37)\\
= \ & 2\frac{1}{n}Tr\left(\partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right)\\
&-4\frac{1}{n}Tr\left(\partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right)\\
&+4\frac{1}{n}Tr\left(\partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right)\\
&-2\frac{1}{n}Tr\left(\partial_{\epsilon,\epsilon}\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right)\\
&+2\frac{1}{n}Tr\left(\partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{K}^{-1}\right)\\
&-4\frac{1}{n}Tr\left(\partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right)\\
&+2\frac{1}{n}Tr\left(\partial_{\epsilon,\epsilon,\psi}\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right).
\end{aligned}$$

Using proposition 5.41, we have $\mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1} = \mathbf{T}\left(f\right)^{-1}\mathbf{T}\left(f_\psi\right)\mathbf{T}\left(f\right)^{-1} \sim_{n\to\infty} \mathbf{T}\left(\frac{f_\psi}{f^2}\right)$ because f and ${}_\psi f$ are C^∞ and f is positive. Hence, as the eigenvalues of $\partial_\epsilon\mathbf{K}$ are uniformly bounded, we obtain, using proposition 5.41 and (5.19),

$$\partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1} \sim_{n\to\infty}\ \partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\ \partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right), \qquad (5.38)$$

and hence

$$\begin{aligned}
&\frac{1}{n}Tr\left(\partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\right) \qquad (5.39)\\
&=\frac{1}{n}Tr\left\{\partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\ \partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\}+o\left(1\right).
\end{aligned}$$

(5.38) is uniform in $\boldsymbol{x}=(x_1,...,x_n)\in[-1,1]^n$ so that (5.39) is uniform in $\boldsymbol{x}=(x_1,...,x_n)\in[-1,1]^n$. Applying the method above for all the terms of (5.37), we obtain

$$\begin{aligned}
&\frac{1}{n}\left\{\frac{\partial^2}{\partial\epsilon^2}Tr\left(\mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi\mathbf{K}\right)\right\}+o\left(1\right)\\
= \ & 2\frac{1}{n}Tr\left\{\partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\ \partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\}-4\frac{1}{n}Tr\left\{\partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{T}\left(\frac{1}{f}\right)\ \partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\}\\
&+4\frac{1}{n}Tr\left\{\partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{1}{f}\right)\ \partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi^2}{f^3}\right)\right\}-2\frac{1}{n}Tr\left\{\partial_{\epsilon,\epsilon}\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi^2}{f^3}\right)\right\}\\
&+2\frac{1}{n}Tr\left\{\partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{T}\left(\frac{1}{f}\right)\ \partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{T}\left(\frac{1}{f}\right)\right\}\\
&-4\frac{1}{n}Tr\left\{\partial_\epsilon\mathbf{K}\ \mathbf{T}\left(\frac{1}{f}\right)\ \partial_{\epsilon,\psi}\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\}+2\frac{1}{n}Tr\left\{\partial_{\epsilon,\epsilon,\psi}\mathbf{K}\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\}+o\left(1\right).
\end{aligned}$$

For a matrix $\mathbf{A}$, we define $\mathbf{A}_x$ by $\left(\mathbf{A}_x\right)_{i,j}=\mathbf{A}_{i,j}\left(X_i-X_j\right)$ and $A_{x,x}$ by $\left(\mathbf{A}_{x,x}\right)_{i,j}=A_{i,j}\left(X_i-X_j\right)^2$, where the X_i are the random perturbations.

We then have, since $\epsilon = 0$,

$$\begin{aligned}
\mathbf{K} &= \mathbf{T}(f), \\
\partial_\psi \mathbf{K} &= \mathbf{T}(f_\psi), \\
\partial_\epsilon \mathbf{K} &= \mathbf{T}_x(\mathrm{i}\ f_t), \\
\partial_{\epsilon,\psi} \mathbf{K} &= \mathbf{T}_x(\mathrm{i}\ f_{t,\psi}), \\
\partial_{\epsilon,\epsilon} \mathbf{K} &= \mathbf{T}_{x,x}(\ f_{t,t})
\end{aligned}$$

and

$$\partial_{\epsilon,\epsilon,\psi} \mathbf{K} = \mathbf{T}_{x,x}(\ f_{t,t,\psi}).$$

With these notations:

$$\begin{aligned}
& \frac{1}{n}\left\{\frac{\partial^2}{\partial\epsilon^2} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right)\right\} \\
= \ & 2\frac{1}{n} Tr\left\{\mathbf{T}_x(f_t)\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\ \mathbf{T}_x(f_t)\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\} \\
& -4\frac{1}{n} Tr\left\{\mathbf{T}_x(f_{t,\psi})\ \mathbf{T}\left(\frac{1}{f}\right)\ \mathbf{T}_x(f_t)\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\} \\
& +4\frac{1}{n} Tr\left\{\mathbf{T}_x(f_t)\ \mathbf{T}\left(\frac{1}{f}\right)\ \mathbf{T}_x(f_t)\ \mathbf{T}\left(\frac{f_\psi^2}{f^3}\right)\right\} \\
& -2\frac{1}{n} Tr\left\{\mathbf{T}_{x,x}(f_{t,t})\ \mathbf{T}\left(\frac{f_\psi^2}{f^3}\right)\right\} \\
& +2\frac{1}{n} Tr\left\{\mathbf{T}_x(f_{t,\psi})\ \mathbf{T}\left(\frac{1}{f}\right)\ \mathbf{T}_x(f_{t,\psi})\ \mathbf{T}\left(\frac{1}{f}\right)\right\} \\
& -4\frac{1}{n} Tr\left\{\mathbf{T}_x(f_t)\ \mathbf{T}\left(\frac{1}{f}\right)\ \mathbf{T}_x(f_{t,\psi})\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\} \\
& +2\frac{1}{n} Tr\left\{\mathbf{T}_{x,x}(f_{t,t,\psi})\ \mathbf{T}\left(\frac{f_\psi}{f^2}\right)\right\} + o(1).
\end{aligned}$$

Hence, using propositions 5.43 and 5.45 below, we obtain the following, where the mean value $\mathbb{E}[.]$ is with respect to the perturbation vector $\boldsymbol{X}$, and where we also use the notation $\mathbb{E}[.]$ when $\epsilon = 0$ to unify the cases $\epsilon = 0$ and $\epsilon \neq 0$.

$$
\begin{aligned}
& \lim_{n\to+\infty} \mathbb{E}\left[\frac{1}{n}\left\{\frac{\partial^2}{\partial\epsilon^2} Tr\left(\mathbf{K}^{-1}\,\partial_\psi \mathbf{K}\,\mathbf{K}^{-1}\,\partial_\psi \mathbf{K}\right)\right\}\right] \\
= \;& 2\left\{\frac{1}{3}M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_t\; f_t\; f_\psi}{f^2}\right) + \frac{1}{3}M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_t\; f_t\; f_\psi}{f^2}\right)\right\} \\
& -4\left\{\frac{1}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_{t,\psi}\; f_t\; f_\psi}{f^2}\right) + \frac{1}{3}M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_{t,\psi}\; f_t}{f}\right)\right\} \\
& +4\left\{\frac{1}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_t\; f_t\; f_\psi^2}{f^3}\right) + \frac{1}{3}M\left(\frac{f_\psi^2}{f^3}\right) M\left(\frac{f_t\; f_t}{f}\right)\right\} \\
& -2\frac{2}{3}M\left(\frac{f_{t,t}\; f_\psi^2}{f^3}\right) \\
& +2\left\{\frac{1}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_{t,\psi}\; f_{t,\psi}}{f}\right) + \frac{1}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_{t,\psi}\; f_{t,\psi}}{f}\right)\right\} \\
& -4\left\{\frac{1}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_t\; f_{t,\psi}\; f_\psi}{f^2}\right) + \frac{1}{3}M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_t\; f_{t,\psi}}{f}\right)\right\} \\
& +2\frac{2}{3}M\left(\frac{f_{t,t,\psi}\; f_\psi}{f^2}\right), \\
= \;& \frac{4}{3}M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_t^2\; f_\psi}{f^2}\right) \\
& -\frac{8}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_{t,\psi}\; f_t\; f_\psi}{f^2}\right) - \frac{8}{3}M\left(\frac{f_\psi}{f^2}\right) M\left(\frac{f_{t,\psi}\; f_t}{f}\right) \\
& +\frac{4}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_t^2\; f_\psi^2}{f^3}\right) + \frac{4}{3}M\left(\frac{f_\psi^2}{f^3}\right) M\left(\frac{f_t^2}{f}\right) \\
& -\frac{4}{3}M\left(\frac{f_{t,t}\; f_\psi^2}{f^3}\right) \\
& +\frac{4}{3}M\left(\frac{1}{f}\right) M\left(\frac{f_{t,\psi}^2}{f}\right) \\
& +\frac{4}{3}M\left(\frac{f_{t,t,\psi}\; f_\psi}{f^2}\right).
\end{aligned}
$$

□

Expression of the second derivative of the Fisher information with respect to ϵ

In proposition 5.42 we give the expression of the second derivative w.r.t ϵ of the (modified) Fisher information $Tr\left(\mathbf{K}^{-1}\,\partial_\psi \mathbf{K}\,\mathbf{K}^{-1}\,\partial_\psi \mathbf{K}\right)$.

Proposition 5.42.

$$\begin{aligned}
&\frac{\partial^2}{\partial\epsilon^2} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
= \;& 2Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -4Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +4Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& -4Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon,\psi} \mathbf{K}\right).
\end{aligned}$$

Proof of proposition 5.42. We use $\frac{\partial}{\partial\epsilon} Tr\left(\mathbf{M}^2\right) = 2Tr\left(\mathbf{M}\ \frac{\partial}{\partial\epsilon}\mathbf{M}\right)$. Then:

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
= \;& 2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\left(-\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K} + \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right)\right) \\
= \;& -2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right).
\end{aligned} \tag{5.40}$$

We use $\frac{\partial}{\partial\epsilon} Tr\left(\mathbf{ABCDEF}\right) = Tr\left(\frac{\partial}{\partial\epsilon}\mathbf{A}\ \mathbf{B}\ \mathbf{C}\ \mathbf{D}\ \mathbf{E}\ \mathbf{F} + ... + \mathbf{A}\ \mathbf{B}\ \mathbf{C}\ \mathbf{D}\ \mathbf{E}\ \frac{\partial}{\partial\epsilon}\mathbf{F}\right)$. Then

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
= \;& -Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
= \;& -Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right)
\end{aligned} \tag{5.41}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
= \;& -Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) + Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& -Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) + Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon,\psi} \mathbf{K}\right).
\end{aligned} \tag{5.42}$$

Using (5.40), (5.41) and (5.42), and using $Tr\left(\mathbf{AB}\right) = Tr\left(\mathbf{BA}\right)$ we obtain

$$\begin{aligned}
& \frac{\partial^2}{\partial\epsilon^2} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
= \ & 2Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +4Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -2\left\{\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right\} \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon,\psi} \mathbf{K}\right),
\end{aligned}$$

so that,

$$\begin{aligned}
& \frac{\partial^2}{\partial\epsilon^2} Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
= \ & 2Tr\left(\mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -4Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +4Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& -2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& -4Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_\epsilon \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\psi} \mathbf{K}\right) \\
& +2Tr\left(\mathbf{K}^{-1}\ \partial_\psi \mathbf{K}\ \mathbf{K}^{-1}\ \partial_{\epsilon,\epsilon,\psi} \mathbf{K}\right).
\end{aligned}$$

□

Two convergence results

We state and prove two convergence results used in the proof of proposition 5.18.

Proposition 5.43. *Let f_1, f_2, f_3 and f_4 be 2π-periodic and C^∞ functions on $[-\pi, \pi]$. Furthermore we suppose that f_1 and f_3 are odd and that f_2 and f_4 are even. Then*

$$\begin{aligned}
& \mathbb{E}\left[\frac{1}{n} Tr\left\{\mathbf{T}_x\left(\mathrm{i}f_1\right)\ \mathbf{T}\left(f_2\right)\ \mathbf{T}_x\left(\mathrm{i}f_3\right)\ \mathbf{T}\left(f_4\right)\right\}\right] \\
\rightarrow_{n\rightarrow\infty} \ & \frac{1}{3} M\left(f_2\right) M\left(f_1 f_3 f_4\right) + \frac{1}{3} M\left(f_4\right) M\left(f_1 f_2 f_3\right).
\end{aligned}$$

Proof of proposition 5.43. We calculate

$$\begin{aligned} Tr\left(\mathbf{ABCD}\right) &= \sum_{i,j=1}^{n} \left(\mathbf{AB}\right)_{i,j} \left(\mathbf{CD}\right)_{j,i} \\ &= \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} A_{i,k} B_{k,i} \right) \left(\sum_{l=1}^{n} C_{j,l} D_{l,i} \right) \\ &= \sum_{i,j,k,l=1}^{n} A_{i,k} B_{k,j} C_{j,l} D_{l,i}. \end{aligned}$$

Then we obtain the following, where, again, the mean value $\mathbb{E}[.]$ is with respect to the perturbation vector $\boldsymbol{X}$.

$$\begin{aligned} & \mathbb{E}\left[Tr\left\{\mathbf{T}_x\left(\mathrm{i}f_1\right)\ \mathbf{T}\left(f_2\right)\ \mathbf{T}_x\left(\mathrm{i}f_3\right)\ \mathbf{T}\left(f_4\right)\right\}\right] \qquad (5.43) \\ = & \mathbb{E}\left\{ \sum_{i,j,k,l=1}^{n} \left(X_i - X_k\right) T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,j} \left(X_j - X_l\right) T\left(\mathrm{i}f_3\right)_{j,l} T\left(f_4\right)_{l,i} \right\} \\ = & \mathbb{E}\left\{ \sum_{i,j,k,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,j} T\left(\mathrm{i}f_3\right)_{j,l} T\left(f_4\right)_{l,i} \left(X_i X_j - X_k X_j - X_i X_l + X_k X_l\right) \right\} \\ = & \frac{1}{3} \sum_{i,k,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,i} T\left(\mathrm{i}f_3\right)_{i,l} T\left(f_4\right)_{l,i} \\ & -\frac{1}{3} \sum_{i,j,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,j} T\left(f_2\right)_{j,j} T\left(\mathrm{i}f_3\right)_{j,l} T\left(f_4\right)_{l,i} \\ & -\frac{1}{3} \sum_{i,j,k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,j} T\left(\mathrm{i}f_3\right)_{i,j} T\left(f_4\right)_{i,i} \\ & +\frac{1}{3} \sum_{i,j,k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,j} T\left(\mathrm{i}f_3\right)_{j,k} T\left(f_4\right)_{k,i}. \end{aligned}$$

Then

$$\begin{aligned} & \frac{1}{n} \sum_{i,k,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,i} T\left(\mathrm{i}f_3\right)_{i,l} T\left(f_4\right)_{l,i} \\ = & \frac{1}{n} \sum_{i,k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,i} \left\{ \sum_{l=1}^{n} T\left(\mathrm{i}f_3\right)_{i,l} T\left(f_4\right)_{l,i} \right\} \\ = & \frac{1}{n} \sum_{i,k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,i} \left(T\left(\mathrm{i}f_3\right) T\left(f_4\right)\right)_{i,i} \\ = & \frac{1}{n} \sum_{i=1}^{n} \left\{T\left(\mathrm{i}f_3\right) T\left(f_4\right)\right\}_{i,i} \left\{ \sum_{k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,i} \right\} \\ = & \frac{1}{n} \sum_{i=1}^{n} \left\{T\left(\mathrm{i}f_3\right) T\left(f_4\right)\right\}_{i,i} \left\{T\left(\mathrm{i}f_1\right) T\left(f_2\right)\right\}_{i,i}. \end{aligned}$$

We use lemma 5.44 for the convergence of this last term.

Lemma 5.44. *For* $||\mathbf{A}'_n - \mathbf{A}_n||_2 \to 0$, $||\mathbf{B}'_n - \mathbf{B}_n||_2 \to 0$, $sup_{i,j,n} \left|\left(\mathbf{A}_n\right)_{i,j}\right| < \infty$ *and* $sup_{i,j,n} \left|\left(\mathbf{B}'_n\right)_{i,j}\right| < \infty$, $\left|\frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{A}'_n\right)_{i,i} \left(\mathbf{B}'_n\right)_{i,i} - \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{A}_n\right)_{i,i} \left(\mathbf{B}_n\right)_{i,i}\right| \to 0$.

Proof of lemma 5.44.

$$
\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{A}'_n\right)_{i,i} \left(\mathbf{B}'_n\right)_{i,i} - \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{A}_n\right)_{i,i} \left(\mathbf{B}_n\right)_{i,i} \right|^2 \\
\leq \quad & \frac{1}{n^2} n \sum_{i=1}^{n} \left\{ \left(\mathbf{A}'_n\right)_{i,i} \left(\mathbf{B}'_n\right)_{i,i} - \left(\mathbf{A}_n\right)_{i,i} \left(\mathbf{B}_n\right)_{i,i} \right\}^2, \quad \text{by Cauchy-Schwartz,} \\
\leq \quad & \frac{1}{n} \sum_{i,j=1}^{n} \left\{ \left(\mathbf{A}'_n\right)_{i,j} \left(\mathbf{B}'_n\right)_{i,j} - \left(\mathbf{A}_n\right)_{i,j} \left(\mathbf{B}_n\right)_{i,j} \right\}^2 \\
\leq \quad & 2 \frac{1}{n} \sum_{i,j=1}^{n} \left\{ \left(\mathbf{A}'_n\right)_{i,j} \left(\mathbf{B}'_n\right)_{i,j} - \left(\mathbf{A}_n\right)_{i,j} \left(\mathbf{B}'_n\right)_{i,j} \right\}^2 \\
& + 2 \frac{1}{n} \sum_{i,j=1}^{n} \left\{ \left(\mathbf{A}_n\right)_{i,j} \left(\mathbf{B}'_n\right)_{i,j} - \left(\mathbf{A}_n\right)_{i,j} \left(\mathbf{B}_n\right)_{i,j} \right\}^2 \\
\leq \quad & 2 \sup_{i,j,n} \left| \left(\mathbf{B}'_n\right)_{i,j} \right| \frac{1}{n} \sum_{i,j=1}^{n} \left\{ \left(\mathbf{A}'_n\right)_{i,j} - \left(\mathbf{A}_n\right)_{i,j} \right\}^2 \\
& + 2 \sup_{i,j,n} \left| \left(\mathbf{A}_n\right)_{i,j} \right| \frac{1}{n} \sum_{i,j=1}^{n} \left\{ \left(\mathbf{B}'_n\right)_{i,j} - \left(\mathbf{B}_n\right)_{i,j} \right\}^2 \\
\leq \quad & 2 \sup_{i,j,n} \left| \left(\mathbf{B}'_n\right)_{i,j} \right| . ||\mathbf{A}'_n - \mathbf{A}_n||_2 \\
& + 2 \sup_{i,j,n} \left| \left(\mathbf{A}_n\right)_{i,j} \right| . ||\mathbf{B}'_n - \mathbf{B}_n||_2.
\end{aligned}
$$

□

We use lemma 5.44 with $\mathbf{A}'_n = \mathbf{T}\left(\mathrm{i}f_1\right)\mathbf{T}\left(f_2\right)$, $\mathbf{A}_n = \mathbf{T}\left(\mathrm{i}f_1 f_2\right)$, $\mathbf{B}'_n = \mathbf{T}\left(\mathrm{i}f_3\right)\mathbf{T}\left(f_4\right)$ and $\mathbf{B}_n = \mathbf{T}\left(\mathrm{i}f_3 f_4\right)$. It is shown in proposition 5.41 that $||\mathbf{A}'_n - \mathbf{A}_n||_2 \to 0$ and $||\mathbf{B}'_n - \mathbf{B}_n||_2 \to 0$. As $\mathrm{i}f_1 f_2$ is C^∞, the coefficients of $\mathbf{T}\left(\mathrm{i}f_1 f_2\right)$ are uniformly bounded. Finally $\left\{\mathbf{T}\left(\mathrm{i}f_1\right)\mathbf{T}\left(f_2\right)\right\}_{i,j} \leq \sup_{i,j,n} \left|T\left(\mathrm{i}f_1\right)_{i,j}\right| \sum_{k\in\mathbb{Z}} \left|T\left(f_2\right)_{k,j}\right|$ which is uniformly bounded because $\mathrm{i}f_1$ and f_2 are C^∞.

Hence

$$
\begin{aligned}
& \frac{1}{n} \sum_{i,k,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,i} T\left(\mathrm{i}f_3\right)_{i,l} T\left(f_4\right)_{l,i} \qquad (5.44) \\
= \quad & \frac{1}{n} \sum_{i=1}^{n} \left\{\mathbf{T}\left(\mathrm{i}f_3\right)\mathbf{T}\left(f_4\right)\right\}_{i,i} \left\{\mathbf{T}\left(\mathrm{i}f_1\right)\mathbf{T}\left(f_2\right)\right\}_{i,i} \\
= \quad & \frac{1}{n} \sum_{i=1}^{n} \left\{\mathbf{T}\left(\mathrm{i}f_3 f_4\right)\right\}_{i,i} \left\{\mathbf{T}\left(\mathrm{i}f_1 f_2\right)\right\}_{i,i} + o\left(1\right) \\
\underset{n\to+\infty}{\to} \quad & M\left(\mathrm{i}f_3 f_4\right) M\left(\mathrm{i}f_1 f_2\right) \\
= \quad & 0, \quad \text{because } f_3 f_4 \text{ is odd.}
\end{aligned}
$$

We show similarly

$$
\frac{1}{n} \sum_{i,j,k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,j} T\left(\mathrm{i}f_3\right)_{j,k} T\left(f_4\right)_{k,i} \to 0. \qquad (5.45)
$$

Then

$$\begin{aligned}
& \frac{1}{n}\sum_{i,j,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,j} T\left(f_2\right)_{j,j} T\left(\mathrm{i}f_3\right)_{j,l} T\left(f_4\right)_{l,i} \qquad (5.46)\\
= \;& M\left(f_2\right)\frac{1}{n}\sum_{i,j,l=1}^{n} T\left(\mathrm{i}f_1\right)_{i,j} T\left(\mathrm{i}f_3\right)_{j,l} T\left(f_4\right)_{l,i}\\
= \;& M\left(f_2\right)\frac{1}{n}\sum_{i,j=1}^{n} T\left(\mathrm{i}f_1\right)_{i,j}\left\{\sum_{l=1}^{n} T\left(\mathrm{i}f_3\right)_{j,l} T\left(f_4\right)_{l,i}\right\}\\
= \;& M\left(f_2\right)\frac{1}{n}\sum_{i,j=1}^{n} T\left(\mathrm{i}f_1\right)_{i,j}\left\{T\left(\mathrm{i}f_3\right) T\left(f_4\right)\right\}_{j,i}\\
= \;& M\left(f_2\right)\frac{1}{n} Tr\left\{\mathbf{T}\left(\mathrm{i}f_1\right)\mathbf{T}\left(\mathrm{i}f_3\right)\mathbf{T}\left(f_4\right)\right\}\\
\to \;& M\left(f_2\right) M\left(\mathrm{i}f_1\mathrm{i}f_3 f_4\right), \quad \text{using proposition 5.41,}\\
= \;& -M\left(f_2\right) M\left(f_1 f_3 f_4\right).
\end{aligned}$$

We show similarly

$$\frac{1}{n}\sum_{i,j,k=1}^{n} T\left(\mathrm{i}f_1\right)_{i,k} T\left(f_2\right)_{k,j} T\left(\mathrm{i}f_3\right)_{i,j} T\left(f_4\right)_{i,i} \to -M\left(f_4\right) M\left(f_1 f_2 f_3\right). \qquad (5.47)$$

We conclude with (5.43), (5.44), (5.45), (5.46) and (5.47) □

Proposition 5.45. *Let f_1 and f_2 be 2π-periodic, C^∞, functions on $[-\pi,\pi]$, with f_1 odd. Then*

$$\mathbb{E}\left[\frac{1}{n} Tr\left\{\mathbf{T}_{x,x}\left(f_1\right)\mathbf{T}\left(f_2\right)\right\}\right] \to \frac{2}{3} M\left(f_1 f_2\right).$$

Proof of proposition 5.45.

$$\begin{aligned}
& \mathbb{E}\left[\frac{1}{n} Tr\left\{\mathbf{T}_{x,x}\left(f_1\right)\mathbf{T}\left(f_2\right)\right\}\right]\\
= \;& \mathbb{E}\left\{\frac{1}{n}\sum_{i,j=1}^{n} T\left(f_1\right)_{i,j}\left(X_i - X_j\right)^2 T\left(f_2\right)_{j,i}\right\}\\
= \;& \frac{1}{n}\frac{2}{3}\sum_{i,j=1}^{n} T\left(f_1\right)_{i,j} T\left(f_2\right)_{j,i}\\
= \;& \frac{2}{3}\frac{1}{n} Tr\left\{\mathbf{T}\left(f_1\right)\mathbf{T}\left(f_2\right)\right\}\\
\to \;& \frac{2}{3} M\left(f_1 f_2\right), \text{using proposition 5.41.}
\end{aligned}$$

□

5.7.3 Proofs for section 5.5

Proof of proposition 5.20. We show, rather similarly to the proof of proposition 5.11,

$$\begin{aligned}
&\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right) \\
&= \mathbb{E}\left[\frac{1}{N^d}\int_{[0,N]^d}(Y(\boldsymbol{x}) - \hat{y}_{\boldsymbol{\psi}}(\boldsymbol{x}))^2 d\boldsymbol{x} - \frac{1}{N^d}\int_{[0,N]^d}(Y(\boldsymbol{x}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{x}))^2 d\boldsymbol{x}\right] \\
&= \mathbb{E}\left[\frac{1}{N^d}\int_{[0,N]^d}(\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{x}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{x}))^2 d\boldsymbol{x}\right] \\
&= \mathbb{E}_{X_1,...,X_{N^d}}\left[\frac{1}{N^d}\int_{[0,N]^d}\mathbb{E}_{Y|X_1,...,X_{N^d}}\left[(\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{x}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{x}))^2\right] d\boldsymbol{x}\right],
\end{aligned}$$

where $\mathbb{E}_{Y|X_1,...,X_{N^d}}$ means that the mean value is calculated conditionally to $X_1, ..., X_{N^d}$, and where the notation Y emphasizes that the only random variable remaining is Y.

Then, with $d(\boldsymbol{x}, X_1, ..., X_{N^d}) = \min_{1\leq i\leq N^d} |\boldsymbol{x} - X_i|_\infty$,

$$\begin{aligned}
&\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right) \\
&\geq \mathbb{E}_{X_1,...,X_{N^d}}\left[\frac{1}{N^d}\int_{[0,N]^d}\mathbf{1}_{d(\boldsymbol{x},X_1,...,X_{N^d})\geq\delta}\mathbb{E}_{Y|X_1,...,X_{N^d}}\left[(\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{x}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{x}))^2\right] d\boldsymbol{x}\right]. \qquad (5.48)
\end{aligned}$$

Let $\mathcal{L}_{N^d+1}$ be the distribution on $(\mathbb{R}^d)^{N^d+1}$ obtained by the following procedure. First, generate independently $X_1, ..., X_{N^d}$, uniformly on $[0, N]^d$, conditionally to the constraint that, for $i \neq j$, $|X_i - X_j|_\infty \geq \delta$. Second, conditionally to $X_1, ..., X_{N^d}$, generate X_{N^d+1} uniformly on $[0, N]^d$, conditionally to the constraint that $d(\boldsymbol{x}, X_1, ..., X_{N^d}) \geq \delta$.

It can be shown that the distribution $\mathcal{L}_{N^d+1}$ can be obtained equivalently by generating independently $X_1, ..., X_{N^d+1}$, uniformly on $[0, N]^d$, conditionally to the constraint that, for $i \neq j$, $|X_i - X_j|_\infty \geq \delta$. Thus, in (5.48), the integrand variable $\boldsymbol{x}$ can be seen as following the same distribution as the $X_1, ..., X_{N^d}$. Indeed, let $P_{X_1,...,X_{N^d}}$ be the probability, given $X_1, ..., X_{N^d}$, that X, following an uniform distribution on $[0, N]^d$, verify $d(X, X_1, ..., X_{N^d}) \geq \delta$. Then, with $X_1, ..., X_{N^d+1} \sim \mathcal{L}_{N^d+1}$, the probability density function of X_{N^d+1} conditionally to $X_1, ..., X_{N^d}$ is $\boldsymbol{x} \to \frac{1}{N^d}\mathbf{1}_{\boldsymbol{x}\in[0,N]^d}\mathbf{1}_{d(\boldsymbol{x},X_1,...,X_{N^d})\geq\delta}\frac{1}{P_{X_1,...,X_{N^d}}}$. Hence we obtain,

$$\begin{aligned}
&\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right) \\
&\geq \mathbb{E}_{X_1,...,X_{N^d+1}\sim\mathcal{L}_{N^d+1}P_{X_1,...,X_{N^d}}}\left[\mathbb{E}_{Y|X_1,...,X_{N^d+1}}(\hat{y}_{\boldsymbol{\psi}}(X_{N^d+1}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(X_{N^d+1}))^2\right].
\end{aligned}$$

Next,

$$P_{X_1,...,X_{N^d+1}} \geq \frac{N^d - (2\delta)^d N^d}{N^d} = 1 - (2\delta)^d.$$

Hence

$$\begin{aligned}
&\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right) \\
&\geq \left(1 - (2\delta)^d\right)\mathbb{E}_{X_1,...,X_{N^d+1}\sim\mathcal{L}_{N^d+1}}\left[\mathbb{E}_{Y|X_1,...,X_{N^d+1}}(\hat{y}_{\boldsymbol{\psi}}(X_{N^d+1}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(X_{N^d+1}))^2\right].
\end{aligned}$$

Now, in the distribution $\mathcal{L}_{X_1,...,X_{N^d+1}}$, the variables $X_1, ..., X_{N^d+1}$ have symmetric roles. Hence

$$\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right)$$
$$\geq \left(1-(2\delta)^d\right)\frac{1}{N^d+1}\sum_{i=1}^{N^d+1}\mathbb{E}\left[(\hat{y}_{i,\boldsymbol{\psi}}-\hat{y}_{i,\boldsymbol{\psi}^{(0)}})^2\right].$$

where $\hat{y}_{i,\boldsymbol{\psi}}$ is the LOO prediction of $Y(X_i)$ according to the covariance hyper-parameter $\boldsymbol{\psi}$ and the observations $Y(X_1),...,Y(X_{i-1}),Y(X_{i+1}),...,Y(X_{N^d+1})$.

Let $\mathbf{K}_{\boldsymbol{\psi}}$ be the $(N^d+1)\times(N^d+1)$ covariance matrix of Y at $X_1,...,X_{N^d+1}$. With $X_1,...,X_{N^d+1}\sim\mathcal{L}_{N^d+1}$, the condition $\min_{i\neq j}|X_i-X_j|_\infty\geq\delta$ ensures that the singular values of the matrices $\mathbf{K}_{\boldsymbol{\psi}}$, $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ and $\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_{i_1}...\partial\psi_{i_k}}$ can be upper-bounded uniformly in n, $\boldsymbol{x}$ and $\boldsymbol{\psi}$ similarly to lemma 5.28. Thus, we can show, in the same way as in the proof of proposition 5.11, that there exists a constant $B<+\infty$ so that

$$\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right)$$
$$\geq B\left(1-(2\delta)^d\right)\frac{1}{N^d+1}\sum_{i\neq j}\mathbb{E}\left[\left(K_{\boldsymbol{\psi}}(X_i-X_j)-K_{\boldsymbol{\psi}^{(0)}}(X_i-X_j)\right)^2\right].$$

Now, with X,X' being two random variables on $[0,N]^d$, following independent uniform distributions, conditionally to the constraint that $|X-X'|_\infty\geq\delta$, we obtain

$$\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right)$$
$$\geq B\left(1-(2\delta)^d\right)N^d\mathbb{E}_{X,X'}\left[\left(K_{\boldsymbol{\psi}}(X-X')-K_{\boldsymbol{\psi}^{(0)}}(X-X')\right)^2\right].$$

The probability density distribution of $X-X'$ at $\boldsymbol{t}\in[-N,N]^d\backslash[-\delta,\delta]^d$ is

$$\frac{1}{N^d}\prod_{i=1}^d\left(1-\frac{|t_i|}{N}\right)\frac{1}{P(|X-X'|_\infty\geq\delta)}\geq\frac{1}{N^d}\prod_{i=1}^d\left(1-\frac{|t_i|}{N}\right).$$

Hence,

$$\mathbb{E}\left(MSE_{\boldsymbol{\psi}} - MSE_{\boldsymbol{\psi}^{(0)}}\right)$$
$$\geq B\left(1-(2\delta)^d\right)\int_{[-N,N]^d\backslash[-\delta,\delta]^d}\left(K_{\boldsymbol{\psi}}(\boldsymbol{t})-K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})\right)^2\prod_{i=1}^d\left(1-\frac{|t_i|}{N}\right)d\boldsymbol{t}$$
$$\to_{N\to+\infty} B\left(1-(2\delta)^d\right)\int_{\mathbb{R}^d\backslash[-\delta,\delta]^d}\left(K_{\boldsymbol{\psi}}(\boldsymbol{t})-K_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})\right)^2 d\boldsymbol{t},$$

using the dominated convergence theorem on $\mathbb{R}^d$. □

Proof of proposition 5.21. Let us first show (5.13). Consider a consistent estimator $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}^{(0)}$. Since $|\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}^{(0)}|=o_p(1)$, it is sufficient to show that $\sup_{1\leq i\leq p,\boldsymbol{\psi}\in\Psi}|\frac{\partial}{\partial\psi_i}E_{\epsilon,\boldsymbol{\psi}}|=O_p(1)$.

Consider a fixed n. Because the trajectory $Y(\boldsymbol{t})$ is almost surely continuous on $[0,N_{1,n}]^d$, because for every $\boldsymbol{\psi}\in\Psi$, $1\leq i\leq p$, $\frac{\partial}{\partial\psi_i}K_{\boldsymbol{\psi}}(\boldsymbol{t})$ is continuous with respect to $\boldsymbol{t}$ and because, from (5.3), $\sup_{\boldsymbol{\psi}\in\Psi,1\leq i\leq p}\left|\frac{\partial}{\partial\psi_i}K_{\boldsymbol{\psi}}(\boldsymbol{t})\right|$ is bounded, we can almost surely exchange integration and derivation w.r.t. ψ_i in the expression of $E_{\epsilon,\boldsymbol{\psi}}$. Thus, we have almost surely

$$\begin{aligned}\frac{\partial}{\partial\psi_i}E_{\epsilon,\boldsymbol{\psi}} &= \frac{1}{N_{1,n}^d}\int_{[0,N_{1,n}]^d}\frac{\partial}{\partial\psi_i}\left((Y(\boldsymbol{t})-\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t}))^2\right)d\boldsymbol{t}\\ &= \frac{2}{N_{1,n}^d}\int_{[0,N_{1,n}]^d}(Y(\boldsymbol{t})-\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t}))\left(-\frac{\partial\boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}+\boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\boldsymbol{y}d\boldsymbol{t},\end{aligned}$$

with $\left(\boldsymbol{k}_{\boldsymbol{\psi}}(\boldsymbol{t})\right)_j = K_{\boldsymbol{\psi}}(\boldsymbol{v}^{(j)} + \epsilon \boldsymbol{x}^{(j)} - \boldsymbol{t})$. Then

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left|\frac{\partial}{\partial\psi_i}E_{\epsilon,\boldsymbol{\psi}}\right|\right)\\
&\leq \frac{2}{N_{1,n}^d}\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{|Y(\boldsymbol{t})-\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t})|\left|\left(\frac{\partial \boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}-\boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\boldsymbol{y}\right|\right\}\right)d\boldsymbol{t}\\
&\leq \frac{2}{N_{1,n}^d}\sqrt{\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{(Y(\boldsymbol{t})-\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t}))^2\right\}\right)}\\
&\quad\times\sqrt{\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{\left(\left(\frac{\partial \boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}-\boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\boldsymbol{y}\right)^2\right\}\right)d\boldsymbol{t}}\\
&\leq \frac{2}{N_{1,n}^d}\sqrt{2\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\{Y(\boldsymbol{t})^2\}\right)+2\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{(\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t}))^2\right\}\right)}\\
&\quad\times\sqrt{\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{\left(\left(\frac{\partial \boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}-\boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\boldsymbol{y}\right)^2\right\}\right)d\boldsymbol{t}}\\
&= \frac{2}{N_{1,n}^d}\sqrt{2K_{\boldsymbol{\psi}^{(0)}}(0)N_{1,n}^d+2\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{(\hat{y}_{\boldsymbol{\psi}}(\boldsymbol{t}))^2\right\}\right)}\\
&\quad\times\sqrt{\int_{[0,N_{1,n}]^d}\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{\left(\left(\frac{\partial \boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}-\boldsymbol{k}_{\boldsymbol{\psi}}^t(\boldsymbol{t})\mathbf{K}_{\boldsymbol{\psi}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\psi}}}{\partial\psi_i}\mathbf{K}_{\boldsymbol{\psi}}^{-1}\right)\boldsymbol{y}\right)^2\right\}\right)d\boldsymbol{t}}. \qquad (5.49)
\end{aligned}
$$

In (5.49), the two supremums can be written

$$\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)^2\right\}\right),$$

with $\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})$ a column vector of size n, not depending on $\boldsymbol{y}$.

Fix $\boldsymbol{t}\in[0,N_{1,n}]^d$. We now use Sobolev embedding theorem (see e.g [Tar07]) on the space Ψ, equipped with the Lebesgue measure. This theorem implies that for $f:\Psi\to\mathbb{R}$, $\sup_{\boldsymbol{\psi}\in\Psi}|f(\boldsymbol{\psi})|\leq C_p\int_\Psi\left(|f(\boldsymbol{\psi})|^p+\sum_{j=1}^p\left|\frac{\partial}{\partial\psi_j}f(\boldsymbol{\psi})\right|^p\right)d\boldsymbol{\psi}$, with C_p a finite constant depending only on p and Ψ. By applying this inequality to the C^1 function of $\boldsymbol{\psi}$, $\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)^2$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)^2\right\}\right)\\
&\leq \mathbb{E}\left(C_p\int_\Psi\sum_{i=1}^p\left|\frac{\partial}{\partial\psi_i}\left(\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)\right)^2\right|^p d\boldsymbol{\psi}\right)+\mathbb{E}\left(C_p\int_\Psi\left|\left(\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)\right)^2\right|^p d\boldsymbol{\psi}\right)\\
&= 2C_p\sum_{i=1}^p\int_\Psi\mathbb{E}\left(\left|\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)\left(\frac{\partial}{\partial\psi_i}\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\right)\boldsymbol{y}\right)\right|^p\right)d\boldsymbol{\psi}+C_p\int_\Psi\mathbb{E}\left(\left\{\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right\}^{2p}\right)d\boldsymbol{\psi}\\
&\leq 2C_p\sum_{i=1}^p\sqrt{\int_\Psi\mathbb{E}\left(\left\{\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)\right\}^{2p}\right)d\boldsymbol{\psi}}\sqrt{\int_\Psi\mathbb{E}\left(\left\{\left(\frac{\partial}{\partial\psi_i}\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\right)\boldsymbol{y}\right)\right\}^{2p}\right)d\boldsymbol{\psi}}\\
&\quad+C_p\int_\Psi\mathbb{E}\left(\left\{\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right\}^{2p}\right)d\boldsymbol{\psi}.
\end{aligned}
$$

There exists a constant C'_p, depending only on p so that, for Z a centered Gaussian variable, $\mathbb{E}(Z^{2p}) = C'_p Var(Z)^p$. Thus, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left\{\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right)^2\right\}\right) \\
&\leq 2C_pC'_p\sum_{i=1}^{p}\sqrt{\int_\Psi\left[\mathbb{E}\left(\left\{\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right\}^2\right)\right]^p d\boldsymbol{\psi}}\sqrt{\int_\Psi\left[\mathbb{E}\left(\left\{\frac{\partial}{\partial\psi_i}\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\right)\boldsymbol{y}\right\}^2\right)\right]^p d\boldsymbol{\psi}} \\
&\quad + C_pC'_p\int_\Psi\left[\mathbb{E}\left(\left\{\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right\}^2\right)\right]^p d\boldsymbol{\psi}. \qquad (5.50)
\end{aligned}$$

Fix $1 \leq i \leq p$ in (5.50). By using (5.3), a slight modification of lemma 5.24 and lemma 5.23, $\sup_{\boldsymbol{\psi}\in\Psi,\boldsymbol{t}\in[0,N_{1,n}]^d}|\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})|^2 \leq A$ and $\sup_{\boldsymbol{\psi}\in\Psi,\boldsymbol{t}\in[0,N_{1,n}]^d}|\frac{\partial}{\partial\psi_i}\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})|^2 \leq A$, independently of n and $\boldsymbol{x}$ and for a constant $A < +\infty$. Thus, in (5.50), $\mathbb{E}\left(\left\{\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\boldsymbol{y}\right\}^2\right) = \mathbb{E}_{\boldsymbol{X}}\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\mathbf{K}_{\boldsymbol{\psi}^{(0)}}\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\right) \leq AB$, with $B = \sup_{n,\boldsymbol{x}}||\mathbf{K}_{\boldsymbol{\psi}^{(0)}}||$. We show in the same way $\mathbb{E}\left(\left\{\frac{\partial}{\partial\psi_i}\left(\boldsymbol{w}_{\boldsymbol{\psi}}(\boldsymbol{t})^t\right)\boldsymbol{y}\right\}^2\right) \leq AB$. Hence, from (5.49) and (5.50), we have shown that, for $1 \leq i \leq p$,

$$\mathbb{E}\left(\sup_{\boldsymbol{\psi}\in\Psi}\left|\frac{\partial}{\partial\psi_i}E_{\epsilon,\boldsymbol{\psi}}\right|\right)$$

is bounded independently of n. Hence $\sup_{\boldsymbol{\psi}\in\Psi,1\leq i\leq p}|\frac{\partial}{\partial\psi_i}E_{\epsilon,\boldsymbol{\psi}}| = O_p(1)$, which proves (5.13).

Let us now prove (5.14).

$$\begin{aligned}
\mathbb{E}\left(E_{\epsilon,\boldsymbol{\psi}^{(0)}}\right) &= \mathbb{E}\left(\frac{1}{(N_{1,n})^d}\int_{[0,N_{1,n}]^d}\left(Y(\boldsymbol{t}) - \hat{y}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})\right)^2 d\boldsymbol{t}\right) \\
&= \frac{1}{(N_{1,n})^d}\int_{[0,N_{1,n}]^d}\mathbb{E}_{\boldsymbol{X}}\left(1 - \boldsymbol{k}^t_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})\mathbf{K}^{-1}_{\boldsymbol{\psi}^{(0)}}\boldsymbol{k}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})\right)d\boldsymbol{t}.
\end{aligned}$$

Now, let $\tilde{\mathbf{K}}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})$ be the covariance matrix of $(Y(\boldsymbol{t}), y_1, ..., y_n)^t$, under covariance function $K_{\boldsymbol{\psi}^{(0)}}$. Then, because of the virtual Leave-One-Out formulas of proposition 2.35,

$$\begin{aligned}
\mathbb{E}\left(E_{\epsilon,\boldsymbol{\psi}^{(0)}}\right) &= \frac{1}{(N_{1,n})^d}\int_{[0,N_{1,n}]^d}\mathbb{E}_{\boldsymbol{X}}\left(\frac{1}{(\tilde{\mathbf{K}}^{-1}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t}))_{1,1}}\right)d\boldsymbol{t} \\
&\geq \frac{1}{N^d_{1,n}}\sum_{i=1}^{n}\int_{\prod_{k=1}^d[\left(\boldsymbol{v}^{(i)}\right)_k+\epsilon+\frac{1}{2}(\frac{1}{2}-\epsilon),\left(\boldsymbol{v}^{(i)}\right)_k+1-\epsilon-\frac{1}{2}(\frac{1}{2}-\epsilon)]}\mathbb{E}_{\boldsymbol{X}}\left(\frac{1}{(\tilde{\mathbf{K}}^{-1}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t}))_{1,1}}\right)d\boldsymbol{t}
\end{aligned}$$

Now, for $\boldsymbol{t} \in \prod_{k=1}^d[\left(\boldsymbol{v}^{(i)}\right)_k + \epsilon + \frac{1}{2}(\frac{1}{2}-\epsilon), \left(\boldsymbol{v}^{(i)}\right)_k + 1 - \epsilon - \frac{1}{2}(\frac{1}{2}-\epsilon)]$, $\inf_{n,1\leq j\leq n,\boldsymbol{x}\in S^n_X}|\boldsymbol{t} - \boldsymbol{v}^{(j)} - \epsilon\boldsymbol{x}^{(j)}|_\infty \geq \frac{1}{2}\left(\frac{1}{2}-\epsilon\right)$. Thus, we can adapt proposition 5.26 to show that the eigenvalues of $\tilde{\mathbf{K}}_{\boldsymbol{\psi}^{(0)}}(\boldsymbol{t})$ are larger than $A > 0$, independently of n, $\boldsymbol{x}$ and $\boldsymbol{t} \in \cup_{1\leq i\leq n}\prod_{k=1}^d[\left(\boldsymbol{v}^{(i)}\right)_k + \epsilon + \frac{1}{2}(\frac{1}{2} - \epsilon), \left(\boldsymbol{v}^{(i)}\right)_k + 1 - \epsilon - \frac{1}{2}(\frac{1}{2}-\epsilon)]$. This yields

$$\mathbb{E}\left(E_{\epsilon,\boldsymbol{\psi}^{(0)}}\right) \geq \frac{A}{N^d_{1,n}}N^d_{1,n}\left(\frac{1}{2}-\epsilon\right)^d,$$

which concludes the proof.

□

Chapter 6

Cross Validation and Maximum Likelihood with misspecified family of covariance functions

This chapter is inspired by the article [Bac13].

6.1 Introduction

In this chapter 6, we aim at further comparing the ML and CV estimations of the covariance hyper-parameters. The conclusion of chapter 5, addressing the increasing-domain asymptotic framework, is that ML is preferable, in the well-specified case, when the true covariance function of the Gaussian process does belong to the parametric set used for estimation. Concerning fixed-domain asymptotics, we have seen in chapter 4, that only microergodic hyper-parameters have an asymptotic influence on the Kriging predictions, and that these hyper-parameters can be consistently estimated by ML (although this is not proved yet for all the classical covariance function families).

Further comparisons have been carried out between ML and CV in the well-specified case. Concerning theoretical results, [Ste90b] showed that for the estimation of a signal-to-noise ratio parameter of a Brownian motion, CV has twice the asymptotic variance of ML. For the case of the estimation of a smoothness and a signal-to-noise ratio parameter, of a covariance function of a Gaussian process, [Ste93] shows that Modified Maximum Likelihood (MML) yields smaller asymptotic variances than Generalized Cross Validation (GCV). It is also shown that the two corresponding prediction errors of the Gaussian process are asymptotically equal, but with a smaller second-order term for MML than for GCV.

Several numerical results are also available, coming either from Monte Carlo studies as in [SWN03, ch.3] or deterministic studies as in [MS04]. These numerical comparisons show an advantage of ML over CV. In both the above numerical studies, the interpolated functions are smooth, and the covariance structures are adapted, being Gaussian in [MS04] and having a free smoothness parameter in [SWN03].

We believe that a framework complementary to the well-specified framework presented above is also relevant in practice. This framework corresponds to the case when a parametric estimation is carried out, within a covariance function set, and when the true underlying covariance function does not belong to this set. We call this the model misspecification case, or the misspecified framework.

In a context of spline approximation methods, situations similar to the misspecified-framework we propose here are studied in [Ste93] and [Kou03]. In [Ste93], in a numerical finite-sample study, GCV is more robust than MML, for selecting the order and smoothness parameters in a spline approximation method, to changes in the predictand function when the spline model remains the same. In [Kou03], an asymptotic comparison between a CV-similar and a Generalized Maximum Likelihood (GML) method is carried-out. The results obtained show that there is much more loss of efficiency in using inappropriately the GML method than the CV-similar method.

In fixed-domain asymptotics, the misspecified framework is all the more relevant when the true covariance function is orthogonal (in the sense of Gaussian measures, see chapter 4) to the covariance functions of the misspecified set. This orthogonality may arise in practice. Indeed, for instance, for two covariance functions of the Matérn class to be equivalent, it is necessary that their smoothness parameters are equal (see chapter 4). Yet, it is common practice, especially for the analysis of computer experiment data, to enforce the smoothness parameter to an arbitrary value (see e.g [MS04]). A misspecified smoothness parameter can have dramatic consequences, as observed in numerical experiments in [Vaz05] chapter 5.3.3. [Ste99], chapter 3 also studies the negative impact of a misspecified smoothness parameter, for a modified version of the Matérn model.

In view of the discussion above, this chapter 6 aims at comparing ML and CV in the misspecified case. We use a two-step approach. In the first step, we consider a parametric family of stationary covariance functions in which only the global variance hyper-parameter is free. In this framework, we carry out a detailed and quantitative finite-sample comparison, using the closed-form expressions for the estimated variances for both the ML and CV methods. For the second step we study the general case in which the global variance hyper-parameter and the correlation hyper-parameters are free and estimated from data. We perform extensive numerical experiments on analytical functions, with various misspecifications, and we compare the Kriging models obtained with the ML and CV estimated hyper-parameters.

Chapter 6 is organized as follows. In section 6.2, we address the case of the estimation of a single variance parameter. In subsection 6.2.1 we detail the statistical framework, we introduce an original quality criterion for a variance estimator, and we give a closed-form formula of this criterion for a large family of estimators. In subsection 6.2.2 we numerically apply the closed-form formulas of subsection 6.2.1 and we study their dependences with respect to model misspecification and number of observation points. We highlight our main result that when the correlation model is misspecified, CV does better compared to ML. Finally in section 6.3 we illustrate this result on the Ishigami analytical function and then generalize it, on the Ishigami and Morris analytical functions, to the case where the correlation hyper-parameters are estimated as well.

6.2 Estimation of a single variance parameter

6.2.1 Theoretical framework

Correlation function error and the Risk criterion

We consider a Gaussian process Y, indexed by a set $\mathcal{D}$. Y is zero-mean, stationary, with unit variance, and its correlation function is denoted by R_1. A Kriging model is built for Y, for which it is assumed that Y is centered and that its covariance function belongs to the set $\mathcal{C}$, with

$$\mathcal{C} = \left\{\sigma^2 R_2, \sigma^2 \in \mathbb{R}^+\right\}, \tag{6.1}$$

where $R_2(\boldsymbol{x})$ is a given stationary correlation function. Throughout this chapter, $\mathbb{E}_i$, Var_i, Cov_i and $\sim_i$, $i \in \{1,2\}$, denote means, variances, covariances and probability distributions taken with respect to the distribution of Y with mean zero, variance one, and the correlation function R_i. We observe Y on the points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$. In this framework, the hyper-parameter σ^2 is estimated from the data $\boldsymbol{y} = (y_1, ..., y_n)^t = (Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))^t$ using an estimator $\hat{\sigma}^2$. This estimation does not affect the Kriging prediction (2.9) of $y_0 = Y(\boldsymbol{x}^{(0)})$, for a new point $\boldsymbol{x}^{(0)}$, which we denote by $\hat{y}(\boldsymbol{x}^{(0)})$:

$$\hat{y}(\boldsymbol{x}^{(0)}) := \mathbb{E}_2(y_0|\boldsymbol{y}) = \boldsymbol{r}_2^t \mathbf{R}_2^{-1} \boldsymbol{y}, \tag{6.2}$$

where $(\boldsymbol{r}_i)_j = R_i(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(0)})$ and $(\mathbf{R}_i)_{j,k} = R_i(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)})$, $i \in \{1,2\}$, $1 \leq j,k \leq n$. The conditional mean square error of this non-optimal prediction is

$$\begin{aligned} \mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right] &= (\hat{y}(\boldsymbol{x}^{(0)}) - \mathbb{E}_1(y_0|\boldsymbol{y}))^2 + Var_1(y_0|\boldsymbol{y}) \\ &= (\boldsymbol{r}_1^t \mathbf{R}_1^{-1}\boldsymbol{y} - \boldsymbol{r}_2^t \mathbf{R}_2^{-1}\boldsymbol{y})^2 + 1 - \boldsymbol{r}_1^t \mathbf{R}_1^{-1}\boldsymbol{r}_1. \end{aligned} \tag{6.3}$$

However, using the covariance family $\mathcal{C}$, we use the classical Kriging predictive variance expression $\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})$ in (2.10), that is

$$\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)}) := \hat{\sigma}^2 Var_2(y_0|\boldsymbol{y}) = \hat{\sigma}^2\left(1 - \boldsymbol{r}_2^t \mathbf{R}_2^{-1}\boldsymbol{r}_2\right). \tag{6.4}$$

As we are interested in the accuracy of the predictive variances obtained from an estimator $\hat{\sigma}^2$, the following notion of Risk can be formulated.

Definition 6.1. *For an estimator $\hat{\sigma}^2$ of σ^2, we call Risk at $\boldsymbol{x}^{(0)}$ and denote by $\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}}$ the quantity*

$$\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} = \mathbb{E}_1\left[\left(\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right] - \hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})\right)^2\right].$$

If $\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}}$ is small, then this means that the predictive variance $\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})$ is a correct prediction of the conditional mean square error (6.3) of the prediction $\hat{y}(\boldsymbol{x}^{(0)})$. Note that when $R_1 = R_2$ the minimizer of the Risk at every $\boldsymbol{x}^{(0)}$ is $\hat{\sigma}^2 = 1$. When $R_1 \neq R_2$, an estimate of σ^2 different from 1 can improve the predictive variance, partly compensating for the correlation function error.

To complete this section, we give the closed-form expression of the Risk of an estimator that can be written as a quadratic form of the observations, which is the case for all classical estimators, including the ML and CV estimators of σ^2 in proposition 3.21 and (3.15).

Proposition 6.2. *Let $\hat{\sigma}^2$ be an estimator of σ^2 of the form $\boldsymbol{y}^t\mathbf{M}\boldsymbol{y}$ with $\mathbf{M}$ an $n \times n$ matrix. Denoting*

$$f(\mathbf{A}, \mathbf{B}) = Tr(\mathbf{A})Tr(\mathbf{B}) + 2Tr(\mathbf{AB}),$$

for $\mathbf{A}$, $\mathbf{B}$ $n \times n$ real matrices,

$$\mathbf{M}_0 = (\mathbf{R}_2^{-1}\boldsymbol{r}_2 - \mathbf{R}_1^{-1}\boldsymbol{r}_1)(\boldsymbol{r}_2^t\mathbf{R}_2^{-1} - \boldsymbol{r}_1^t\mathbf{R}_1^{-1})\mathbf{R}_1,$$

$$\mathbf{M}_1 = \mathbf{M}\mathbf{R}_1,$$

$$c_1 = 1 - \boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{r}_1$$

and

$$c_2 = 1 - \boldsymbol{r}_2^t\mathbf{R}_2^{-1}\boldsymbol{r}_2,$$

we have:

$$\begin{aligned}\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} &= f(\mathbf{M}_0, \mathbf{M}_0) + 2c_1Tr(\mathbf{M}_0) - 2c_2f(\mathbf{M}_0, \mathbf{M}_1) \\ &\quad + c_1^2 - 2c_1c_2Tr(\mathbf{M}_1) + c_2^2 f(\mathbf{M}_1, \mathbf{M}_1).\end{aligned}$$

Proof. Using the definition of the Risk, the expression of $\hat{\sigma}^2$, (6.3) and (6.4), we get:

$$\begin{aligned}\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} &= \mathbb{E}_1\big[(\boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{y} - \boldsymbol{r}_2^t\mathbf{R}_2^{-1}\boldsymbol{y})^2 + 1 - \boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{r}_1 \\ &\quad - \boldsymbol{y}^t\mathbf{M}\boldsymbol{y}(1 - \boldsymbol{r}_2^t\mathbf{R}_2^{-1}\boldsymbol{r}_2)\big]^2 \\ &= \mathbb{E}_1\big[\boldsymbol{y}^t(\mathbf{R}_2^{-1}\boldsymbol{r}_2 - \mathbf{R}_1^{-1}\boldsymbol{r}_1)(\boldsymbol{r}_2^t\mathbf{R}_2^{-1} - \boldsymbol{r}_1^t\mathbf{R}_1^{-1})\boldsymbol{y} \\ &\quad + 1 - \boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{r}_1 - \boldsymbol{y}^t\mathbf{M}\boldsymbol{y}(1 - \boldsymbol{r}_2^t\mathbf{R}_2^{-1}\boldsymbol{r}_2)\big]^2.\end{aligned}$$

Then, writing $\boldsymbol{y} = \mathbf{R}_1^{\frac{1}{2}}\boldsymbol{z}$ with $\boldsymbol{z} \sim_1 \mathcal{N}(0, \mathbf{I}_n)$, we get:

$$\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} = \mathbb{E}_1\left(\boldsymbol{z}^t\tilde{\mathbf{M}}_0\boldsymbol{z} + c_1 - c_2\boldsymbol{z}^t\tilde{\mathbf{M}}_1\boldsymbol{z}\right)^2, \tag{6.5}$$

with

$$\tilde{\mathbf{M}}_0 = \mathbf{R}_1^{\frac{1}{2}}(\mathbf{R}_2^{-1}\boldsymbol{r}_2 - \mathbf{R}_1^{-1}\boldsymbol{r}_1)(\boldsymbol{r}_2^t\mathbf{R}_2^{-1} - \boldsymbol{r}_1^t\mathbf{R}_1^{-1})\mathbf{R}_1^{\frac{1}{2}}$$

and

$$\tilde{\mathbf{M}}_1 = \mathbf{R}_1^{\frac{1}{2}}\mathbf{M}\mathbf{R}_1^{\frac{1}{2}}.$$

To compute this expression, we use the following lemma.

Lemma 6.3. *Let $\boldsymbol{z} \sim_1 \mathcal{N}(0, \mathbf{I}_n)$, and $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ real symmetric matrices. Then:*

$$\mathbb{E}_1(\boldsymbol{z}^t\mathbf{A}\boldsymbol{z}\boldsymbol{z}^t\mathbf{B}\boldsymbol{z}) = f(\mathbf{A}, \mathbf{B}).$$

Proof of lemma 6.3. This lemma corresponds to (5.20), since $\mathbb{E}(\boldsymbol{z}^t\mathbf{A}\boldsymbol{z}) = Tr(\mathbf{A})$. □

Using the lemma and expanding (6.5) yields

$$\begin{aligned}\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} &= f(\tilde{\mathbf{M}}_0, \tilde{\mathbf{M}}_0) + 2c_1Tr(\tilde{\mathbf{M}}_0) - 2c_2f(\tilde{\mathbf{M}}_0, \tilde{\mathbf{M}}_1) \\ &\quad + c_1^2 - 2c_1c_2Tr(\tilde{\mathbf{M}}_1) + c_2^2 f(\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_1).\end{aligned} \tag{6.6}$$

Finally, based on $Tr(\mathbf{AB}) = Tr(\mathbf{BA})$, we can replace $\tilde{\mathbf{M}}_0$ and $\tilde{\mathbf{M}}_1$ by $\mathbf{M}_0$ and $\mathbf{M}_1$ in (6.6), which completes the proof. □

It seems difficult at first sight to conclude from proposition 6.2 whether one estimator is better than another one, given a correlation function error and a set of observation points. Therefore, in subsection 6.2.2, we numerically analyze the Risk for the ML and CV estimators of the variance for several designs of experiments. Before that, we recall the ML and CV estimators of σ^2, and we confirm that ML is more efficient when there is no correlation function error.

The ML and CV estimators of the variance parameter

In the framework of section 6.2, the ML estimator $\hat{\sigma}^2_{ML}$ of σ^2 (see proposition 3.21) is

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \boldsymbol{y}^t \mathbf{R}_2^{-1} \boldsymbol{y}. \tag{6.7}$$

Let us now recall the CV estimator of σ^2. The principle is that, given a value σ^2 specifying the covariance function used among the set $\mathcal{C}$, we can, for $1 \leq i \leq n$, compute $\hat{y}_i := \mathbb{E}_2(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$ and $\sigma^2 \hat{c}_i^2 := \sigma^2 Var_2(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$. The Cross Validation estimate of σ^2 is hence, from (3.15),

$$\hat{\sigma}^2_{LOO} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{c}_i^2}. \tag{6.8}$$

By means of the virtual LOO formulas of proposition 2.35, we obtain the following vector-matrix closed-form expression of (6.8),

$$\hat{\sigma}^2_{LOO} = \frac{1}{n} \boldsymbol{y}^t \mathbf{R}_2^{-1} \left[Diag(\mathbf{R}_2^{-1}) \right]^{-1} \mathbf{R}_2^{-1} \boldsymbol{y}.$$

In chapter 5, we have not addressed the expansion-asymptotic results for $\hat{\sigma}^2_{LOO}$ in the well-specified case. In proposition 6.4, we show that, when $R_1 = R_2$, this estimator is consistent. This is expected, because we have seen in chapter 5 that, under mild conditions, all correlation hyper-parameters are consistently estimated by CV.

Proposition 6.4. *Assume $\mathcal{D} = \mathbb{R}^d$ and that the observation points constitute a sequence $(\boldsymbol{x}^{(i)})_{i \in \mathbb{N}^*}$ verifying, for a constant $\delta > 0$, $|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}| \geq \delta$ for $i \neq j$. Assume $R_2 = R_1$ has a positive continuous Fourier transform and satisfies, for a constant $c < +\infty$ and for all $\boldsymbol{t} \in \mathbb{R}^d$,*

$$|R_2(\boldsymbol{t})| \leq \frac{c}{(1 + |\boldsymbol{t}|)^{d+1}}.$$

Then $\hat{\sigma}^2_{LOO}$ converges in the mean square sense to one as $n \to +\infty$.

Proof. Introducing $\boldsymbol{z} = \mathbf{R}_2^{-\frac{1}{2}} \boldsymbol{y} \sim_2 \mathcal{N}(0, \mathbf{I}_n)$ yields:

$$\hat{\sigma}^2_{LOO} = \frac{1}{n} \boldsymbol{z}^t \mathbf{R}_2^{-\frac{1}{2}} \left[Diag(\mathbf{R}_2^{-1}) \right]^{-1} \mathbf{R}_2^{-\frac{1}{2}} \boldsymbol{z},$$

Then,

$$\begin{aligned} \mathbb{E}_2(\hat{\sigma}^2_{LOO}) &= \frac{1}{n} Tr \left(\mathbf{R}_2^{-\frac{1}{2}} \left[Diag(\mathbf{R}_2^{-1}) \right]^{-1} \mathbf{R}_2^{-\frac{1}{2}} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{(\mathbf{R}_2^{-1})_{i,i}}{(\mathbf{R}_2^{-1})_{i,i}} \\ &= 1. \end{aligned}$$

Furthermore,

$$Var_2(\hat{\sigma}^2_{LOO}) = \frac{2}{n^2} Tr\left(\mathbf{R}_2^{-1}\left[Diag(\mathbf{R}_2^{-1})\right]^{-1}\mathbf{R}_2^{-1}\left[Diag(\mathbf{R}_2^{-1})\right]^{-1}\right). \tag{6.9}$$

Then, with $\lambda_{min}(\mathbf{R})$ and $\lambda_{max}(\mathbf{R})$ the smallest and largest eigenvalues of a symmetric positive matrix $\mathbf{R}$,

$$\begin{aligned}\lambda_{max}(\left[Diag(\mathbf{R}_2^{-1})\right]^{-1}) &= \max_{1\leq i\leq n}\frac{1}{(\mathbf{R}_2^{-1})_{i,i}}\\ &\leq \frac{1}{\lambda_{min}(\mathbf{R}_2^{-1})}\\ &= \lambda_{max}(\mathbf{R}_2).\end{aligned} \tag{6.10}$$

Hence, from (6.9) and (6.10),

$$\frac{2}{n}\left(\frac{\lambda_{min}(\mathbf{R}_2)}{\lambda_{max}(\mathbf{R}_2)}\right)^2 \leq Var_2(\hat{\sigma}^2_{LOO}) \leq \frac{2}{n}\left(\frac{\lambda_{max}(\mathbf{R}_2)}{\lambda_{min}(\mathbf{R}_2)}\right)^2. \tag{6.11}$$

Because there exists a positive minimum distance between two different observation points, it can be shown, similarly to the proof of lemma 5.28, that $0 < \inf_n \lambda_{min}(\mathbf{R}_2) \leq \sup_n \lambda_{max}(\mathbf{R}_2) < +\infty$. This implies the proposition because of (6.11). □

ML is preferable when $R_2 = R_1$

When $R_1 = R_2$, we will show that ML is more efficient than CV. Indeed, first notice that

$$\begin{aligned}\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} &= \mathbb{E}_1\left((1-\boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{r}_1) - \hat{\sigma}^2(1-\boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{r}_1)\right)^2\\ &= (1-\boldsymbol{r}_1^t\mathbf{R}_1^{-1}\boldsymbol{r}_1)^2\mathbb{E}_1((\hat{\sigma}^2-1)^2),\end{aligned} \tag{6.12}$$

so that the Risk of definition 6.1 is proportional to the quadratic error in estimating the true $\sigma^2 = 1$. We calculate $\mathbb{E}_1(\hat{\sigma}^2_{ML}) = \mathbb{E}_1(\hat{\sigma}^2_{LOO}) = 1$, hence both estimators are unbiased.

Concerning their variances, let us first recall the Cramér-Rao bound (see chapter 3) for the estimation of σ^2. As we are in the case $\sigma^2 = 1$, for an unbiased estimator $\hat{\sigma}^2$ of σ^2:

$$Var_1(\hat{\sigma}^2) \geq \mathbb{E}_1^{-1}\left[\left(\frac{\partial}{\partial\sigma^2}\left(\ln l(\boldsymbol{y},\sigma^2)\right)_{\sigma^2=1}\right)^2\right],$$

with, $l(\boldsymbol{y},\sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{\boldsymbol{y}^t\mathbf{R}_1^{-1}\boldsymbol{y}}{2\sigma^2}\right)$, the likelihood of the observations. We then calculate the Cramér-Rao bound:

$$\begin{aligned}\mathbb{E}_1^{-1}\left[\left(\frac{\partial}{\partial\sigma^2}\left(\ln l(\boldsymbol{y},\sigma^2)\right)_{\sigma^2=1}\right)^2\right] &= \mathbb{E}_1^{-1}\left[\left(\frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\ln\sigma^2 - \frac{\boldsymbol{y}^t\mathbf{R}_1^{-1}\boldsymbol{y}}{2\sigma^2}\right)_{\sigma^2=1}\right)^2\right]\\ &= \mathbb{E}_1^{-1}\left[\frac{n^2}{4} + \frac{1}{4}(\boldsymbol{y}^t\mathbf{R}_1^{-1}\boldsymbol{y})^2 - \frac{n}{2}\boldsymbol{y}^t\mathbf{R}_1^{-1}\boldsymbol{y}\right]\\ &= \left(\frac{n^2}{4} + \frac{n^2+2n}{4} - \frac{n^2}{2}\right)^{-1}\\ &= \frac{2}{n},\end{aligned}$$

where we used lemma 6.3 with $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$ to show $\mathbb{E}_1\left((\boldsymbol{y}^t\mathbf{R}_1^{-1}\boldsymbol{y})^2\right) = n^2 + 2n$. Hence, the Cramér-Rao bound of the statistical model $\mathcal{C}$ is $\frac{2}{n}$ when $\sigma^2 = 1$.

Now, on the one hand the variance of the ML estimator is

$$\begin{aligned} Var_1(\hat{\sigma}^2_{ML}) &= Var_1(\frac{1}{n}\boldsymbol{y}^t\mathbf{R}_1^{-1}\boldsymbol{y}) \\ &= \frac{1}{n^2}Var_1(\sum_{i=1}^n z_i^2) \\ &= \frac{2}{n}, \end{aligned}$$

with $\boldsymbol{z} = \mathbf{R}_1^{-\frac{1}{2}}\boldsymbol{y} \sim_1 \mathcal{N}(0, \mathbf{I}_n)$. Thus, the ML estimator reaches the Cramér-Rao bound.

On the other hand:

$$\begin{aligned} Var_1(\hat{\sigma}^2_{LOO}) &= \frac{2}{n^2}Tr(\mathbf{R}_1^{-1}\left[Diag(\mathbf{R}_1^{-1})\right]^{-1}\mathbf{R}_1^{-1}\left[Diag(\mathbf{R}_1^{-1})\right]^{-1}) \\ &= \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{(\mathbf{R}_1^{-1})_{i,j}^2}{(\mathbf{R}_1^{-1})_{i,i}(\mathbf{R}_1^{-1})_{j,j}}. \end{aligned}$$

Hence $Var_1(\hat{\sigma}^2_{LOO}) \geq \frac{2}{n^2}\sum_{i=1}^n \frac{(\mathbf{R}_1^{-1})_{i,i}^2}{(\mathbf{R}_1^{-1})_{i,i}(\mathbf{R}_1^{-1})_{i,i}} = \frac{2}{n}$, the Cramér-Rao bound. Let us also notice that, roughly speaking, when $\mathbf{R}_2^{-1} = \mathbf{R}_1^{-1}$ becomes close to being diagonal, $Var_1(\hat{\sigma}^2_{LOO})$ becomes closer to the Cramér-Rao bound $\frac{2}{n}$.

However $Var_1(\hat{\sigma}^2_{LOO})$ is only upper-bounded by 2 (because $\mathbf{R}_1^{-1}$ is a covariance matrix). Furthermore $Var_1(\hat{\sigma}^2_{LOO})$ can be arbitrarily close to 2. To see this, consider the following statistical model where $\mathbf{R}_1 = \mathbf{R}_2$ are stationary covariance matrices:

$$\mathbf{R}_1 = \mathbf{R}_2 = \frac{n-1+\epsilon}{n-1}\mathbf{I} - \frac{\epsilon}{n-1}\mathbf{J},$$

where $\mathbf{J}$ is the $n \times n$ matrix with all coefficients being 1 and $\epsilon \in [0, 1)$.

Using the formula $(a\mathbf{I} + b\mathbf{J})^{-1} = \frac{1}{a}\mathbf{I} - \frac{b}{a(a+nb)}\mathbf{J}$ (lemma B.3.3 in [SWN03]), we obtain

$$\mathbf{R}_2^{-1} = (n-1)\left(\frac{1}{n-1+\epsilon}\mathbf{I} + \frac{\epsilon}{(n-1+\epsilon)(n-1+\epsilon-n\epsilon)}\mathbf{J}\right).$$

Thus, we can calculate explicitly the variance of the CV estimator,

$$\begin{aligned} Var_1(\hat{\sigma}^2_{LOO}) &= \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{(\mathbf{R}_2^{-1})_{i,j}^2}{(\mathbf{R}_2^{-1})_{i,i}(\mathbf{R}_2^{-1})_{j,j}} \\ &= \frac{2}{n} + \frac{2(n-1)}{n}\frac{\frac{\epsilon^2}{(n-1+\epsilon)^2(n-1+\epsilon-n\epsilon)^2}}{\left(\frac{1}{n-1+\epsilon} + \frac{\epsilon}{(n-1+\epsilon)(n-1+\epsilon-n\epsilon)}\right)^2} \\ &= \frac{2}{n} + \frac{2(n-1)}{n}\frac{\epsilon^2}{(n-1+\epsilon-n\epsilon+\epsilon)^2} \\ &= \frac{2}{n} + \frac{2(n-1)}{n}\frac{\epsilon^2}{(\epsilon+(n-1)(1-\epsilon))^2}. \end{aligned}$$

Hence, for ϵ arbitrarily close to 1, $Var_1(\hat{\sigma}^2_{LOO})$ is arbitrarily close to 2.

As a conclusion, when $R_1 = R_2$, ML is more efficient to estimate the variance parameter. The object of subsection 6.2.2 is to study the case $R_1 \neq R_2$ numerically.

6.2.2 Numerical results

All the numerical experiments are carried out with the numerical software Scilab [GBC+99]. We use the Mersenne Twister pseudo random number generator of M. Matsumoto and T. Nishimura, which is the default pseudo random number generator in Scilab for large-size random simulations.

Criteria for comparison

Pointwise criteria We define two quantitative criteria that will be used to compare the ML and CV assessments of the predictive variance at prediction point $\boldsymbol{x}^{(0)}$.

The first criterion is the Risk on Target Ratio (RTR),

$$RTR(\boldsymbol{x}^{(0)}) = \frac{\sqrt{\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}}}}{\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right]}, \tag{6.13}$$

with $\hat{\sigma}^2$ being either $\hat{\sigma}^2_{ML}$ or $\hat{\sigma}^2_{LOO}$.

From definition 6.1 we obtain

$$RTR(\boldsymbol{x}^{(0)}) = \frac{\sqrt{\mathbb{E}_1\left[\left(\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right] - \hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})\right)^2\right]}}{\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right]}. \tag{6.14}$$

The numerator of (6.14) is the root mean square error in predicting the random quantity $\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right]$ (the target in the RTR acronym) with the predictor $\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})$. Using $\mathbb{E}(\mathbb{E}(X_1|X_2)) = \mathbb{E}(X_1)$ for two random variables X_1 and X_2, the denominator of (6.14) is the mean of the predictand $\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right]$. Hence, the RTR in (6.13) is a relative prediction error, which can be easily interpreted.

We have the following bias-variance decomposition of the Risk,

$$\begin{aligned}\mathcal{R}_{\hat{\sigma}^2,\boldsymbol{x}^{(0)}} &= \left(\underbrace{\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right] - \mathbb{E}_1\left[\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})\right]}_{\text{bias}}\right)^2 \\ &\quad + \underbrace{Var_1\left(\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right] - \hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})\right)}_{\text{variance}}.\end{aligned} \tag{6.15}$$

Hence the second criterion is the Bias on Target Ratio (BTR) and is the relative bias

$$BTR(\boldsymbol{x}^{(0)}) = \frac{\left|\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right] - \mathbb{E}_1\left(\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})\right)\right|}{\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right]}. \tag{6.16}$$

The following equation summarizes the link between RTR and BTR:

$$\left(\underbrace{RTR}_{\text{relative error}}\right)^2 = \left(\underbrace{BTR}_{\text{relative bias}}\right)^2 + \underbrace{\frac{Var_1\left(\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right] - \hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(0)})\right)}{\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right]^2}}_{\text{relative variance}}. \tag{6.17}$$

Case of no correlation function misspecification Let us now study more particularly the RTR and BTR criteria in the case where $R_1 = R_2$. When $R_1 = R_2$, $\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2|\boldsymbol{y}\right]$ does not depend on $\boldsymbol{y}$. Therefore, the RTR and BTR simplify into $RTR(\boldsymbol{x}^{(0)}) = \sqrt{\mathbb{E}_1\left[(\hat{\sigma}^2 - 1)^2\right]}$ and $BTR(\boldsymbol{x}^{(0)}) = |1 - \mathbb{E}_1(\hat{\sigma}^2)|$. Hence, the RTR and BTR are the mean square error and the bias in the estimation of the true variance $\sigma^2 = 1$, and $RTR^2 = BTR^2 + Var_1(\hat{\sigma}^2)$.

Integrated criteria We now define the two integrated versions of RTR and BTR over the prediction space $\mathcal{D}$. Assume $\mathcal{D}$ is equipped with a probability measure μ. Then we define

$$IRTR = \sqrt{\int_{\mathcal{D}} RTR^2(\boldsymbol{x}^{(0)}) d\mu(\boldsymbol{x}^{(0)})} \tag{6.18}$$

and

$$IBTR = \sqrt{\int_{\mathcal{D}} BTR^2(\boldsymbol{x}^{(0)}) d\mu(\boldsymbol{x}^{(0)})}. \tag{6.19}$$

Hence we have the equivalent of (6.17) for IRTR and IBTR:

$$IRTR^2 = IBTR^2 + \int_{\mathcal{D}} \frac{Var_1\left(\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2 | \boldsymbol{y}\right] - \hat{\sigma}^2 \hat{c}^2(\boldsymbol{x}^{(0)})\right)}{\mathbb{E}_1\left[(\hat{y}(\boldsymbol{x}^{(0)}) - y_0)^2\right]^2} d\mu(\boldsymbol{x}^{(0)}). \tag{6.20}$$

Designs of experiments studied

We consider three different kinds of Designs Of Experiments (DOEs) of n observation points on the prediction space $\mathcal{D} = [0,1]^d$.

The first DOE is the Simple Random Sampling (SRS) design and consists of n independent observation points with uniform distributions on $[0,1]^d$. This design may not be optimal from a Kriging prediction point of view, as it is likely to contain relatively large areas without observation points. However, it is a convenient design for the estimation of covariance hyper-parameters because it may contain some points with small spacing. It is noted in [Ste99], chapter 6.9 that such points can dramatically improve the estimation of the covariance hyper-parameters. The conclusion of chapter 5, on the impact of the spatial sampling on estimation, is also an argument in favor of using some observation points with small spacing.

The second DOE is the Latin Hypercube Sampling Maximin (LHS-Maximin) design (see e.g [SWN03]). This design is one of the most widespread non-iterative designs in Kriging. A LHS design is a set of n points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ so that, for $1 \leq k \leq d$ and $1 \leq i \leq n$, there is exactly one j so that $x_k^{(j)} \in [\frac{i-1}{n}, \frac{i}{n}]$. Intuitively, the one-dimensional projections of a LHS design are rather uniformly spread on $[0,1]$. Then, a LHS-Maximin design is a LHS-design that maximizes

$$\min_{i \neq j} |\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}|^2, \tag{6.21}$$

which has the advantage of avoiding almost equal observation points, which would give a redundant information on the values of the Gaussian process Y.

(6.21) is difficult to optimize numerically, because the input space is the set of the LHS designs, which is a subset of $[0,1]^{nd}$. (6.21) is thus a very high dimensional optimization problem. To address the optimization problem (6.21), we generate randomly 1000 LHS designs, and keep the one that maximizes (6.21). To generate randomly a LHS design, we generate d random permutations of $\{1, ..., n\}$: $i_1^k, ..., i_n^k$, $1 \leq k \leq d$. The LHS design we generate is then defined by, for $1 \leq j \leq n$ and $1 \leq k \leq d$, $x_k^{(j)} = \frac{i_j^k - 1}{n} + X_{j,k}$ where the nd $X_{j,k}$ are iid random variables following the uniform distribution on $[0, \frac{1}{n}]$.

Let us notice that this method for generating LHS-Maximin DOEs is the method used by the Matlab function *lhsdesign(...,'maximin',k)* which generates k LHS designs with default $k = 5$. Notice also that other optimization methods can be used to address (6.21). We refer to table 1 of [VVB10] for a review.

The third DOE is a deterministic sparse regular grid. It is built according to the Smolyak's construction ([Smo63] and see e.g. [GG98], [NR96]) of the family of one-dimensional regular grids $G_k = \{\frac{1}{2^k}, ..., \frac{2^k - 1}{2^k}\}$, for $k \in \mathbb{N}^*$ varying. For a given level l, the DOE obtained from the

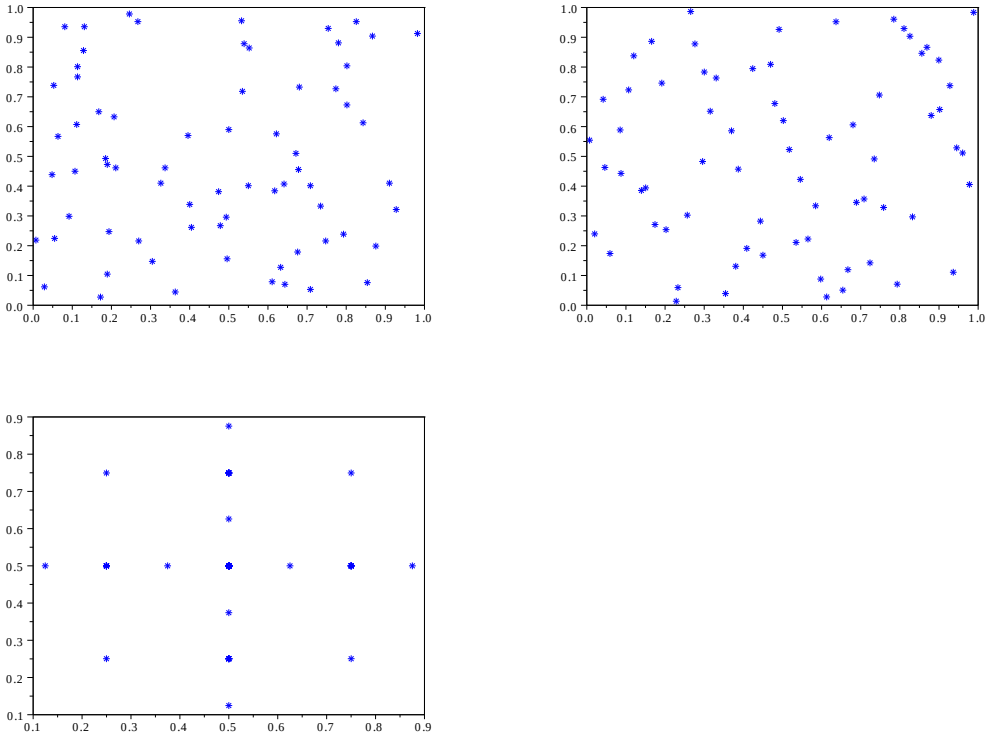

Figure 6.1: For $d = 5$ and $n = 70$, example of a SRS DOE (top-left), a LHS-Maximin DOE (top-right) and the deterministic sparse grid (bottom, $n = 71$). Projection on the first 2 base vectors. The SRS and LHS-Maximin DOEs are much less regular than the regular grid. Although the projections of the SRS and LHS-Maximin DOEs are similar, the criterion (6.21) is 0.16^2 for the SRS DOE and 0.26^2 for the LHS-Maximin DOE.

Smolyak's construction is as follows:

$$\bigcup_{\substack{k_1,\ldots,k_d\in\mathbb{N}^* \\ k_1+\ldots+k_d\leq l+d-1}} G_{k_1}\times\ldots\times G_{k_d}. \tag{6.22}$$

(6.22) is a sparse subset of the fully tensorized regular grid $G_l \times ... \times G_l$. (6.22) uses much fewer observation points than the fully tensorized regular grid, and remains thus tractable for larger dimension d. The Smolyak's construction is classically used for numerical interpolation and integration. For integration of smooth functions, the decay rate of the error, as a function of the number of observation points, is faster for the Smolyak's construction (6.22) than for the fully tensorized regular grid (see e.g. [Nou09] for details). For dimension $d = 5$ and level $l = 3$, the Smolyak's construction yields $n = 71$ observation points. We show in figure 6.1 the projection of this sparse grid on the first two base vectors.

The three DOEs are representative of the classical DOEs that can be used for interpolation of functions, going from the most irregular ones (SRS) to the most regular ones (sparse grid). In figure 6.1, we plot, for $n = 70$ and $d = 5$, the projections on the two first base vectors of two realizations of the SRS and LHS-Maximin DOEs and of the regular grid. The SRS and LHS-Maximin DOEs are much less regular than the regular grid. The projections on the two first

base vectors of the SRS and LHS-Maximin DOEs are similar. However, this can be misleading, since this kind of projection is not representative of the distance between different 5-dimensional observation points. Inspecting the criterion (6.21), we see that its value is 0.16^2 for the SRS DOE and 0.26^2 for the LHS-Maximin DOE.

Families of correlation functions studied

We first study the isotropic Matérn correlation function family of chapter 3, parameterized by the vector of correlation lengths $\boldsymbol{\ell} = (\ell_1, ..., \ell_d)$ and the smoothness parameter ν. We recall that R is Matérn $(\boldsymbol{\ell}, \nu)$ when

$$R(\boldsymbol{h}) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu}|\boldsymbol{h}|_{\boldsymbol{\ell}}\right)^{\nu} K_{\nu}\left(2\sqrt{\nu}|\boldsymbol{h}|_{\boldsymbol{\ell}}\right), \tag{6.23}$$

with $|\boldsymbol{h}|_{\boldsymbol{\ell}} = \sqrt{\sum_{i=1}^{d} \frac{h_i^2}{\ell_i^2}}$, Γ the Gamma function and K_{ν} the modified Bessel function of second order.

We also study the power-exponential correlation function family of chapter 3, parameterized by the vector of correlation lengths $\boldsymbol{\ell} = (\ell_1, ..., \ell_d)$ and the power p. We recall that R is power-exponential $(\boldsymbol{\ell}, p)$ when

$$R(\boldsymbol{h}) = \exp\left(-\sum_{i=1}^{d}\left(\frac{|h_i|}{\ell_i}\right)^p\right). \tag{6.24}$$

Remark 6.5. *In this chapter 6, we study an isotropic (up to a scaling of the axis) covariance function (the isotropic Matérn model) and a tensor-product covariance function (the power-exponential covariance model). While the main goal of chapter 6 is to compare ML and CV, it could also be interesting, in future work, to compare tensority and isotropy more specifically. Especially, the two tensor-product and isotropic versions of the Matérn model could be compared. This distinction between the two versions could have a significant impact for DOEs that rely strongly on the choice of axis, such as the sparse regular grid.*

Influence of the model error

We study the influence of the model error, i.e. the difference between R_1 and R_2. For different pairs R_1, R_2, we generate $n_p = 50$ SRS and LHS learning samples, and the deterministic sparse grid presented above. We compare the empirical means of the two integrated criteria IRTR and IBTR for the different DOEs and for ML and CV. IRTR and IBTR are calculated on a large test sample of size 5000. We take $n = 70$ for the learning sample size (actually $n = 71$ for the regular grid) and $d = 5$ for the dimension.

For the pairs R_1, R_2, we consider the three following cases. First, R_1 is power-exponential $((1.2, ..., 1.2), 1.5)$ and R_2 is power-exponential $((1.2, ..., 1.2), p_2)$ with varying p_2. Second, R_1 is Matérn $((1.2, ..., 1.2), 1.5)$ and R_2 is Matérn $((1.2, ..., 1.2), \nu_2)$ with varying ν_2. Finally, R_1 is Matérn $((1.2, ..., 1.2), 1.5)$ and R_2 is Matérn $((\ell_2, ..., \ell_2), 1.5)$ with varying ℓ_2.

In figure 6.2, we plot the results for the SRS DOE. We clearly see that when the model error becomes large, CV becomes more efficient than ML in the sense of IRTR. Looking at (6.20), one can see that the IRTR is composed of IBTR and of an integrated relative variance term. When R_2 becomes different from R_1, the IBTR contribution increases faster than the

integrated relative variance contribution, especially for ML. Hence, the main reason why CV is more robust than ML to model misspecification is that its bias increases more slowly with the model misspecification.

In figure 6.3 we plot the results for the LHS-Maximin DOE. The results are similar to those of the SRS DOE. They also appear to be slightly more pronounced, the IRTR of CV being smaller than the IRTR of ML for a smaller model error.

In figure 6.4, we plot the results for the regular grid DOE. The results are radically different from the ones obtained with the SRS and LHS-Maximin designs. The first comment is that the assessment of predictive variances is much more difficult in case of model misspecification (the minimum, between ML and CV, of IRTR for the SRS and LHS-Maximin designs is smaller than that of the regular grid and the difference is even stronger when considering the maximum). This is especially true for misspecifications on the exponent for the power-exponential correlation function and on the smoothness parameter for the Matérn function. The second comment is that this time CV appears to be less robust than ML to model misspecification. In particular, its bias increases faster than ML bias with model misspecification and can be very large. Indeed, having observation points that are on a regular grid, CV estimates a σ^2 hyper-parameter adapted only to predictions on the regular grid. Because of the correlation function misspecification, this does not generalize at all to predictions outside the regular grid. Hence, CV is efficient to assess predictive variances at the points of the regular grid but not to assess predictive variances outside the regular grid. This is less accentuated for ML because ML estimates a general-purpose σ^2 and not a σ^2 for the purpose of assessing predictive variances at particular points. Furthermore, it is noted in [IBFM10] that removing a point from a highly structured DOE breaks its structure, which may yield overpessimistic CV results.

We conclude from these numerical results that, for the SRS and LHS-Maximin designs of experiments, CV is more robust to model misspecification. It is the contrary for the regular grid, for the structural reasons presented above. This being said, we do not consider the regular grid anymore in the following numerical results and only consider the SRS and LHS-Maximin designs. Let us finally notice that the regular grid is not particularly a Kriging-oriented DOE. Indeed, for instance, for $n = 71$, it remains only 17 distinct points when projecting on the first two base vectors (figure 6.1).

Influence of the number of points

Using the same procedure as for the influence of the model error presented above, we still set $d = 5$ and we vary the learning sample size n. The pair R_1, R_2 is fixed in the three following different cases. First, R_1 is power-exponential $((1.2, ..., 1.2), 1.5)$ and R_2 is power-exponential $((1.2, ..., 1.2), 1.7)$. Second, R_1 is Matérn $((1.2, ..., 1.2), 1.5)$ and R_2 is Matérn $((1.2, ..., 1.2), 1.8)$. Finally, R_1 is Matérn $((1.2, ..., 1.2), 1.5)$ and R_2 is Matérn $((1.8, ..., 1.8), 1.5)$. This time, we do not consider integrated quantities of interest and focus on the prediction on the point $\boldsymbol{x}^{(0)}$ having all its components set to $\frac{1}{2}$ (center of domain).

In figure 6.5 we plot the results for the SRS DOE. The first comment is that, as n increases, the BTR does not vanish, but seems to reach a limit value. This limit value is smaller for CV for the three pairs R_1, R_2. Recalling from (6.17) that RTR is the sum of BTR and of a relative

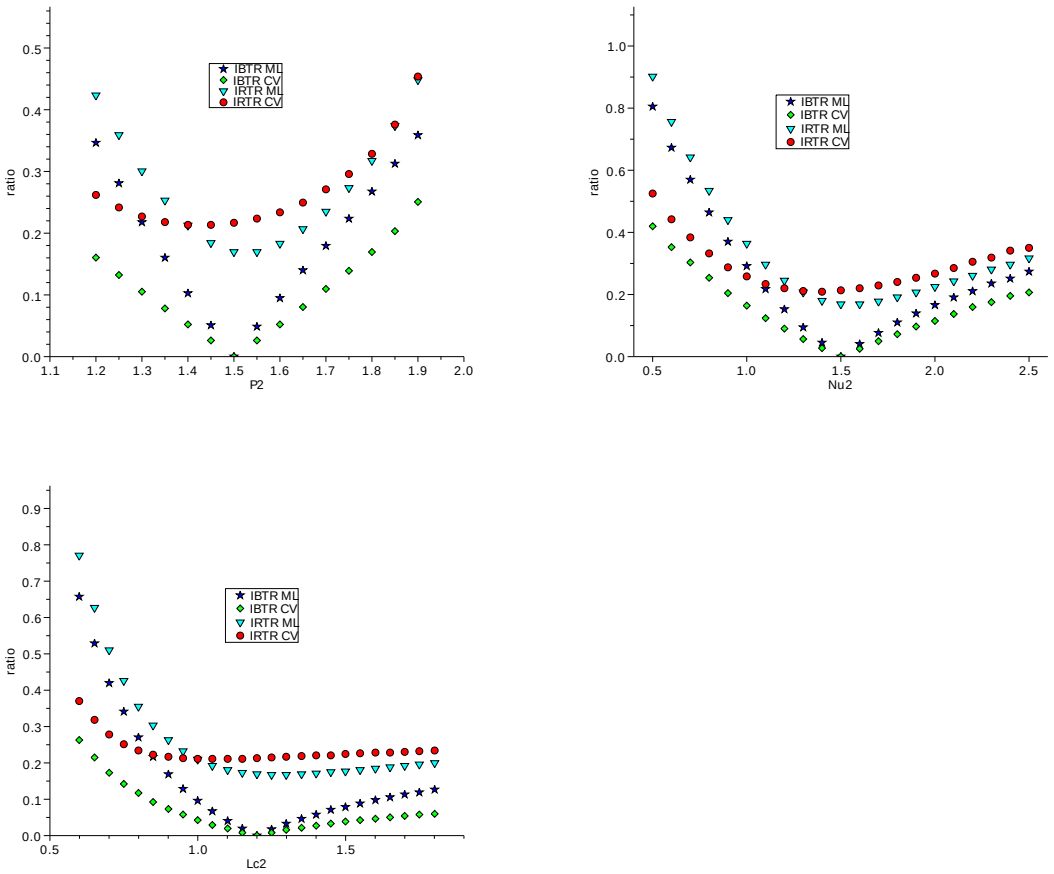

Figure 6.2: Influence of the model error for the SRS DOE. Plot of the IRTR and IBTR integrated criteria for ML and CV. Top-left: power-exponential correlation function with error on the exponent, the true exponent is $p_1 = 1.5$ and the model exponent p_2 varies in $[1.2, 1.9]$. Top-right: Matérn correlation function with error on the smoothness parameter, the true smoothness parameter is $\nu_1 = 1.5$ and the model smoothness parameter ν_2 varies in $[0.5, 2.5]$. Bottom: Matérn correlation function with error on the correlation length, the true correlation length is $\ell_1 = 1.2$ and the model correlation length ℓ_2 varies in $[0.6, 1.8]$. ML is optimal when there is no model error while CV is more robust to model misspecifications.

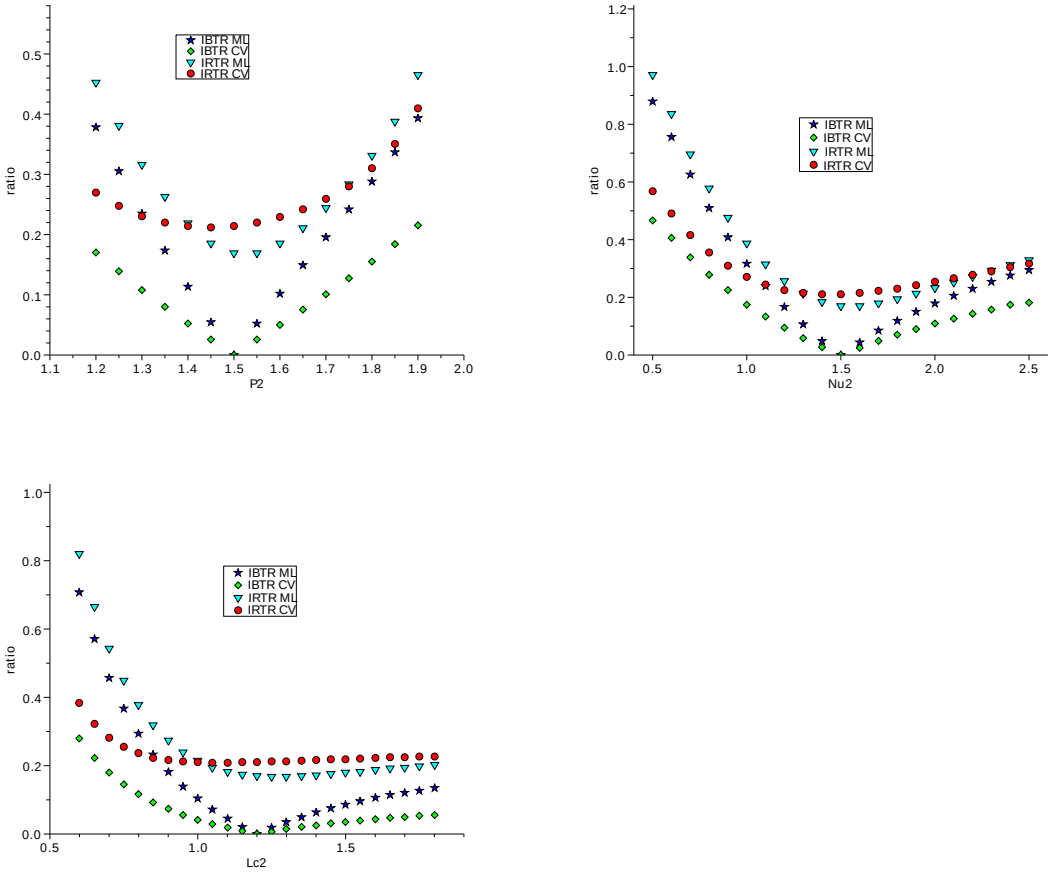

Figure 6.3: Same setting as in figure 6.2, but with the LHS-Maximin DOE. ML is optimal when there is no model error while CV is more robust to model misspecifications.

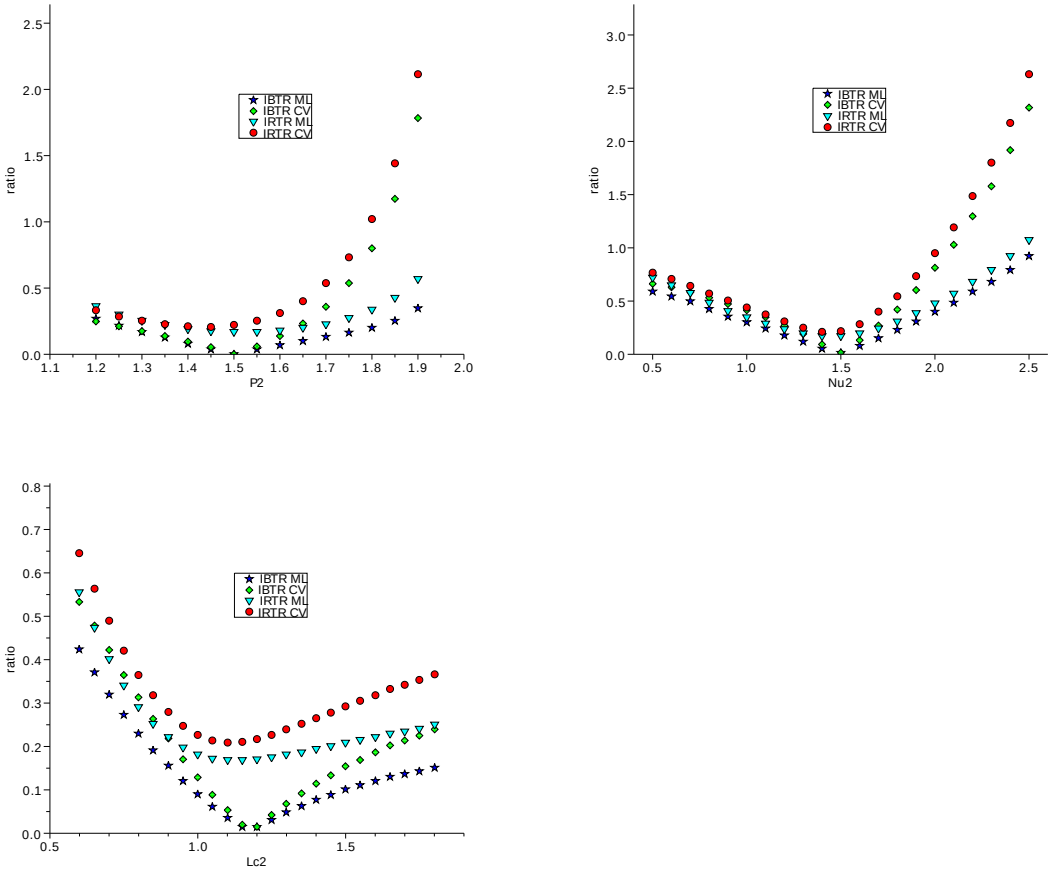

Figure 6.4: Same setting as in figure 6.2 but with the regular sparse grid DOE. The results are radically different from the ones obtained with the SRS and LHS-Maximin DOEs. This time CV is less robust to misspecifications of the correlation function.

variance term, we observe that this relative variance term decreases and seems to vanish when n increases (because BTR becomes closer to RTR). The decrease is much slower for the error on the correlation length than for the two other errors on the correlation function. Furthermore, the relative variance term decreases more slowly for CV than for ML. Finally, because CV is better than ML for the BTR and worse than ML for the relative variance, and because the contribution of BTR to RTR increases with n, the ratio of the RTR of ML over the RTR of CV increases with n. This ratio can be smaller than 1 for very small n and eventually becomes larger than 1 as n increases (meaning that CV does better than ML).

In figure 6.6 we plot the results for the LHS-Maximin DOE. The results are similar to those of the SRS DOE. The RTR of CV is smaller than the RTR of ML for a slightly smaller n. This confirms the results above on the influence of the model error, where the model error for which the IRTR of ML reaches the IRTR of CV is smaller for LHS-Maximin than for SRS.

6.3 Estimation of variance and correlation hyper-parameters

The first goal of this section 6.3 is to illustrate the results of section 6.2 on the estimation of the variance hyper-parameter on analytical functions, instead of realizations of Gaussian processes, as was the case in section 6.2. Indeed, the study of section 6.2 is more related to the theory of Kriging (we work on Gaussian processes) while this section is more related to the application of Kriging (modeling of deterministic functions as realizations of Gaussian processes). The second goal of this section 6.3 is to generalize section 6.2 to the case where correlation hyper-parameters are estimated from data.

6.3.1 Procedure

ML and CV estimations of covariance hyper-parameters

We consider a set of observations $(\boldsymbol{x}^{(1)}, y_1), ..., (\boldsymbol{x}^{(n)}, y_n)$ as in section 6.2, and the family $\left\{\sigma^2 R_{\boldsymbol{\theta}}, \sigma^2 > 0, \boldsymbol{\theta} \in \Theta\right\}$ of stationary covariance functions, with $R_{\boldsymbol{\theta}}$ a stationary correlation function, and Θ a finite-dimensional set. We denote by $\mathbb{E}_{\boldsymbol{\theta}}$ and $Var_{\boldsymbol{\theta}}$ the means and variances with respect to the distribution of a stationary Gaussian process with mean zero, variance one and correlation function $R_{\boldsymbol{\theta}}$. We denote by $\mathbf{R}_{\boldsymbol{\theta}}$ the correlation matrix of the training sample with correlation function $R_{\boldsymbol{\theta}}$, that is $(\mathbf{R}_{\boldsymbol{\theta}})_{i,j} = R_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})$.

The ML estimate of $(\sigma^2, \boldsymbol{\theta})$ is, as we have seen in chapter 3,

$$\hat{\boldsymbol{\theta}}_{ML} \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} |\mathbf{R}_{\boldsymbol{\theta}}|^{1/n} \hat{\sigma}^2_{ML}(\mathbf{R}_{\boldsymbol{\theta}}), \tag{6.25}$$

with $\hat{\sigma}^2_{ML}(\mathbf{R}_{\boldsymbol{\theta}})$ as in (6.7), and

$$\hat{\sigma}^2_{ML} = \hat{\sigma}^2_{ML}(\mathbf{R}_{\hat{\boldsymbol{\theta}}_{ML}}).$$

For CV we recall the estimation for the hyper-parameter $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_{LOO} \in \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i,\boldsymbol{\theta}})^2, \tag{6.26}$$

with $\hat{y}_{i,\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$.

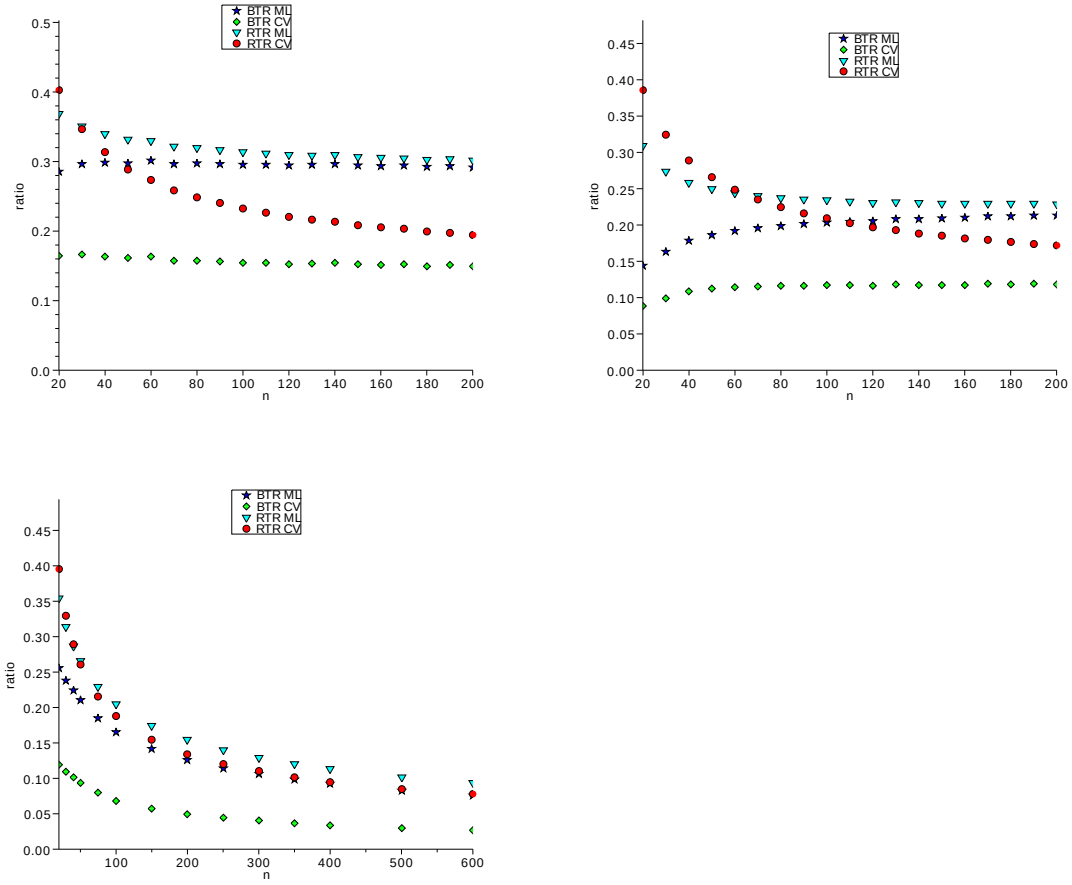

Figure 6.5: Influence of the number n of observation points for the SRS DOE. Plot of the RTR and BTR criteria for prediction at the center of the domain and for ML and CV. Top-left: power-exponential correlation function with error on the exponent, the true exponent is $p_1 = 1.5$ and the model exponent is $p_2 = 1.7$. Top-right: Matérn correlation function with error on the smoothness parameter, the true smoothness parameter is $\nu_1 = 1.5$ and the model smoothness parameter is $\nu_2 = 1.8$. Bottom: Matérn correlation function ($\nu = \frac{3}{2}$) with error on the correlation length, the true correlation length is $\ell_1 = 1.2$ and the model correlation length is $\ell_2 = 1.8$.

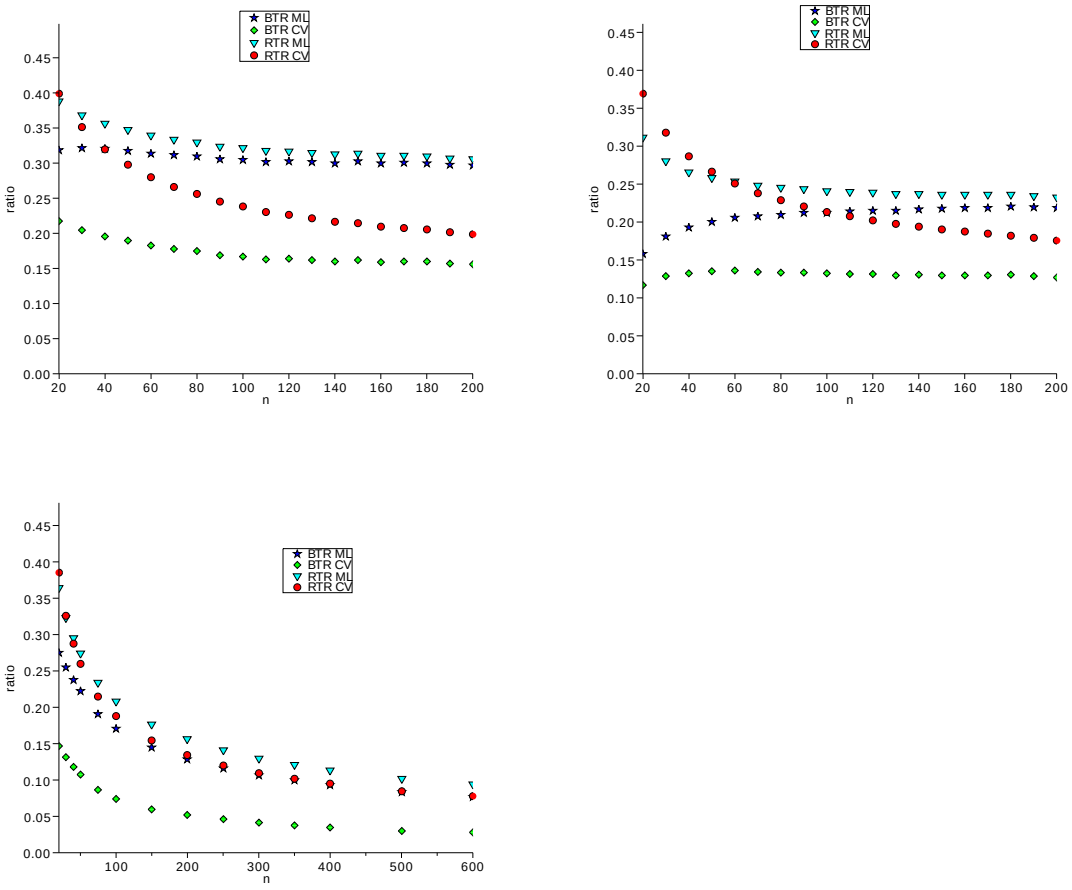

Figure 6.6: Same setting as in figure 6.5, but with the LHS-Maximin DOE.

We remember from chapter 3 that, notably after using proposition 2.35 for the estimator $\hat{\boldsymbol{\theta}}_{LOO}$, the functions to minimize in (6.25) and (6.26) are respectively:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\log|\mathbf{R}_{\boldsymbol{\theta}}| + \log\left(\boldsymbol{y}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}\right) \tag{6.27}$$

and

$$LOO(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{y}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}Diag(\mathbf{R}_{\boldsymbol{\theta}}^{-1})^{-2}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}. \tag{6.28}$$

We have seen in chapter 3 that we dispose of the closed-form expressions of the gradients of $\mathcal{L}(\boldsymbol{\theta})$ and $LOO(\boldsymbol{\theta})$, as functions of the first-order derivatives of the correlation function. The evaluations of the two functions and their gradients have similar computational complexities of the order of $O(n^3)$.

Once we have the closed-form expressions of the gradients at hand, our optimization procedure is based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton optimization method (see e.g. chapter 6 of [NW06]), implemented in the Scilab function *optim*. Since the functions $\mathcal{L}(\boldsymbol{\theta})$ and $LOO(\boldsymbol{\theta})$ may have multiple local minima, the BFGS method is run several times, by taking the initial points in a LHS-Maximin design. The presence of multiple local minima is discussed e.g in [MS04]. An important point is that, when $\boldsymbol{\theta}$ is a correlation length, we recommend to use its logarithm to run the optimization. Indeed a correlation length acts as a multiplier in the correlation, so that using its log ensures that a given perturbation has the same importance, whether applied to a large or a small correlation length. Furthermore, when one wants to explore the space of correlation lengths uniformly, as is the case with a LHS design, directly using the correlation lengths may give too much emphasis on large correlation lengths, which is avoided by using their log.

Another important issue is the numerical inversion of the correlation matrix. This issue is even more significant when the correlation matrix is ill-conditioned, which happens when the correlation function is smooth (Gaussian or Matérn with a large smoothness parameter). To tackle this issue we recommend to use the numerical nugget effect. More specifically, for a given correlation matrix $\mathbf{R}_{\boldsymbol{\theta}}$, we actually compute $\mathbf{R}_{\boldsymbol{\theta}} + \tau^2\mathbf{I}_n$, with $\tau^2 = 10^{-8}$ in our simulations. A detailed analysis of the influence of the nugget effect on the hyper-parameter estimation and on the Kriging prediction is carried out in [AC12]. Notice also that we have seen in chapter 4 that the numerical nugget effect ensures a fixed-domain asymptotic consistency of the Kriging predictions, for both the cases where the observations come with measurement errors or not. However, for the CV estimation of $\boldsymbol{\theta}$, when the correlation function belongs to the Gaussian family, or the Matérn family with large smoothness parameter, another structural problem appears. For $\hat{\sigma}^2_{LOO}$ very large, as the overall predictive variance term $\hat{\sigma}^2_{LOO}(1 - \boldsymbol{r}_{\boldsymbol{\theta}}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}})$ has the same order of magnitude as the squared observations, the term $1 - \boldsymbol{r}_{\boldsymbol{\theta}}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}}$ is very small. Hence, a fixed numerical error on the inversion of $\mathbf{R}_{\boldsymbol{\theta}}$, however small it is, may cause the term $1 - \boldsymbol{r}_{\boldsymbol{\theta}}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}}$ to be negative. This is what we observe for the CV case when fitting e.g the correlation lengths of a Gaussian correlation function. The heuristic scheme is that large correlation lengths are estimated, which yields large $\hat{\sigma}^2_{LOO}$, which yields small $(1 - \boldsymbol{r}_{\boldsymbol{\theta}}\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{r}_{\boldsymbol{\theta}})$, so possibly negative ones. Notice however that the relative errors of the Kriging prediction terms $\boldsymbol{r}_{\boldsymbol{\theta}}^t\mathbf{R}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}$ are correct. It is noted in [MS04] p.7 that CV may overestimate correlation lengths. Hence, to have appropriate predictive variances, one has to ensure that the estimated

correlation lengths are not too large. Two possible solutions are to penalize either too large correlation lengths or too large $\hat{\sigma}^2_{LOO}$ in the minimization of $LOO(\boldsymbol{\theta})$. We choose here the second solution because our experience is that the ideal penalty on the correlation lengths, both ensuring reliable predictive variance computation and having a minimal effect on the $\boldsymbol{\theta}$ estimation, depends on the DOE substantially. In practice, we use a penalty for $\hat{\sigma}^2_{LOO}$ starting at 1000 times the empirical variance $\frac{1}{n}\boldsymbol{y}^t\boldsymbol{y}$. This penalty is needed only for CV when the correlation function is Gaussian or Matérn with free smoothness parameter.

Prediction criteria

We consider a deterministic function f on $[0,1]^d$. We generate $n_p = 100$ LHS-Maximin training samples of the form $\boldsymbol{x}^{(a,1)}, f(\boldsymbol{x}^{(a,1)}), ..., \boldsymbol{x}^{(a,n)}, f(\boldsymbol{x}^{(a,n)})$. We denote $y_i^{(a)} = f(\boldsymbol{x}^{(a,i)})$, $i = 1, ..., n_p$. From each training sample, we estimate σ^2 and $\boldsymbol{\theta}$ with the ML and CV methods presented above.

We consider simple Kriging in this section 6.3, except in the end of subsection 6.3.2 where we consider universal Kriging. We are interested in two criteria, based on the Kriging prediction with estimated covariance parameters, on a large Monte Carlo test sample $\boldsymbol{x}^{(t,1)}, f(\boldsymbol{x}^{(t,1)}), ...,$ $\boldsymbol{x}^{(t,n_t)}, f(\boldsymbol{x}^{(t,n_t)})$ on $[0,1]^d$ ($n_t = 10000$). We denote $y_{t,i} = f(\boldsymbol{x}^{(t,i)})$, $\hat{y}(\boldsymbol{x}^{(t,i)}) = \mathbb{E}_{\hat{\boldsymbol{\theta}}}(y_{t,i}|\boldsymbol{y}^{(a)})$ and $\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(t,i)}) = \hat{\sigma}^2 Var_{\hat{\boldsymbol{\theta}}}(y_{t,i}|\boldsymbol{y}^{(a)})$, where $\hat{\sigma}^2$ and $\hat{\boldsymbol{\theta}}$ come from either the ML or CV method.

The first criterion is the Mean Square Error (MSE). It evaluates the prediction capability of the estimated correlation function $R_{\hat{\boldsymbol{\theta}}}$:

$$\frac{1}{n_t}\sum_{i=1}^{n_t}(y_{t,i} - \hat{y}(\boldsymbol{x}^{(t,i)}))^2. \tag{6.29}$$

The second criterion is the Predictive Variance Adequation (PVA):

$$\left|\log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}\frac{(y_{t,i} - \hat{y}(\boldsymbol{x}^{(t,i)}))^2}{\hat{\sigma}^2\hat{c}^2(\boldsymbol{x}^{(t,i)})}\right)\right|. \tag{6.30}$$

This criterion evaluates the quality of the predictive variances given by the estimated covariance hyper-parameters $\hat{\sigma}^2, \hat{\boldsymbol{\theta}}$. The smaller the PVA is, the better it is because the predictive variances are globally of the same order than the prediction errors, so that the confidence intervals are reliable. We use the logarithm in order to give the same weight to relative overestimation and to relative underestimation of the prediction errors.

We finally average the two criteria over the n_p training samples.

Analytical functions studied

We study the two following analytical functions. The first one, for $d = 3$, is the Ishigami function:

$$f(x_1, x_2, x_3) = \tag{6.31}$$
$$\sin(-\pi + 2\pi x_1) + 7\sin((-\pi + 2\pi x_2))^2 + 0.1(-\pi + 2\pi x_3)^4 \sin(-\pi + 2\pi x_1).$$

The second one, for $d = 10$, is a simplified version of the Morris function [Mor91],

$$f(\boldsymbol{x}) = \sum_{i=1}^{10} w_i(\boldsymbol{x}) + \sum_{1 \le i < j \le 6} w_i(\boldsymbol{x}) w_j(\boldsymbol{x}) + \sum_{1 \le i < j < k \le 5} w_i(\boldsymbol{x}) w_j(\boldsymbol{x}) w_k(\boldsymbol{x}) + w_1(\boldsymbol{x}) w_2(\boldsymbol{x}) w_3(\boldsymbol{x}) w_4(\boldsymbol{x}),$$

$$\text{with} \qquad w_i(\boldsymbol{x}) = \begin{cases} 2\left(\frac{1.1 x_i}{x_i + 0.1} - 0.5\right), & \text{if } i = 3, 5, 7 \\ 2(x_i - 0.5) & \text{otherwise.} \end{cases}$$

Both the Ishigami and Morris functions are smooth functions. For the Morris function, the low-index components have the largest influence since they appear in most of the sums in the expression of the Morris function. Furthermore, notice the two different expressions for the $w_i(\boldsymbol{x})$, depending on the index i. The Morris function is hence anisotropic.

6.3.2 Results and discussion

Results with enforced correlation lengths

We work with the Ishigami function, with $n = 100$ observation points. For the correlation function family, we study the tensorized exponential and Gaussian families (power-exponential family of (6.24) with enforced $p = 1$ for exponential and $p = 2$ for Gaussian).

For each of these two correlation models, we enforce three vectors $\boldsymbol{\ell}$ of correlation lengths for R: an arbitrary isotropic correlation length, a well-chosen isotropic correlation length and three well-chosen correlation lengths along the three dimensions. To obtain a well-chosen isotropic correlation length, we generate $n_p = 100$ LHS-Maximin DOEs, for which we estimate the correlation length by ML and CV as described above. We calculate each time the MSE on a test sample of size 10000 and the well-chosen correlation length is the one with the smallest MSE among the $2n_p$ estimated correlation lengths. The three well-chosen correlation lengths are obtained similarly. The three vectors of correlation lengths yield an increasing prediction quality.

The results are presented in table 6.1. Comparing line 3 against line 6, we see that the Gaussian family is more appropriate than the exponential one for the Ishigami function. Indeed, it yields the smallest MSE among the cases when one uses three different correlation lengths, and the PVA is quite small as well. This could be anticipated since the Ishigami function is smooth, so a Gaussian correlation model (smooth trajectories) is more adapted than an exponential one (rough trajectories).

Notice, nevertheless, from lines $1, 2$ and $4, 5$, that the prediction results for the Gaussian model appear significantly more sensitive to non-optimal choices of the correlation lengths. Indeed, the prediction error becomes larger than that of the exponential model for the cases of the arbitrary and well-chosen isotropic correlation length.

Finally, we see that CV yields much smaller PVAs than ML in line 1, 2, 3 and 4, in the cases when the correlation function is not appropriate. For line 6, which is the most appropriate correlation function, ML yields a PVA comparable to CV and for line 5, ML PVA is smaller than CV PVA. All these comments are in agreement with the main result of subsection 6.2.2:

Correlation model	Enforced hyper-parameters	MSE	PVA
exponential	$[1, 1, 1]$	2.01	$ML : 0.50\ CV : 0.20$
exponential	$[1.3, 1.3, 1.3]$	1.94	$ML : 0.46\ CV : 0.23$
exponential	$[1.20, 5.03, 2.60]$	1.70	$ML : 0.54\ CV : 0.19$
Gaussian	$[0.5, 0.5, 0.5]$	4.19	$ML : 0.98\ CV : 0.35$
Gaussian	$[0.31, 0.31, 0.31]$	2.03	$ML : 0.16\ CV : 0.23$
Gaussian	$[0.38, 0.32, 0.42]$	1.32	$ML : 0.28\ CV : 0.29$

Table 6.1: Mean of the MSE and PVA criteria for the Ishigami function for different fixed correlation models. The MSE is the same between ML and CV as the same correlation function is used. When the correlation model is misspecified, the MSE is large and CV does better than ML for the PVA criterion.

The ML estimation of σ^2 is more appropriate when the correlation function is well-specified while the CV estimation is more appropriate when the correlation function is misspecified.

Results with estimated correlation lengths

We use the exponential and Gaussian models, as when the correlation lengths were enforced, as well as the Matérn model of (6.23). We distinguish two subcases for the vector $\boldsymbol{\ell}$ of correlation lengths. In **Case i** we estimate a single isotropic correlation length, while in **Case a** we estimate d correlation lengths for the d dimensions.

The numerical optimization problem We first discuss the optimization of $\mathcal{L}(\boldsymbol{\theta})$ and $LOO(\boldsymbol{\theta})$ in (6.27) and (6.28), in the case of the Ishigami function with $n = 70$ observation points and with the Gaussian model for Case a. One of the n_p LHS-Maximin DOEs is thus randomly selected and fixed. The dimension of the optimization problem is 3 and the variables are $\ln \ell_1, \ln \ell_2, \ln \ell_3$. We restrict the optimization in the subset $[\ln 0.1, \ln 100]^3$.

In figure 6.7, we plot the level sets of $(\ln \ell_1, \ln \ell_2) \to \min_{\ell_3} f(\ell_1, \ell_2, \ell_3)$, where f is either $\mathcal{L}$ or LOO. We first observe that the two criteria functions have several local minima (we distinguish at least two for both functions). This observation is true in our general experience in optimizing ML and CV criteria throughout the PhD thesis. As a consequence, it can not be overstated that we recommend not to use an optimization method that is only local. Specifically, using a single BFGS algorithm with arbitrary starting point can result in only reaching a local minimum of the criterion to minimize. As we have discussed, we run n_r BFGS methods, where the n_r initial points constitute a space-filling design of experiment of $[\ln 0.1, \ln 100]^3$. We use $n_r = 150$ in the present illustration.

The second comment for figure 6.7 is that the ML criterion appears convex in a large area around its global minimizer. This is not the case for CV, where we distinguish two other local minimizers close to the global minimizer. The CV criterion is hence, somehow, more difficult to optimize than ML, using a local search-based optimization method. This fact is also general in our experience.

In figure 6.7, for the optimization of $\mathcal{L}$ and LOO, we also plot the localization of the 50

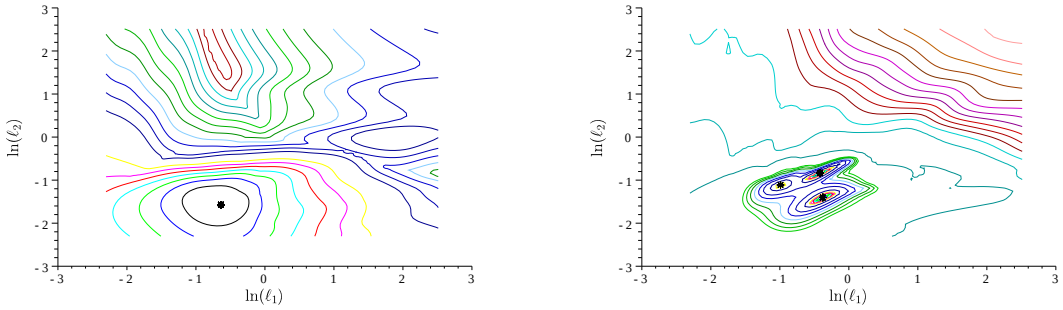

Figure 6.7: Estimation of the correlation length vector $\boldsymbol{\ell}$ for the Gaussian model for Case a. Case of the Ishigami function where a LHS-Maximin DOE of $n = 70$ observation points is considered. Plot of the level sets of the criteria $(\ln \ell_1, \ln \ell_2) \to \min_{\ell_3} \mathcal{L}(\ell_1, \ell_2, \ell_3)$ (left) and $(\ln \ell_1, \ln \ell_2) \to \min_{\ell_3} \ln(LOO(\ell_1, \ell_2, \ell_3))$ (right). The black stars are the 50 terminal points of the 50 BFGS runs that yield the smallest criteria among 150 runs. The optimization is more difficult for CV than ML, with the BFGS method with multiple starting points that we use. For ML, the 50 best runs all reach the global minimizer, while for CV they are distributed between the global minimizer and 2 local minimizers.

terminal points of the 50 BFGS runs that yield the smallest criterion values, among the $n_r = 150$ BFGS runs. For the ML case, the 50 terminal points are equal to the global minimizer. For the CV case, they are distributed between 3 local minimizers (including the global one) that are relatively close to one another. This is a confirmation that the optimization for the CV criterion is more difficult than for ML.

When we observe all the 150 terminal points, they converge to local minimizers, converge to boundary points of the optimization domain, or do not converge, in the case, for instance, when the BFGS method reaches a determined maximum number of iteration and stops. Finally, for the ML case, 57 of the 150 BFGS runs converge to the global minimizer, against 16 for the CV case.

In figure 6.8, we plot the equivalent of figure 6.7, but for the exponential covariance model. We see that, contrary to figure 6.7, both the ML and CV criterion functions are strongly unimodal. As a consequence, we have also observed that all the BFGS runs converge to the global minimizer. Generally speaking, we have noticed, throughout the PhD thesis, that the optimization problem is more difficult with the Gaussian model, than with the Matérn model with fixed and relatively small smoothness parameter. See also [Ste99], p.173.

Prediction results of the estimated hyper-parameters The discussion on the numerical optimization problem being concluded, in table 6.2 we now present the prediction results of the estimated hyper-parameters for the Ishigami and Morris functions, with $n = 100$ observation points. We address the exponential, Gaussian, and Matérn with free smoothness parameter models. For both the Ishigami and Morris functions, the Gaussian model yields smaller MSEs than the exponential model. Indeed, both functions are smooth. Over the different DOEs, we observe that the estimated Matérn smoothness hyper-parameters are large, so that the MSEs and

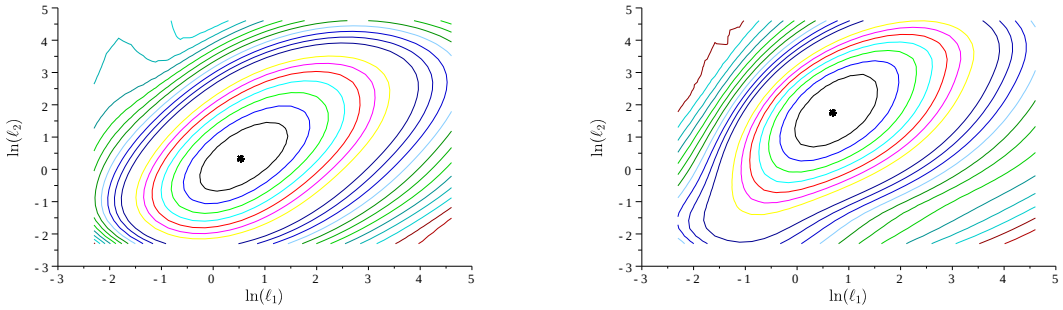

Figure 6.8: Same setting as in figure 6.7, but for the estimation of the correlation length vector $\boldsymbol{\ell}$ for the exponential model for Case a. For both ML (left) and CV (right), the criterion functions are strongly unimodal.

the PVAs for the Matérn model are similar to those of the Gaussian model. Let us notice that for the Ishigami function, the relatively large number $n = 100$ of observation points is required for the Gaussian correlation model to be more adapted than the exponential one. Indeed, in table 6.3, we show the same results with $n = 70$ where the Gaussian model yields relatively larger MSEs and substantially larger PVAs. Our interpretation is that the linear interpolation yielded by the exponential correlation function can be sufficient, even for a smooth function, if there are not enough observation points. We also notice that, generally, estimating different correlation lengths (Case a) yields a smaller MSE than estimating one isotropic correlation length (Case i). In our simulations this is always true except for the Ishigami function with the exponential model. Indeed, we see in table 6.1 that we get a relatively small benefit for the Ishigami function from using different correlation lengths. Here, this benefit is compensated by an error in the estimation of the 3 correlation lengths with $n = 100$ observation points. The overall conclusion is that the Gaussian and Matérn correlation models are more adapted than the exponential one, and that using different correlation lengths is more adapted than using an isotropic one, provided that there are enough data to estimate these correlation lengths.

In the exponential case, for both Cases i and a, CV always yields a smaller PVA than ML and yields a MSE that is smaller or similar. In Case a, for the Gaussian and Matérn correlation functions, the most adapted ones, ML always yields MSEs and PVAs smaller than CV or similar. Furthermore, for the Morris function with Matérn and Gaussian correlation functions, going from Case i to Case a enhances the advantage of ML over CV.

From the discussion above, we conclude that the numerical experiments yield results, for the deterministic functions considered here, that are in agreement with the conclusion of section 6.2: ML is optimal for the best adapted correlation models, while CV is more robust in cases of model misspecification.

Case of universal Kriging

So far, the case of simple Kriging has been considered, for which the underlying Gaussian process is considered centered. The case of universal Kriging, presented in chapter 2, can equally be studied. We recall that, in the universal Kriging case, the Gaussian process is considered to

Function	Correlation model	MSE	PVA
Ishigami	exponential Case i	ML: 1.99 CV: 1.97	ML: 0.35 CV: 0.23
Ishigami	exponential Case a	ML: 2.01 CV: 1.77	ML: 0.36 CV: 0.24
Ishigami	Gaussian Case i	ML: 2.06 CV: 2.11	ML: 0.18 CV: 0.22
Ishigami	Gaussian Case a	ML: 1.50 CV: 1.53	ML: 0.53 CV: 0.50
Ishigami	Matérn Case i	ML: 2.19 CV: 2.29	ML: 0.18 CV: 0.23
Ishigami	Matérn Case a	ML: 1.69 CV: 1.67	ML: 0.38 CV: 0.41
Morris	exponential Case i	ML: 3.07 CV: 2.99	ML: 0.31 CV: 0.24
Morris	exponential Case a	ML: 2.03 CV: 1.99	ML: 0.29 CV: 0.21
Morris	Gaussian Case i	ML: 1.33 CV: 1.36	ML: 0.26 CV: 0.26
Morris	Gaussian Case a	ML: 0.86 CV: 1.21	ML: 0.79 CV: 1.56
Morris	Matérn Case i	ML: 1.26 CV: 1.28	ML: 0.24 CV: 0.25
Morris	Matérn Case a	ML: 0.75 CV: 1.06	ML: 0.65 CV: 1.43

Table 6.2: $n = 100$ observation points. Mean of the MSE and PVA criteria over $n_p = 100$ LHS-Maximin DOEs for the Ishigami ($d = 3$) and Morris ($d = 10$) functions for different fixed correlation models. When the model is misspecified, the MSE is large and the CV does better compared to ML for the MSE and PVA criterion.

Function	Correlation model	MSE	PVA
Ishigami	exponential Case a	ML: 3.23 CV: 2.91	ML: 0.27 CV: 0.26
Ishigami	Gaussian Case a	ML: 3.15 CV: 4.13	ML: 0.72 CV: 0.76

Table 6.3: $n = 70$ observation points. Mean of the MSE and PVA criteria over $n_p = 100$ LHS-Maximin DOEs for the Ishigami ($d = 3$) and Morris ($d = 10$) functions for the exponential correlation model. Contrary to the case $n = 100$ of table 6.2, the Gaussian correlation model does not yield smaller MSEs than the exponential one.

Function	Mean function model	Correlation model	MSE	PVA
Ishigami	constant	exponential Case a	ML: 1.96 CV: 1.74	ML: 0.39 CV: 0.24
Ishigami	affine	exponential Case a	ML: 1.98 CV: 1.75	ML: 0.40 CV: 0.24
Ishigami	constant	Gaussian Case a	ML: 1.54 CV: 1.63	ML: 0.54 CV: 0.54
Ishigami	affine	Gaussian Case a	ML: 1.58 CV: 1.78	ML: 0.57 CV: 0.57

Table 6.4: $n = 100$ observation points. Mean of the MSE and PVA criteria over $n_p = 100$ LHS-Maximin DOEs for the Ishigami ($d = 3$) function and the exponential and Gaussian correlation models. The incorporation of the mean function does not change the conclusions of table 6.2.

have a mean at location $\boldsymbol{x}$ of the form $\sum_{i=1}^{p} \beta_i g_i(\boldsymbol{x})$, with known functions g_i and unknown coefficients β_i. For instance a closed-form formula similar to that of proposition 6.2 can be obtained in the same fashion, and virtual LOO formulas are also available (proposition 2.35). We have chosen to focus on the simple Kriging case because we are able to address as precisely as possible the issue of the covariance function class misspecification, the Kriging model depending only on the covariance function choice. Furthermore it is shown in [Ste99] p.138 that the issue of the mean function choice for the Kriging model is much less crucial than that of the covariance function choice.

Nevertheless, for completeness, in table 6.4 we study, for the Ishigami function, the influence of using a universal Kriging model with either a constant ($\boldsymbol{x} \to \beta_1$) or affine ($\boldsymbol{x} \to \beta_1 + \sum_{i=1}^{d} \beta_i x_i$) mean function. The process is the same as for table 6.2. We first see that using a non-zero mean does not improve significantly the Kriging model. It is possible to observe a slight improvement only with the exponential covariance structure, which we can interpret because a smooth mean function makes the Kriging model more adapted to the smooth Ishigami function. On the contrary, for the Gaussian covariance structure, the mean function over-parameterizes the Kriging model and slightly damages its performances. Let us also notice that CV appears to be more sensitive to this over-parameterization, its MSE increasing with the complexity of the mean function. This can be observed similarly in the numerical experiments in [MS04]. The second overall conclusion is that the main finding of section 6.2 and of table 6.2 is confirmed: CV has smaller MSE and PVA for the misspecified exponential structure, while ML is optimal for the Gaussian covariance structure which is the more adapted and yields the smallest MSE.

6.4 Discussion

In this chapter 6, we have carried out a detailed analysis of ML and CV for the estimation of the covariance hyper-parameters of a Gaussian process, with a misspecified parametric family of covariance functions. This analysis has been carried out by using a two-step approach. We have first studied the estimation of a global variance hyper-parameter, for which the correlation function is misspecified. In this framework, we can control precisely the degree of model misspecification and we obtain closed-form expressions for the mismatch indices that we have introduced. We conclude from the numerical study of these formulas that where the model is

misspecified, CV performs better than ML. Second, we have studied the general case where the correlation hyper-parameters are estimated from data, via numerical experiments on analytical functions. We confirm the results of the first step, and generalize them.

We have also noticed that the conclusion above does not hold for the case where the Design Of Experiments is a regular grid. In this case, CV is less robust than ML to model misspecification, for structural reasons that we have pointed out.

Because of its practical interest shown in this chapter 6, the CV estimation method has been implemented in the DiceKriging R package [RGD12].

Part III

Applications to Uncertainty Quantification for Computer Experiments

Chapter 7

Probabilistic modeling of discrepancy between computer model and experiments

In this chapter 7, we consider the case where observations of a physical system can be made, and where a computer model can be used to predict these observations. This framework is part of the field of statistics dedicated to the design and analysis of computer experiments ([SWMW89], [SWN03]). In the analysis of computer experiments, three classical objectives can be formulated, in the case where experimental data are available.

Validation corresponds to answering the question: how well does the computer model approximate the physical system underlying the experimental results? A reference book on model validation is e.g. [Cac03]. In this chapter 7, the validation problem is related to the two following problems: calibration and prediction.

Calibration corresponds to setting the optional parameters of the computer model, so that it reproduces the physical system as well as possible. This can be done in two ways here. First, we can be interested in associating a variability to these optional parameters, in order to model the variability of the experimental results. Second, we can be interested in estimating a single value for these optional parameters that is best adapted to the representation of the physical system.

Prediction corresponds to improving the computer model predictions of the physical system, and quantifying the uncertainty obtained, by assimilating experimental results. This point of view is slightly different from the point of view of validation, because the objective is less to study the validity of the computer model than to complete it by incorporating the information brought by the experimental results. A recent reference on demonstrating, or refuting, the validity of the actual computer model would be [WCT09].

In this chapter 7, we address the calibration and validation problems. These two problems can be summarized in a single objective: modeling the differences between the observations of the physical system and the computer model results. Gaussian process models, that we have thoroughly studied in parts I and II, play a central role in this context, because of their natural

ability to provide a Bayesian *a priori* distribution on deterministic functions. We will have a confirmation of their importance in this chapter 7, and in the rest of part III.

Chapter 7 is organized as follows. In section 7.1, we detail the framework for computer models and experiments. In section 7.2, we review some methods in the literature, in which the discrepancy between the computer model and the experiments are modeled by a variability of the physical system. In section 7.3, we consider the case where these discrepancies are explained by a model error of the computer model. Subsection 7.3.2 presents the associated methods, in the literature, that make no linear approximation for the computer model. In subsection 7.3.3, we present the methods that rely on a linear approximation of the computer model with respect to its model parameters. These are the methods we have retained, both from a methodology point of view in this chapter 7, and for the application case on the FLICA 4 thermal-hydraulic code in chapter 8.

7.1 Framework for computer models and experiments

In this manuscript, a computer model corresponds to a deterministic function f_{mod} of the form

$$f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}) : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}.$$

This computer model is a representation of a physical system, that is a map of the form

$$f_{real}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}.$$

The scalar output of the physical system is the physical variable of interest. It depends, via the map f_{real}, on a vector $\boldsymbol{x}$ of input quantities, that we call **experimental conditions**.

Remark 7.1. *The map $\boldsymbol{x} \to f_{real}(\boldsymbol{x})$ of the physical system can be considered deterministic, in which case it will be simply called a function. It can also be considered random, meaning that exactly knowing the experimental conditions $\boldsymbol{x}$ is not sufficient to exactly know the physical output $f_{real}(\boldsymbol{x})$. This randomness can notably occur when the physical output represented by $\boldsymbol{x} \to f_{real}(\boldsymbol{x})$ is actually $(\boldsymbol{x}, \boldsymbol{x}') \to f'_{real}(\boldsymbol{x}, \boldsymbol{x}')$, where $\boldsymbol{x}'$ is a vector constituted of other physical variables that are not taken into account in the representation $\boldsymbol{x} \to f_{real}(\boldsymbol{x})$. In this case, among different observations of $f_{real}(\boldsymbol{x})$, for the same experimental condition $\boldsymbol{x}$, the physical variables $\boldsymbol{x}'$ that are not taken into account vary, thus yielding different observed values. This variability can be modeled by a randomness when considering only the experimental conditions $\boldsymbol{x}$ in the representation of the physical system.*

In section 7.2, the physical system map $\boldsymbol{x} \to f_{real}(\boldsymbol{x})$ is modeled as random, while in section 7.3 and chapter 8, it is modeled as deterministic.

The components of the vector $\boldsymbol{x}$ of the experimental conditions can be divided into two categories. The first category contains the **control variables**. These variables define the physical system, independently of the environment in which the system is put. In engineering for instance, geometric parameters of the system can often be placed in this category, since they remain fixed regardless of what happens to the system. The second category contains the **environment variables**. These variables are the inputs of the physical system whose values are

not planned in the conception of the system. These variables are likely to be imposed beforehand by other systems. The distinction of the experimental conditions into these two categories is presented for instance in [SWN03] section 2.1. To give an illustration, in the system design phase, the environment variables are set by the future use of the system, while the control variables are the free parameters that may be set through an optimization phase.

The map f_{real} of the physical system can not be known for all the experimental conditions. Hence, this map is approximated by the computer model f_{mod}. The function f_{mod} shares the same input vector $\boldsymbol{x}$ as the physical system and provides the same scalar output. Furthermore, the function f_{mod} can have a second kind of inputs, denoted by the vector $\boldsymbol{\beta}$. The components of this vector are called the **model parameters**. They are the fitting parameters of the computer model f_{mod}. These parameters are unnecessary to carry out an experiment of the physical system, but they are needed to run the computer model. Hence, these quantities are seen as degrees of freedom for the computer model, and allow it to give a good approximation of the physical system. We will see that the term "good approximation" has two possible meanings. If the physical system is random, affecting a probability distribution to the model parameters enables to reproduce this randomness. If the physical system is a deterministic function, varying the model parameter gives different functions $\boldsymbol{x} \to f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$, thus giving more flexibility for the approximation of the physical system function.

Example 7.2. *Let us consider a toy example of a physical system and of an associated computer model (this toy example served as a pedagogic illustration for a training session on the CIRCE method [dC96]). The physical system is a tank, which can move forward and shoot a cannon ball. An experiment consists in making the tank move and shoot, and measuring the distance between the tank position, and the point at which the cannon ball hits the ground. Thus, the variable of interest is the distance P traveled by the cannon ball, and the two experimental conditions are the speed V of the tank, and the angle t between the cannon of the tank and the horizontal line. A schematic is provided in figure 7.1. The two experimental conditions V and t would rather be considered as environment variables, since they are likely to vary over the different utilizations of the tank.*

Now, consider a physical modeling of the shoot, in which a supplementary variable is introduced: the initial speed U of the cannon ball, relatively to the cannon, after the shoot (see figure 7.1). Notice that this speed does not need to be specified to carry out a real shoot. Thus, U is the model parameter. The computer model, in this toy model, neglects air friction, and consider gravitation as the only impacting force, with constant $g = 9.8m.s^{-2}$. Thus, the computer model is

$$(\boldsymbol{x}, \beta) = (V, t, U) \to P = \frac{1}{g}\left(U^2 \sin(2t) + 2UV \sin(t)\right). \tag{7.1}$$

The idea is that, if the speed of the cannon ball U is appropriately specified, (7.1) can provide a good prediction of the measured distance P for real shoots of the tank.

We have n observations of the physical system of the form $\boldsymbol{x}^{(1)}, y_{obs,1}, ..., \boldsymbol{x}^{(n)}, y_{obs,n}$, where $\boldsymbol{x}^{(i)}$ is an experimental condition and $y_{obs,i}$ is the observation of the physical system f_{real} obtained from it. The central question of this chapter 7 is to explain the discrepancies $y_{obs,i} - f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta})$.

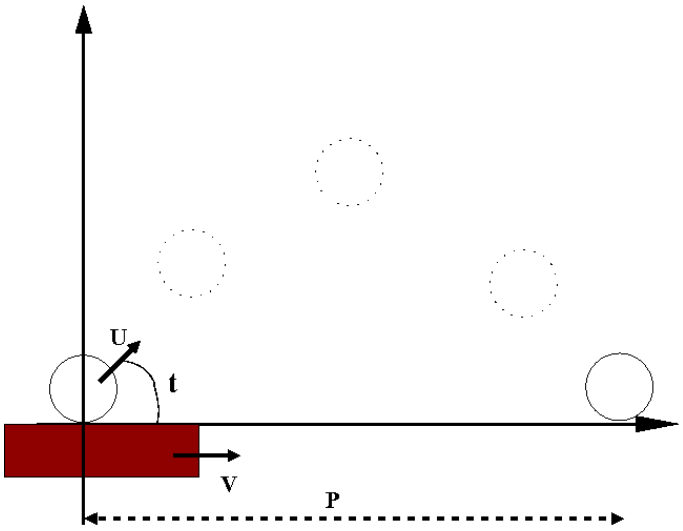

Figure 7.1: Toy example of the cannon ball. A tank can move forward and shoot a cannon ball. The variable of interest is the distance P between the point of shoot, and the point at which the cannon ball hits the ground. The experimental conditions are the speed V of the tank, and the angle t between the cannon of the tank and the horizontal line. The computer model is parameterized by the initial speed U of the cannon ball relatively to the cannon, and is defined by $P = \frac{1}{g}\left(U^2 \sin(2t) + 2UV \sin(t)\right)$, with g the earth gravitation constant.

The most classical and simple explanation is to consider that these discrepancies only come from a misspecification of $\boldsymbol{\beta}$ and measurement errors. More precisely, this corresponds to the case where two hypotheses are made. The first hypothesis is that the physical system f_{real} is a deterministic function and that the computer model is capable of perfectly reproducing it. That is to say, there is a model parameter $\boldsymbol{\beta}^{(0)}$ so that $\forall \boldsymbol{x}, f_{real}(\boldsymbol{x}) = f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}^{(0)})$. The second hypothesis is that the deviations $y_{obs,i} - f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}^{(0)})$ come from uncertainties related to the experiments. These uncertainties have generally two sources. First, the observations are affected by measurement errors. Second, although we do not treat this problem in this thesis, there can be a replicate uncertainty, meaning that the experimental conditions can not be known exactly for a given experiment.

The main limitation of this explanation is the assumption that there exists $\boldsymbol{\beta}^{(0)}$ so that the deviations $f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}^{(0)}) - y_{obs,i}$ come only from uncertainties related to the experiments. Indeed the order of magnitude of these uncertainties is usually known. Hence, when mean error indicators, such as $\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_{obs,i} - f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta})\right)^2$, are too large compared to this order of magnitude, it indicates that there is a problem with the two joint hypotheses discussed above (this can be quantified by Monte Carlo methods).

In this chapter 7, we will discuss two main frameworks to address deviations $y_{obs,i} - f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta})$ that are too large to be explained only by uncertainties related to the experiments.

In section 7.2, it is assumed that, for $1 \leq i \leq n$, $y_{obs,i} = f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}^{(i)}) + \epsilon_i$ where the $\boldsymbol{\beta}^{(i)}$ and the ϵ_i are *iid* and follow two given distributions. Therefore, the error terms $f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}) - y_{obs,i}$ are jointly explained by a variability of the physical system (which is random) and by

measurement errors.

In section 7.3, it is assumed that, for $1 \leq i \leq n$, $y_{obs,i} = f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}^{(0)}) + Z(\boldsymbol{x}^{(i)}) + \epsilon_i$, where $\boldsymbol{\beta}^{(0)}$ is a fixed model parameter, $\boldsymbol{x} \to Z(\boldsymbol{x})$ is a deterministic function and the ϵ_i are *iid* and follow a given distribution. Z is called the model error function. Hence, in section 7.3, it is assumed that the physical system is a deterministic function and that the error terms $f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}) - y_{obs,i}$ are jointly explained by a model error of the computer model f_{mod} and by measurement errors.

7.2 Errors modeled by a variability of the physical system

7.2.1 The general probabilistic model

Based on the general framework of section 7.1, the probabilistic model that we follow in this section 7.2 is the following one.

$$f_{real}(\boldsymbol{x}) = f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}), \tag{7.2}$$

where $\boldsymbol{x}$ is the vector of the experimental conditions and $\boldsymbol{\beta}$ is the random vector of the model parameters. The distribution $\mathcal{L}_{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, on $\mathbb{R}^m$, is unknown.

Remark 7.3. *Let us consider again the toy example 7.2. The model* (7.2) *corresponds to the case where doing two shoots with the same angle t and tank speed V results in two different distances P for the cannon ball. This difference is explained by an intrinsic variability of the shoot process in the cannon.* (7.2) *thus boils down to considering that this intrinsic variability yields a random initial speed U of the cannon ball after the shoot, and that the trajectory of the cannon ball is deterministic once t, V and U are fixed.*

In (7.2), the physical system is random, and the objective is to estimate its distributions $\boldsymbol{x} \to \mathcal{L}_{\boldsymbol{x}}$, where $\mathcal{L}_{\boldsymbol{x}}$ is the distribution of $f_{real}(\boldsymbol{x})$. Once this distribution is estimated, it can be used in any applied statistical analysis, such as, in a risk analysis, evaluating the probability that, for a given experimental condition $\boldsymbol{x}$, the physical system yield an undesirably large value. Furthermore, in a design study, knowing the distributions $\mathcal{L}_{\boldsymbol{x}}$ enables to carry out an optimization under uncertainty of the system parameter part of $\boldsymbol{x}$.

Since the computer model can be run for arbitrary inputs $\boldsymbol{x}, \boldsymbol{\beta}$, the distribution mapping $\boldsymbol{x} \to \mathcal{L}_{\boldsymbol{x}}$, can be estimated if the distribution $\mathcal{L}_{\boldsymbol{\beta}}$ of the model parameters is known. This is called uncertainty propagation ([dRDT08]), because the uncertainty on $\boldsymbol{\beta}$ is propagated in f_{mod} to yield the uncertainty on $f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$. Thus, the objective associated to the framework (7.2) is to estimate the distribution $\mathcal{L}_{\boldsymbol{\beta}}$.

This is done by using the computer model f_{mod} for chosen experimental conditions and model parameters, and by using a set of experimental results. A set of experimental results is of the form $\boldsymbol{x}^{(1)}, y_{obs,1}, ..., \boldsymbol{x}^{(n)}, y_{obs,n}$, where $\boldsymbol{x}^{(i)}$ is an experimental condition and $y_{obs,i} = f_{real}(\boldsymbol{x}^{(i)}) + \epsilon_i$ is an observation of the physical system for this experimental condition. It is assumed that the ϵ_i are independent measure errors, following a centered Gaussian distribution with known variance σ^2_{mes}. Thus, we have

$$y_{obs,i} = f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta}^{(i)}) + \epsilon_i \tag{7.3}$$

where the $y_{obs,i}$ are iid, the $\boldsymbol{x}^{(i)}$ are deterministic and observed, the ϵ_i are iid with known $\mathbf{N}(0, \sigma^2_{mes})$ distribution and the $\boldsymbol{\beta}^{(i)}$ are iid, unobserved and follow an unknown distribution $\mathcal{L}_{\boldsymbol{\beta}}$ which is to be estimated.

As in many distribution estimation problem, the question of whether using a parametric or non-parametric estimation method arises. Classically, roughly speaking, using a non-parametric method enables potentially to approximate the true distribution $\mathcal{L}_{\boldsymbol{\beta}}$ as accurately as possible. However, due to the limited number of observations, a non-parametric estimation method is much more subject to their variability than a parametric one, so that it can eventually yield a more imprecise estimation. The two errors that we have described, the approximation error and the prediction error, are hence antagonistic, so that a trade-off has to be found between them.

In this section 7.2, we will discuss the parametric estimation of $\mathcal{L}_{\boldsymbol{\beta}}$. The distribution $\mathcal{L}_{\boldsymbol{\beta}}$ is assumed to be multidimensional Gaussian, with unknown mean vector $\boldsymbol{m}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. Hence, the goal becomes to build estimators $\hat{\boldsymbol{m}}(\boldsymbol{y}_{obs}) = \hat{\boldsymbol{m}}(y_{obs,1}, ..., y_{obs,n})$ and $\hat{\boldsymbol{\Sigma}}(\boldsymbol{y}_{obs}) = \hat{\boldsymbol{\Sigma}}(y_{obs,1}, ..., y_{obs,n})$ for the mean vector and the covariance matrix.

7.2.2 Non-linear methods

By non-linear methods, we mean that we do not make any approximation for the computer model function $(\boldsymbol{x}, \boldsymbol{\beta}) \to f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$.

Maximum Likelihood methods

For given $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$, $y_{obs,i}$ in (7.3) always follows a non-degenerate distribution on $\mathbb{R}$ when $\sigma^2_{mes} > 0$. Thus, Maximum Likelihood is feasible. The probability function of $y_{obs,i}$ is written as, with $p_{\boldsymbol{m},\boldsymbol{\Sigma}}(.)$ being a probability density function given $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$,

$$\begin{aligned} & p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i}) \\ = & \int_{\mathbb{R}^m} p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i}, \boldsymbol{\beta}) d\boldsymbol{\beta} \\ = & \int_{\mathbb{R}^m} p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i}|\boldsymbol{\beta}) p_{\boldsymbol{m},\boldsymbol{\Sigma}}(\boldsymbol{\beta}) d\boldsymbol{\beta} \\ = & \frac{1}{(\sqrt{2\pi})^{m+1}\sigma_{mes}\sqrt{|\boldsymbol{\Sigma}|}} \\ & \int_{\mathbb{R}^m} \exp\left(-\frac{(y_{obs,i} - f_{mod}(\boldsymbol{x}_i, \boldsymbol{\beta}))^2}{2\sigma^2_{mes}}\right) \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\right) d\boldsymbol{\beta}. \end{aligned} \tag{7.4}$$

Thus, by writing the log-Likelihood of each of the independent $y_{obs,i}$, the Maximum Likelihood estimation of $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ is

$$(\hat{\boldsymbol{m}}_{ML}, \hat{\boldsymbol{\Sigma}}_{ML}) \in \underset{\boldsymbol{m},\boldsymbol{\Sigma}}{\operatorname{argmin}} L(\boldsymbol{m}, \boldsymbol{\Sigma}),$$

with

$$L(\boldsymbol{m}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} L_i(\boldsymbol{m}, \boldsymbol{\Sigma}), \tag{7.5}$$

with

$$L_i(\boldsymbol{m}, \boldsymbol{\Sigma}) = \tag{7.6}$$
$$\ln\left(\frac{1}{\sqrt{|\boldsymbol{\Sigma}|}}\int_{\mathbb{R}^m}\exp\left(-\frac{(y_{obs,i}-f_{mod}(\boldsymbol{x}_i,\boldsymbol{\beta}))^2}{2\sigma^2_{mes}}\right)\exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{m})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right)d\boldsymbol{\beta}\right)$$

Because of the very general form of f_{mod}, the integral terms in (7.5) and (7.6) are not explicit. If the computer model f_{mod} is cheap to run, they can be evaluated numerically, making the ML estimation computationally feasible. Let us also mention the existence of the Stochastic Expectation Maximization algorithm, that, roughly speaking, aims at optimizing (7.5) without calculating it exactly. We refer to chapter 1 of the PhD thesis [Fu12] and to [Bar10] on this subject.

If the computer model f_{mod} is expensive, one non-prohibitive solution, as mentioned in [Fu12], is to replace it by a cheaper metamodel $\hat{f}_{mod}$ before optimizing (7.5) and (7.6).

Bayesian methods

In the thesis [Fu12], Bayesian methods are preferred to the ML method (7.5), because of their ability to take into account expert knowledge, especially when the number n of observations is small.

A Bayesian model considers $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ as random vector and matrix, where the randomness corresponds to a lack of knowledge, and not to a variability. This randomness is hence different in nature from the randomness of $\boldsymbol{\beta}$, which really varies among the different observations $y_{obs,i}$. The a priori distribution of $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ is chosen by the practitioner, according to available expert knowledge, and is thus assumed fixed and known in all the mathematical developments. We refer to [Rob01] for an introduction to Bayesian statistics.

Treating $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ as random variables with known distribution makes it natural to consider the conditional distribution

$$p(\boldsymbol{m}, \boldsymbol{\Sigma}|\boldsymbol{y}_{obs}) \tag{7.7}$$

as gathering all the information relative to their estimation. For instance the mean of (7.7) can be considered as their estimation, and the variance of (7.7) can be considered as an indicator of the uncertainty of this estimation.

Considering the probability density function of a random symmetric matrix $\boldsymbol{\Sigma}$ is done by bijectively mapping S, the set of the $m \times m$ symmetric matrices, with $\mathbb{R}^{\frac{m(m+1)}{2}}$. The bijective mapping corresponds to extracting the $\frac{m(m+1)}{2}$ upper-diagonal coefficients of a symmetric matrix.

Once the probability density functions for symmetric matrices in S are defined, the conditional distribution (7.7) becomes,

$$\begin{aligned} p(\boldsymbol{m}, \boldsymbol{\Sigma}|\boldsymbol{y}_{obs}) &= \frac{1}{Z_{\boldsymbol{y}_{obs}}} p(\boldsymbol{y}_{obs}|\boldsymbol{m}, \boldsymbol{\Sigma}) p(\boldsymbol{m}, \boldsymbol{\Sigma}) \\ &= \frac{1}{Z_{\boldsymbol{y}_{obs}}} \left\{\prod_{i=1}^{n} p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i})\right\} p(\boldsymbol{m}, \boldsymbol{\Sigma}), \end{aligned} \tag{7.8}$$

with

$$\begin{aligned} Z_{\boldsymbol{y}_{obs}} &= \int_{\mathbb{R}^m \times S} p(\boldsymbol{y}_{obs}|\boldsymbol{m}, \boldsymbol{\Sigma}) p(\boldsymbol{m}, \boldsymbol{\Sigma}) d\boldsymbol{m} d\boldsymbol{\Sigma} \\ &= \int_{\mathbb{R}^m \times S} \{\prod_{i=1}^n p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i})\} p(\boldsymbol{m}, \boldsymbol{\Sigma}) d\boldsymbol{m} d\boldsymbol{\Sigma} \end{aligned}$$

and with $p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i})$ as in (7.4).

Remark 7.4. *In* (7.8)*, we have a classical Bayes' rule, that is summarized in equation (2.5) of [RW06] by the formulation*

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal} \quad \text{likelihood}},$$

where the likelihood term is the pdf of $\boldsymbol{y}_{obs}$ given $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$, the prior term is the unconditional pdf of $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$, the posterior term is the conditional pdf of $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ given $\boldsymbol{y}_{obs}$ and the marginal likelihood term is the intergral over $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ of the likelihood times the prior. Note that the marginal likelihood term is also the unconditional pdf of $\boldsymbol{y}_{obs}$.

Similarly to ML, the conditional pdf (7.8) of $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ depends on the non-explicit integrals (7.4) involving the computer model f_{mod}. If the function f_{mod} is cheap, Monte Carlo Markov Chains (MCMC) algorithms can compute a sample $\boldsymbol{m}^{(1)}, \boldsymbol{\Sigma}^{(1)}, ..., \boldsymbol{m}^{(mc)}, \boldsymbol{\Sigma}^{(mc)}$ whose empirical distribution is approximately the conditional distribution (7.8). We refer to [RC99] for a general introduction to MCMC algorithms and to the chapter 1 of [Fu12] for their utilization in the present context.

Now, if the computer model f_{mod} is time-costly, the approach followed in [Fu12] is to combine MCMC algorithms with a Kriging approximation of f_{mod}. We refer to [Fu12] for further details on this approach.

7.2.3 Methods based on a linearization of the computer model

The advantage of linearization-based methods is to make explicit the terms $p_{\boldsymbol{m},\boldsymbol{\Sigma}}(y_{obs,i})$ of (7.4), which are interpreted as likelihood functions or as conditional distributions, depending on whether we are in the frequentist of Bayesian framework.

The computer model is thus approximated linearly with respect to its model parameters (within the range of values that is under consideration). Hence the computer model is considered of the form $f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}) = f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}_{nom}) + \sum_{i=1}^m h_i(\boldsymbol{x})(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{nom,i})$ where $\boldsymbol{\beta}_{nom}$ is the nominal vector around which the linear approximation is made. We choose, for simplicity reasons, to remove the perfectly known quantities $\boldsymbol{\beta}_{nom}$ and $f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}_{nom})$. Indeed, up to a shift with respect to $\boldsymbol{\beta}$ and f_{mod}, we can consider that $\boldsymbol{\beta}_{nom} = 0$ and $f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}_{nom}) = 0$. We then have

$$\forall \boldsymbol{x} : \ f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}) = \sum_{i=1}^m h_i(\boldsymbol{x}) \boldsymbol{\beta}_i. \tag{7.9}$$

The linear approximation is justified by a Taylor series expansion when the true covariance matrix $\boldsymbol{\Sigma}$ of the model parameter $\boldsymbol{\beta}$ is small. If it is not the case, as stated in chapter 1 of [Fu12], the estimation of $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ can be misleading.

Nevertheless, the linear approximation simplifies the treatment, because, denoting $\boldsymbol{h}(\boldsymbol{x}^{(i)}) = (h_1(\boldsymbol{x}^{(i)}), ..., h_m(\boldsymbol{x}^{(i)}))^t$, we have

$$y_{obs,i} = \boldsymbol{h}(\boldsymbol{x}^{(i)})^t \boldsymbol{\beta} + \epsilon_i,$$

so that

$$\mathcal{L}(y_{obs,i}|\boldsymbol{m},\boldsymbol{\Sigma}) = \mathcal{N}\left(\boldsymbol{h}(\boldsymbol{x}^{(i)})^t\boldsymbol{m}, V_i(\boldsymbol{\Sigma})\right), \tag{7.10}$$

with

$$V_i(\boldsymbol{\Sigma}) = \boldsymbol{h}(\boldsymbol{x}^{(i)})^t\boldsymbol{\Sigma}\boldsymbol{h}(\boldsymbol{x}^{(i)}) + \sigma^2_{mes}. \tag{7.11}$$

Thus, the ML estimator of subsection 7.2.2 becomes

$$(\hat{\boldsymbol{m}}_{ML}, \hat{\boldsymbol{\Sigma}}_{ML}) \in \underset{\boldsymbol{m},\boldsymbol{\Sigma}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ln\left(V_i(\boldsymbol{\Sigma})\right) + \frac{(y_{obs,i} - \boldsymbol{h}(\boldsymbol{x}^{(i)})^t\boldsymbol{m})^2}{V_i(\boldsymbol{\Sigma})}, \tag{7.12}$$

with $V_i(\boldsymbol{\Sigma})$ as in (7.11). The Likelihood function requires to compute the gradients of f_{mod} w.r.t $\boldsymbol{\beta}$ for all the $\boldsymbol{x}^{(i)}$. This is done prior to the optimization in (7.12), so that, naturally, the linearization-based ML involves fewer evaluations of f_{mod} than the non-linear ML of subsection 7.2.2. When the gradients are calculated, (7.12) could be numerically optimized directly, since the likelihood criterion is evaluated with $O(n)$ operations. In [dC96], an Expectation-Maximization algorithm is proposed for optimizing (7.12). The obtained method for estimating $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ is named the "CIRCE" method. It has been widely used in the system thermal-hydraulic domain, and especially with the CATHARE computer model.

Finally, notice that, though we have presented the case where the measure error variance σ^2_{mes} is known, this parameter can be estimated as well by ML, thus yielding an optimization problem similar to (7.12). In the CIRCE method, the parameter σ^2_{mes} can similarly be estimated.

In the Bayesian framework, the conditional distribution (7.8) of the non-linear case becomes

$$p(\boldsymbol{m},\boldsymbol{\Sigma}|\boldsymbol{y}_{obs}) = \tag{7.13}$$
$$\frac{1}{Z_{\boldsymbol{y}_{obs}}}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi V_i(\boldsymbol{\Sigma})}}\exp\left(\frac{-\left[y_{obs,i} - \boldsymbol{h}(\boldsymbol{x}^{(i)})^t\boldsymbol{m}\right]^2}{2V_i(\boldsymbol{\Sigma})}\right)\right\} p(\boldsymbol{m},\boldsymbol{\Sigma}),$$

where

$$Z_{\boldsymbol{y}_{obs}} =$$
$$\int_{\mathbb{R}^m\times S}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi V_i(\boldsymbol{\Sigma})}}\exp\left(\frac{-\left[y_{obs,i} - \boldsymbol{h}(\boldsymbol{x}^{(i)})^t\boldsymbol{m}\right]^2}{2V_i(\boldsymbol{\Sigma})}\right)\right\} p(\boldsymbol{m},\boldsymbol{\Sigma})d\boldsymbol{m}d\boldsymbol{\Sigma},$$

and with $V_i(\boldsymbol{\Sigma})$ as in (7.11).

Similarly to (7.12), (7.13) requires to compute the gradients of f_{mod} once and for all, before the conditional distribution is computed. Thus, the conditional distribution is rather cheap to compute, since the terms in the integrals in (7.13) can be computed with $O(n)$ operations. Hence, classical integral evaluation methods, or classical MCMC algorithms are likely to work well.

7.3 Errors modeled by a model error process

7.3.1 The general probabilistic model

The statistical model of this section 7.3 is based on two main ideas:

- The physical system $\boldsymbol{x} \to f_{real}(\boldsymbol{x})$ does not necessarily belong to the set of computer model functions $\{\boldsymbol{x} \to f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})\}$. We model the difference between the physical system and the correctly parameterized computer model by an error function that is called the **model error**. The notion of correctly parameterized computer model is explained below.

- The model error function is not observable everywhere, and hence is unknown for the majority of the experimental conditions. This lack of knowledge is modeled by the introduction of a stochastic framework for this function, that is to say, it is represented by a realization of a Gaussian process $Z(\boldsymbol{x})$. Thus, as we have discussed in subsection 2.1.1, we model an unknown deterministic function as a realization of a Gaussian process. Being the sum of the correctly parameterized computer model and of the model error function, the physical system itself is a realization of a Gaussian process. Hence, we do not use the notation $f_{real}(\boldsymbol{x})$ anymore for the physical system. Instead, we denote it by the Gaussian process $Y_{real}(\boldsymbol{x})$.

Motivated by these two ideas, the Gaussian process statistical model of section 7.3 is defined by the two following equations

$$Y_{real}(\boldsymbol{x}) = f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}) + Z(\boldsymbol{x}) \tag{7.14}$$

and

$$Y_{obs}(\boldsymbol{x}) = Y_{real}(\boldsymbol{x}) + \epsilon(\boldsymbol{x}). \tag{7.15}$$

Where:

- $Y_{real}(\boldsymbol{x})$ is the Gaussian process of the physical system.

- $Z(\boldsymbol{x})$ is the model error process. It is assumed to be Gaussian and centered. Its covariance function is generally not considered known, and will be estimated from data, similarly to chapter 3.

- $\boldsymbol{\beta}$ is the correct parameter of the computer model. We call it the correct parameter because, Z being centered, the computer model parameterized by $\boldsymbol{\beta}$ is the mean function of the physical system.

- $Y_{obs}(\boldsymbol{x})$ is the observed output of the physical system for the experimental conditions $\boldsymbol{x}$. This observation is the sum of the variable of interest and of a measurement error $\epsilon(\boldsymbol{x})$. $\epsilon(\boldsymbol{x})$ follows a Gaussian centered distribution, and is independent from one experiment to another. The variance of ϵ is in general constant, in which case we denote it by σ^2_{mes}, and can be assumed known.

Remark 7.5. *Let us consider the toy example 7.2. Assume that the computer model remains the analytical function*

$$(\boldsymbol{x}, \beta) = (V, t, U) \to P = \frac{1}{g}\left(U^2 \sin(2t) + 2UV \sin(t)\right),$$

which assumes that the only acting force is the gravitation. Assume that the air friction force is actually non-negligible, and is of the form $-\gamma \overrightarrow{v}$ for a speed vector $\overrightarrow{v}$. Assume also that a cannon

shoot always yields the same initial speed U_0 of the cannon ball. Then, the physical system is still a deterministic function of V and t, which can correspond to the realization of a Gaussian process. U_0 can be interpreted as the correct model parameter, but the model obtained from it is not a perfect representation of the physical system. The air friction coefficient γ, that is not taken into account by the computer model, causes a model error function, that, by difference, is a deterministic function. This deterministic function can be modeled as the realization of a Gaussian process.

For the model error Z in (7.14), we recommend to use a continuous covariance function, such as the ones presented in chapter 2. These covariance functions make $Z(\boldsymbol{x})$ and $Z(\boldsymbol{x}+\boldsymbol{\delta_x})$ dependent for small $\boldsymbol{\delta_x}$. There are two reasons for doing so.

- The physical system is generally continuous with respect to the experimental conditions, and so is the computer model. Hence, as a difference, the model error process Z must be a process with continuous trajectories. This is generally the case when the covariance function is continuous (see chapter 2).
- Similarly, it is expected that if the computer model makes a certain error for a given experimental point, then it will do a similar error for a nearby experimental point. This principle is taken into account by a continuous covariance function.

Concerning the correct model parameter $\boldsymbol{\beta}$, we consider both a frequentist or Bayesian framework. The Bayesian framework allows to take into account expert judgments for the model parameter $\boldsymbol{\beta}$. This is done by modeling the constant but unknown correct model parameter $\boldsymbol{\beta}$ as a random vector. The distribution of this random vector is known, and chosen according to the degree of knowledge one has about the model parameter $\boldsymbol{\beta}$. We use a Gaussian distribution for the Bayesian modeling of $\boldsymbol{\beta}$. Hence, we distinguish two cases:

No prior information case: $\boldsymbol{\beta}$ is a vector of unknown constants.

Prior information case: $\boldsymbol{\beta}$ is a Gaussian vector, with known mean vector $\boldsymbol{\beta}_{prior}$ and covariance matrix $\mathbf{Q}_{prior}$.

From the Gaussian process modeling (7.14) and (7.15), we are interested in solving the two following problems:

1. **Calibration.** It is the problem of estimating the correct model parameter $\boldsymbol{\beta}$, or equivalently finding the most accurate computer model function $\boldsymbol{x} \to f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$ to represent the physical system.
2. **Prediction.** For a new experimental condition $\boldsymbol{x}^{(new)}$, we want to predict the physical system, and associate a measure of uncertainty to this prediction. The main idea is that the physical system is not predicted solely by the calibrated computer model, because we are also able to infer the value of the model error at $\boldsymbol{x}^{(new)}$.

The calibration and prediction methods depend on the approximations made on the computer model function. In subsection 7.3.2, no approximations are made. In subsection 7.3.3, a linear approximation of the computer model, with respect to the model parameters, is made.

7.3.2 Non-linear methods

Estimation of the covariance function of the model error

For the calibration and prediction tasks to be carried out, it is first necessary to estimate the covariance function of the model error Z and the variance σ^2_{mes} of the measurement error ϵ.

The variance σ^2_{mes} of the measurement errors can generally be specified from physical expertise. This is the case we will consider here. If it is not the case, this function can, for example, be estimated in the same way as the model error covariance function.

We denote by K_{mod} the covariance function of the model error. Generally there is no expert judgment available concerning K_{mod}. Indeed, physical knowledge is used in the conception of the computer model, and hence may not help to know the shape of the error of the computer model. Therefore, K_{mod} has to be selected in the parametric set similar to definition 3.18,

$$\mathcal{K}_{mod} = \{K_{mod,\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\},$$

with $K_{mod,\boldsymbol{\psi}}$ a stationary covariance function and $\Psi \subset \mathbb{R}^p$. $\mathcal{K}_{mod}$ can be one of the classical covariance function families that are presented in chapter 3.

We have n observations of the physical system of the form $\boldsymbol{x}^{(1)}, y_{obs,1}, ..., \boldsymbol{x}^{(n)}, y_{obs,n}$, where $\boldsymbol{x}^{(i)}$ is an experimental condition and $y_{obs,i} = f_{obs}(\boldsymbol{x}^{(i)})$. In the frequentist framework for $\boldsymbol{\beta}$, the likelihood function of $\boldsymbol{y}_{obs}$ is a function of both $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ and is as follows.

$$l(\boldsymbol{\beta}, \boldsymbol{\psi}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{|\mathbf{K}_{\boldsymbol{\psi}}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})})^t \mathbf{K}_{\boldsymbol{\psi}}^{-1} (\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})})\right), \tag{7.16}$$

where $\boldsymbol{m}^{(\boldsymbol{\beta})}$ is the $n \times 1$ vector defined by $m_i^{(\boldsymbol{\beta})} = f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta})$ and $\mathbf{K}_{\boldsymbol{\psi}} := \mathbf{K}_{mod,\boldsymbol{\psi}} + \sigma^2_{mes}\mathbf{I}_n$, with $\mathbf{K}_{mod,\boldsymbol{\psi}}$ the $n \times n$ matrix defined by $(\mathbf{K}_{mod,\boldsymbol{\psi}})_{i,j} = K_{mod,\boldsymbol{\psi}}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})$.

Because of the general form of the computer model function f_{mod} in (7.16), the likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\psi})$ must generally be maximized jointly with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. This yields the ML estimator $\hat{\boldsymbol{\psi}}_{ML}$.

Similarly to section 7.2, two cases are considered, depending if the computer model is expensive to run or not. If the computer model is not expensive to run, (7.16) can be directly maximized numerically. If the computer model is expensive to run, one possible solution is to build a Kriging model of it, jointly with respect to $\boldsymbol{x}$ and $\boldsymbol{\beta}$, from a limited number n_m of computer model results. Also, new computer model results could be added iteratively, with well-chosen inputs $\boldsymbol{x}, \boldsymbol{\beta}$, in a spirit similar to Kriging-based optimization [JSW98]. To our knowledge, this problem has not been much addressed in the literature. Indeed, classical references on computer model calibration ([KO01], [HKC+04], [BBP+07]), when a model error is taken into account, rather consider the Bayesian framework for $\boldsymbol{\beta}$.

In the Bayesian framework for $\boldsymbol{\beta}$, let $p(\boldsymbol{\beta})$ be the probability density function of $\boldsymbol{\beta}$, following a $\mathcal{N}(\boldsymbol{\beta}_{prior}, \mathbf{Q}_{prior})$ distribution. We consider the fully-Bayesian case where an *a priori* probability distribution $p(\boldsymbol{\psi})$ is also specified for $\boldsymbol{\psi}$. Indeed, this is the case in the references [KO01], [HKC+04], [BBP+07]. Then, the conditional distribution of $\boldsymbol{\beta}, \boldsymbol{\psi}$ given $\boldsymbol{y}_{obs}$ has the pdf

$$\frac{l(\boldsymbol{\beta}, \boldsymbol{\psi}) p(\boldsymbol{\beta}) p(\boldsymbol{\psi})}{\int_{\mathbb{R}^{m+p}} l(\boldsymbol{\beta}, \boldsymbol{\psi}) p(\boldsymbol{\beta}) p(\boldsymbol{\psi}) d\boldsymbol{\beta} d\boldsymbol{\psi}}, \tag{7.17}$$

with $l(\boldsymbol{\beta}, \boldsymbol{\psi})$ as in (7.16). Similarly to the frequentist case, if the computer model f_{mod} is cheap, then the numerator in (7.17) is rather cheap to compute, so that MCMC methods can be carried out directly. These MCMC methods yield a sample $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\psi}^{(1)}, ..., \boldsymbol{\beta}^{(mc)}, \boldsymbol{\psi}^{(mc)})$ following approximately the conditional distribution (7.17). The empirical mean of the $\boldsymbol{\psi}^{(i)}$ is thus the fully-Bayesian estimation of $\boldsymbol{\psi}$.

Now, if the computer model is expensive to run, the aforementioned references [KO01], [HKC+04], [BBP+07] propose to build a Kriging model for it. A Bayesian probability distribution is also associated to the covariance hyper-parameters of this second Kriging model. This results in a rather complex fully Bayesian framework, for which, by using MCMC methods, it is nevertheless tractable to compute an approximation of the conditional distribution of $\boldsymbol{\psi}$ conditionally to the vector of experimental results $\boldsymbol{y}_{obs}$, and to a vector of results of the computer model f_{mod}. Again, the mean of this approximate conditional distribution constitutes the Bayesian estimation of $\boldsymbol{\psi}$. We refer to the aforementioned references for details of this treatment.

In the Bayesian framework for $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, whether the computer model is expensive to run or not, the MCMC method yields a sample $\boldsymbol{\psi}^{(1)}, ..., \boldsymbol{\psi}^{(mc)}$ from the distribution (7.17) for $\boldsymbol{\psi}$. This sample naturally enables us to quantify the uncertainty related to the estimation of $\boldsymbol{\psi}$. This uncertainty can also naturally be taken into account in the prediction and calibration (see [KO01], [HKC+04], [BBP+07]).

Nevertheless, in section 5.2.1 of [BBP+07] and in section 4.5 of [KO01], it is recommended to consider $\boldsymbol{\psi}$ as known and equal to its Bayesian estimate. We will follow this recommendation in this subsection 7.3.2. Hence, in the sequel of subsection 7.3.2, $\boldsymbol{\psi}$ is considered known and equal to its estimate. Notice that this is similar to chapter 3 where, most classically, the covariance hyper-parameters of a Kriging model are first estimated, and then assumed known and equal to their estimate when addressing Kriging predictions and predictive variances. Notice that this may result in a slight underestimation of the uncertainty associated to the Kriging predictions.

Calibration when the covariance function of the model error is fixed

As we have previously discussed, we consider that the covariance function K_{mod} of the model error process Z is known.

In the frequentist framework for $\boldsymbol{\beta}$, the ML estimation of $\boldsymbol{\beta}$ corresponds to maximizing the likelihood function

$$l(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sqrt{|\mathbf{K}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})})^t \mathbf{K}^{-1} (\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})})\right), \tag{7.18}$$

where $\boldsymbol{m}^{(\boldsymbol{\beta})}$ is the $n \times 1$ vector defined by $m_i^{(\boldsymbol{\beta})} = f_{mod}(\boldsymbol{x}^{(i)}, \boldsymbol{\beta})$ and $\mathbf{K} = \mathbf{K}_{mod} + \sigma^2_{mes} \mathbf{I}_n$, with $\mathbf{K}_{mod}$ the $n \times n$ matrix defined by $(\mathbf{K}_{mod})_{i,j} = K_{mod}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})$. Maximizing (7.18) yields the ML estimator $\hat{\boldsymbol{\beta}}_{ML}$.

Remark 7.6. *We see in* (7.18) *that the ML estimator* $\hat{\boldsymbol{\beta}}_{ML}$ *can be written*

$$\hat{\boldsymbol{\beta}}_{ML} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})})^t \mathbf{K}^{-1} (\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})}).$$

Therefore, the ML estimator of $\boldsymbol{\beta}$ *selects a model parameter yielding the best possible reproduction of the experimental results by the computer model, which is intuitive.*

Remark 7.7. *Quantifying the uncertainty of the ML estimator $\hat{\boldsymbol{\beta}}_{ML}$ in (7.18) is not simple because of the general nature of the computer model function f_{mod}. On the contrary, in the Bayesian framework treated below for calibration, the conditional distribution of $\boldsymbol{\beta}$ naturally yields this uncertainty. In this subsection 7.3.2, we will not discuss the uncertainty related to the frequentist ML estimation of $\boldsymbol{\beta}$, because we do not use this ML estimation elsewhere in the manuscript (when the computer model is not assumed linear with respect to $\boldsymbol{\beta}$). We are not aware either of references on the frequentist estimation of $\boldsymbol{\beta}$ in the non linear case. On the contrary, the references [KO01], [HKC+04] and [BBP+07] treat the Bayesian estimation of $\boldsymbol{\beta}$.*

We refer to the discussion following (7.16) for the optimization of (7.18): if the computer model is cheap, the optimization can be carried out directly, if not, building a Kriging model of the computer model is a possibility.

We now consider the Bayesian estimation of $\boldsymbol{\beta}$, for which details can be found in the references [KO01], [HKC+04] and [BBP+07]. Let $p(\boldsymbol{\beta})$ be the probability density function of $\boldsymbol{\beta}$ following a $\mathcal{N}(\boldsymbol{\beta}_{prior}, \mathbf{Q}_{prior})$ distribution. Then the distribution of $\boldsymbol{\beta}$, conditionally to the vector of observation $\boldsymbol{y}_{obs}$ is

$$\frac{l(\boldsymbol{\beta})p(\boldsymbol{\beta})}{\int_{\mathbb{R}^m} l(\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}}, \tag{7.19}$$

with $l(\boldsymbol{\beta})$ as in (7.18). We refer to the discussion following (7.17) for the computation of (7.19). If the computer model is cheap, MCMC methods can be carried out directly. Else, it is proposed in [KO01], [HKC+04] and [BBP+07] to combine MCMC methods with a Kriging model for the computer model. In both cases, MCMC methods yield a sample $(\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(mc)})$ following approximately the conditional distribution (7.19). All the conditional distribution can be considered for carrying out further uncertainty analysis on $\boldsymbol{\beta}$. If more simple indicators are preferable, the empirical mean of the $\boldsymbol{\psi}^{(i)}$ is the fully-Bayesian estimation of $\boldsymbol{\psi}$, and an associated indicator of the estimation error is the empirical variance of the $\boldsymbol{\psi}^{(i)}$.

Prediction when the covariance function of the model error is fixed

The goal of the prediction is to give the most probable value of the physical system, for a new experimental condition, without doing a real experiment. This most probable value is not necessarily given by the output of the calibrated computer model, because the model error is inferred as well.

Consider a new point $\boldsymbol{x}^{(new)}$, for which we aim at predicting the value of the physical system $Y_{real}(\boldsymbol{x}^{(new)})$, from the observed values at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. We denote by $\boldsymbol{k}(\boldsymbol{x}^{(new)})$ the $n \times 1$ covariance vector of the model error process Z between $(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)})$ and $\boldsymbol{x}^{(new)}$, that is $\left(\boldsymbol{k}(\boldsymbol{x}^{(new)})\right)_i := K_{mod}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(new)})$.

Consider first the frequentist case, with the ML estimator $\hat{\boldsymbol{\beta}}_{ML}$ of (7.18). The most natural prediction method is to consider $\boldsymbol{\beta}$ fixed and equal to the estimate $\hat{\boldsymbol{\beta}}_{ML}$. This being done, the observation vector $y_{obs,1}, ..., y_{obs,n}$ has distribution $\mathcal{N}(\boldsymbol{m}^{(\hat{\boldsymbol{\beta}}_{ML})}, \mathbf{K})$. Furthermore, since the measure error is independent of the model error, the covariance vector between $\boldsymbol{y}_{obs}$ and $Y_{real}(\boldsymbol{x}^{(new)})$ is $\boldsymbol{k}(\boldsymbol{x}^{(new)})$. Thus, we can directly use the simple Kriging equation (2.9), which yields the prediction

$$\hat{y}(\boldsymbol{x}^{(new)}) = f_{mod}(\boldsymbol{x}^{(new)}, \hat{\boldsymbol{\beta}}_{ML}) + \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\hat{\boldsymbol{\beta}}_{ML})}). \tag{7.20}$$

The main idea in (7.20) is that $Y_{real}(\boldsymbol{x}^{(new)})$ is not predicted by the calibrated computer model only, because we are able to infer the value of the model error at $\boldsymbol{x}^{(new)}$. Thus, the predictor (7.20) is composed of the calibrated computer model $f_{mod}(\boldsymbol{x}^{(new)}, \hat{\boldsymbol{\beta}}_{ML})$ and of the inferred model error $\boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\hat{\boldsymbol{\beta}}_{ML})})$. By inspection of (7.20), the inferred model error has the following properties:

- $\boldsymbol{x}^{(new)}$ being fixed, this term is large when the errors $\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\hat{\boldsymbol{\beta}}_{ML})}$ between the experimental results and the calibrated computer model are large.
- The observations being fixed, this term is a linear combination of the components of $\boldsymbol{k}(\boldsymbol{x}^{(new)})$. These elements are usually a decreasing function of the distance between $\boldsymbol{x}^{(new)}$ and the experimental conditions $\boldsymbol{x}^{(i)}$. Hence, if $\boldsymbol{x}^{(new)}$ is far from an experimental condition $\boldsymbol{x}^{(i)}$, then the weight of this experimental result is small in the combination. Hence, the prediction of $Y_{real}(\boldsymbol{x}^{(new)})$ is almost only composed of the calibrated computer model when $\boldsymbol{x}^{(new)}$ is far from any available experimental condition, while the model error inference term is significant when $\boldsymbol{x}^{(new)}$ is in the neighborhood of an available experimental condition (the neighborhood is defined in terms of the correlation function K_{mod}).

Neglecting the uncertainty related to $\hat{\boldsymbol{\beta}}_{ML}$, the prediction mean square error of $\hat{y}(\boldsymbol{x}^{(new)})$ is obtained from (2.10) and is

$$\hat{\sigma}^2(\boldsymbol{x}^{(new)}) = K_{mod}(\boldsymbol{x}^{(new)}, \boldsymbol{x}^{(new)}) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1} \boldsymbol{k}(\boldsymbol{x}^{(new)}).$$

Similarly to remark 7.7, we do not discuss methods for taking the uncertainty related to $\hat{\boldsymbol{\beta}}_{ML}$ into account in the prediction mean square error. Indeed, the prediction in the frequentist case (in the non-linear case of this subsection 7.3.2) is not treated in this manuscript, neither is it (to our knowledge) in the literature.

We now consider the Bayesian case for $\boldsymbol{\beta}$. The Bayesian framework computes, by nature, the conditional distribution of $Y_{real}(\boldsymbol{x}^{(new)})$ given $\boldsymbol{y}_{obs}$, which also takes the uncertainty related to $\boldsymbol{\beta}$ into account in the prediction error. Letting $p(.)$ and $p(.|.)$ denote probability density functions and conditional probability density functions, we have

$$p(Y_{real}(\boldsymbol{x}^{(new)})|\boldsymbol{y}_{obs}) = \int_{\mathbb{R}^m} p(Y_{real}(\boldsymbol{x}^{(new)})|\boldsymbol{y}_{obs}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{y}_{obs}) d\boldsymbol{\beta}. \tag{7.21}$$

$Y_{real}(\boldsymbol{x}^{(new)})$ can be sampled easily conditionally to $(\boldsymbol{y}_{obs}, \boldsymbol{\beta})$, because it follows a Gaussian distribution with mean

$$f_{mod}(\boldsymbol{x}^{(new)}, \boldsymbol{\beta}) + \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{(\boldsymbol{\beta})}) \tag{7.22}$$

and variance

$$K_{mod}(\boldsymbol{x}^{(new)}, \boldsymbol{x}^{(new)}) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1} \boldsymbol{k}(\boldsymbol{x}^{(new)}). \tag{7.23}$$

Also, a sample $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(mc)}$, following approximately the distribution of $\boldsymbol{\beta}$ conditionally to $\boldsymbol{y}_{obs}$, can be obtained by using MCMC techniques, as we have discussed in (7.19). Thus, sampling $(Y_{real}(\boldsymbol{x}^{(new)}))_i$ conditionally to $\boldsymbol{y}_{obs}, \boldsymbol{\beta}^{(i)}$ for $1 \leq i \leq mc$, we obtain the sample $(Y_{real}(\boldsymbol{x}^{(new)}))_1, ..., (Y_{real}(\boldsymbol{x}^{(new)}))_{mc}$ following approximately the full-Bayesian distribution

(7.21). Notice that (7.21) is similar to the first equation of appendix C in [BBP+07], and that the sampling method we discuss is similar to the sampling method discussed there.

Notice also that the sampling method for (7.21) is intuitive because $(Y_{real}(\boldsymbol{x}^{(new)}))_i$ is the sum of $f_{mod}(\boldsymbol{x}^{(new)}, \boldsymbol{\beta}^{(i)})$ and of a conditional realization of the model error (see (7.22) and (7.23)). A conditional realization of $Y_{real}(\boldsymbol{x}^{(new)}))$ is thus composed of a conditional realization of the calibrated code and of a conditional realization of the model error.

The sampling method following (7.21) has the advantage of separating the calibration part and the prediction part. Indeed, when the sample $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(mc)}$ is obtained, it can be stored and used, afterwards, for sampling $Y_{real}(\boldsymbol{x}^{(new)})$ for a large number of experimental conditions $\boldsymbol{x}^{(new)}$. This second sample is done exactly and does not require MCMC methods. Furthermore, the full process $\boldsymbol{x}^{(new)} \rightarrow Y_{real}(\boldsymbol{x}^{(new)})$ can also be sampled, from the same MCMC sample $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(mc)}$, based on subsection 2.2.3.

The question on whether the computer model is cheap to run or not impacts the generation of $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(mc)}$, as we have discussed after (7.19). Predicting $Y_{real}(\boldsymbol{x}^{(new)})$ for a large number of experimental conditions $\boldsymbol{x}^{(new)}$ is also impracticable directly if f_{mod} is expensive to run, because f_{mod} has to be called mc times for each experimental condition $\boldsymbol{x}^{(new)}$. We refer to [KO01], [HKC+04] and [BBP+07] for methods based on a Kriging modeling of f_{mod} to overcome this issue.

7.3.3 Methods based on a linearization of the computer model

Linearization of the computer model

The methods described in subsection 7.3.2 are valid for any computer model function f_{mod}. However, they can be rather computationally expensive to use in practice. Indeed, we have seen that these methods require, in the Bayesian case, to run a MCMC algorithm, and possibly to approximate the computer model by a surrogate model in both the $\boldsymbol{x}$ and $\boldsymbol{\beta}$ domains. In the frequentist case, the methods of subsection 7.3.2 require to numerically solve an optimization problem, involving f_{mod}, with respect to $\boldsymbol{\beta}$, and we have seen that they do not provide natural way to take the uncertainty related to $\boldsymbol{\beta}$ into account in further predictions.

In this subsection 7.3.3, we show that the treatment of the Gaussian process modeling of the model error is much simpler when the computer model is assumed linear with respect to its model parameter $\boldsymbol{\beta}$ (within the range of values that is under consideration).

Hence, in this subsection 7.3.3, and similarly to subsection 7.2.3, we consider computer models of the form $f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}) = f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}_{nom}) + \sum_{i=1}^{m} h_i(\boldsymbol{x})(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{nom,i})$ where $\boldsymbol{\beta}_{nom}$ is the nominal vector around which the linear approximation is made. $\boldsymbol{\beta}_{nom}$ is generally chosen by expert judgment or by previous calibration studies. Similarly to (7.9), we can equivalently have the simpler equation

$$\forall \boldsymbol{x} : \; f_{mod}(\boldsymbol{x}, \boldsymbol{\beta}) = \sum_{i=1}^{m} h_i(\boldsymbol{x})\boldsymbol{\beta}_i. \tag{7.24}$$

The linear approximation is justified by a Taylor series expansion when the uncertainty concerning the correct parameter $\boldsymbol{\beta}$ is small. This linear approximation is frequently made, for example in thermal-hydraulics [dC01, PCD08], or in neutron transport [KHF+06]. A thorough

discussion on the validity of using the linear approximation in the non-linear case is given at the end of this subsection 7.3.3.

We now formulate the problem in vector-matrix form. Assume that n experiments are carried out at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. We denote the $n \times m$ matrix $\mathbf{H}$ of partial derivatives of the computer model with respect to $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)$. $\mathbf{H}$ is defined by

$$H_{i,j} = h_j(\boldsymbol{x}^{(i)}).$$

For the n experiments the equations (7.14) and (7.15) become

$$\boldsymbol{y}_{obs} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{z} + \boldsymbol{\epsilon}, \tag{7.25}$$

with $z_i = Z(\boldsymbol{x}^{(i)})$ and $\epsilon_i = \epsilon(\boldsymbol{x}^{(i)})$. Hence we have a universal Kriging model (see chapter 2). We denote by $\mathbf{K}$ the covariance matrix of the model and measurement error vector $\boldsymbol{z} + \boldsymbol{\epsilon}$.

$$\mathbf{K} := cov(\boldsymbol{z} + \boldsymbol{\epsilon}) = \mathbf{K}_{mod} + \sigma^2_{mes}\mathbf{I}_n. \tag{7.26}$$

Remark 7.8. *In* (7.26)*, the covariance matrix of the measure error vector* $\boldsymbol{\epsilon}$ *is* $\sigma^2_{mes}\mathbf{I}_n$*, because the measure errors are independent. The case of dependent Gaussian measure errors can be treated similarly, by replacing, in* (7.26)*,* $\sigma^2_{mes}\mathbf{I}_n$ *by the covariance matrix* $\mathbf{K}_{mes}$ *of the measure error vector* $\boldsymbol{\epsilon}$*. In this subsection 7.3.2, for concision, we address the case of iid measure errors with variance* σ^2_{mes}*.*

When the matrix $\mathbf{K}$ is fixed, we can compute the *a priori* distribution of the vector of observations. In the no prior information case we have, with $\boldsymbol{\beta}$ an unknown constant,

$$\boldsymbol{y}_{obs} \sim \mathcal{N}(\mathbf{H}\boldsymbol{\beta}, \mathbf{K}). \tag{7.27}$$

In the prior information case, we have, with $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_{prior}, \mathbf{Q}_{prior})$

$$\boldsymbol{y}_{obs} \sim \mathcal{N}(\mathbf{H}\boldsymbol{\beta}_{prior}, \mathbf{H}\mathbf{Q}_{prior}\mathbf{H}^t + \mathbf{K}). \tag{7.28}$$

Thus, the linear approximation (7.24) yields the simple Gaussian distributions (7.27) and (7.28) for $\boldsymbol{y}_{obs}$, contrary to subsection 7.3.2, where the distributions in the frequentist and Bayesian case are general, because of the general nature of f_{mod}.

Estimation of the covariance function of the model error

Similarly to subsection 7.3.2, we consider that the variance σ^2_{mes} of the measure error is known and that the covariance function of the model error Z is estimated from $\boldsymbol{y}_{obs}$ in the parametric family

$$\mathcal{K}_{mod} = \{K_{mod,\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\},$$

with $K_{mod,\boldsymbol{\psi}}$ a stationary covariance function and $\Psi \subset \mathbb{R}^p$.

Since we deal with a classical universal Kriging model for $\boldsymbol{y}_{obs}$, the estimation methods presented in chapter 3 can be adapted to estimate $\boldsymbol{\psi}$. We use preferably the REML method of chapter 3. The advantage of REML is that the estimation of $\boldsymbol{\psi}$ is independent of the estimation of $\boldsymbol{\beta}$. Furthermore, this method enables us to have the same estimation of $\boldsymbol{\psi}$ in both the prior and

no prior information case. Finally, let us notice that $n > m$ is required for the REML method, that is to say there are more experiments than model parameters. In thermal-hydraulics, the field of the application case of chapter 8, this condition holds. Nevertheless, in other fields of Nuclear Engineering, typically in neutron transport [KHF+06], one may have $m \gg n$. In this case, if one wants to address the present model error modeling anyway, it is recommended to work in a fully Bayesian framework, both for the model parameters and the covariance hyper-parameters as described in [SWN03] section 4.1.4, and in subsection 7.3.2 for the non-linear case for f_{mod}. Indeed, the very large number of model parameters makes the uncertainty related to the hyper-parameters of the model error covariance function too large to be neglected, as it is done when these hyper-parameters are fixed to their estimated values.

Let us denote $\mathbf{K}_{\boldsymbol{\psi}} = \mathbf{K}_{mod,\boldsymbol{\psi}} + \sigma_{mes}^2 \mathbf{I}_n$ and let $\mathbf{U}, \mathbf{S}, \mathbf{V}$ be a Singular Value Decomposition of $\mathbf{H}$, with $\mathbf{U}$ of size $n \times m$ so that $\mathbf{U}^t\mathbf{U} = \mathbf{I}_{m,m}$, $\mathbf{S}$ a diagonal matrix of size m, with nonnegative numbers on the diagonal, and $\mathbf{V}$ an orthogonal matrix of size m, so that $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^t$. Then, the REML estimation of $\boldsymbol{\psi}$ is defined by (see chapter 3)

$$\hat{\boldsymbol{\psi}} \in \underset{\boldsymbol{\psi}}{\operatorname{argmin}}\, q(\boldsymbol{\psi}), \tag{7.29}$$

with

$$\begin{aligned} q(\sigma, \boldsymbol{\theta}) &= \ln\left|\mathbf{U}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{U}\right| + \ln|\mathbf{K}_{\boldsymbol{\psi}}| + \boldsymbol{y}_{obs}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y}_{obs} \\ &\quad -\boldsymbol{y}_{obs}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{U}(\mathbf{U}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\mathbf{U})^{-1}\mathbf{U}^t\mathbf{K}_{\boldsymbol{\psi}}^{-1}\boldsymbol{y}_{obs}. \end{aligned} \tag{7.30}$$

We recall (see chapter 3) that it does not matter if $\mathbf{H}$ is ill-conditioned, or even singular, since its singular values are actually not used in the computation of the Restricted Likelihood.

Calibration and prediction

Throughout this subsection, we assume that the covariance function K_{mod} of Z is estimated and fixed and we use the classical Kriging formulas of chapter 2 to solve the calibration and prediction problems. Thus, let $\mathbf{K} = \mathbf{K}_{mod} + \sigma_{mes}^2\mathbf{I}_n$ be the fixed $n \times n$ matrix defined by $K_{i,j} = Cov(z_i + \epsilon_i, z_j + \epsilon_j)$.

In the no prior information case, the calibration problem is the frequentist problem of estimating the unknown parameter $\boldsymbol{\beta}$. From chapter 2, the maximum likelihood estimation of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{y}_{obs}. \tag{7.31}$$

This estimator is unbiased and has covariance matrix

$$cov(\hat{\boldsymbol{\beta}}) = (\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}. \tag{7.32}$$

We see that if there is a $\boldsymbol{\beta}$ so that $\mathbf{H}\boldsymbol{\beta} = \boldsymbol{y}_{obs}$, then we have $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. This means that, if we are in the favorable case when the computer model can perfectly reproduce the experiments, then the Gaussian process calibration of the computer model will achieve this perfect reproduction, as should be expected. Finally, as the random vector $\hat{\boldsymbol{\beta}}$ has Gaussian distribution, its covariance matrix is sufficient to yield confidence ellipsoids for $\boldsymbol{\beta}$.

In the prior information case, from chapter 2, the posterior distribution of $\boldsymbol{\beta}$ given the observations $\boldsymbol{y}_{obs}$ is Gaussian with mean vector

$$\boldsymbol{\beta}_{post} = \boldsymbol{\beta}_{prior} + (\mathbf{Q}_{prior}^{-1} + \mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}\mathbf{H}^t\mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \mathbf{H}\boldsymbol{\beta}_{prior}), \tag{7.33}$$

and covariance matrix

$$\mathbf{Q}_{post} = (\mathbf{Q}_{prior}^{-1} + \mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}. \tag{7.34}$$

We can notice that, similarly to chapter 2, when $\mathbf{Q}_{prior}^{-1} \to 0$, then the prior information case calibration tends to the no prior information case calibration. This is an intuitive fact, because $\mathbf{Q}_{prior}^{-1}$ small corresponds to a small a priori knowledge of $\boldsymbol{\beta}$ and hence should, in the limit case, correspond to an absence of knowledge.

Remark 7.9. *The prior information case calibration of* (7.33) *is classically used in neutron transport [KHF+06], when the linear approximation* (7.24) *of the computer model is also made. In the reference hereabove, no model error is assumed, so that the physical system is predicted by the calibrated computer model only. In thermal-hydraulics, which is the field of the case of application in chapter 8, this hypothesis is not justified. Indeed, computer models can rely on aggregation of correlation models that have no physical justification. We will see in the prediction formulas of* (7.35) *and* (7.37)*, and in the FLICA 4 application case of chapter 8, that modeling the model error allows to significantly improve the predictions of a computer model that is only partially representative of the physical system.*

We now present the prediction formulas. For a new experimental condition $\boldsymbol{x}^{(new)}$, we denote by $\boldsymbol{h}(\boldsymbol{x}^{(new)})$ the $n \times 1$ vector of derivatives of the computer model with respect to $\beta_1, ..., \beta_m$ at $\boldsymbol{x}^{(new)}$. Hence we have $\left(\boldsymbol{h}(\boldsymbol{x}^{(new)})\right)_i = h_i(\boldsymbol{x}^{(new)})$. We also denote $\boldsymbol{k}(\boldsymbol{x}^{(new)})$ the covariance vector of the model error between $(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)})$ and $\boldsymbol{x}^{(new)}$, that is $\left(\boldsymbol{k}(\boldsymbol{x}^{(new)})\right)_i := K_{mod}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(new)})$.

Following (2.15), in the no prior information case, the Best Linear Unbiased Predictor (BLUP) of $Y_{real}(\boldsymbol{x}^{(new)})$ with respect to the vector of observations $\boldsymbol{y}_{obs}$ is

$$\hat{y}(\boldsymbol{x}^{(new)}) = \underbrace{(\boldsymbol{h}(\boldsymbol{x}^{(new)}))^t\hat{\boldsymbol{\beta}}}_{\text{calibrated computer model}} + \underbrace{(\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t\mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \mathbf{H}\hat{\boldsymbol{\beta}})}_{\text{inferred model error}}, \tag{7.35}$$

with $\hat{\boldsymbol{\beta}}$ as in (7.31). As in (7.20), this predictor is composed of the calibrated computer model and of the inferred model error. The inferred model error has the two following properties, as we have discussed for (7.20). For fixed prediction point it is large when the differences between the observations and the calibrated code are large, and for fixed observations, it decays when one moves away from observation points, the distance being defined in terms of the covariance function of the model error.

The mean square error of the BLUP of $Y_{real}(\boldsymbol{x}^{(new)})$ is, see (2.16),

$$\begin{aligned}\hat{\sigma}^2(\boldsymbol{x}^{(new)}) = K_{mod}(\boldsymbol{x}^{(new)}, \boldsymbol{x}^{(new)}) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)}) \\ +(\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t(\mathbf{H}^t\mathbf{K}^{-1}\mathbf{H})^{-1}(\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}^{(new)})).\end{aligned} \tag{7.36}$$

Since only linear combinations have been used, the BLUP has Gaussian distribution and the mean square error allows to build confidence intervals.

In the prior information case, from chapter 2, the posterior distribution of $Y_{real}(\boldsymbol{x}^{(new)})$ given the observations $\boldsymbol{y}_{obs}$ is Gaussian with mean

$$\hat{y}(\boldsymbol{x}^{(new)}) = \underbrace{(\boldsymbol{h}(\boldsymbol{x}^{(new)}))^t \boldsymbol{\beta}_{post}}_{\text{calibrated computer model}} + \underbrace{(\boldsymbol{k}(\boldsymbol{x}^{(new)}))^t \mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \mathbf{H}\boldsymbol{\beta}_{post})}_{\text{inferred model error}}, \tag{7.37}$$

with $\boldsymbol{\beta}_{post}$ from (7.33), and variance

$$\begin{aligned}&\hat{\sigma}^2(\boldsymbol{x}^{(new)}) = \\ &K_{mod}(\boldsymbol{x}^{(new)}, \boldsymbol{x}^{(new)}) - \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1} \boldsymbol{k}(\boldsymbol{x}^{(new)}) \\ &+ \\ &(\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t \mathbf{K}^{-1} \boldsymbol{k}(\boldsymbol{x}^{(new)}))^t (\mathbf{H}^t \mathbf{K}^{-1} \mathbf{H} + \mathbf{Q}_{prior}^{-1})^{-1} (\boldsymbol{h}(\boldsymbol{x}^{(new)}) - \mathbf{H}^t \mathbf{K}^{-1} \boldsymbol{k}(\boldsymbol{x}^{(new)})).\end{aligned} \tag{7.38}$$

We can make the same remarks as for (7.35). Similarly to calibration, the limit when $\mathbf{Q}_{prior}^{-1} \to 0$ of the prediction in the prior information case is the prediction in the no prior information case.

Let us conclude about the advantage of the linear approximation (7.24) from a simplicity point of view. Once the model error function K_{mod} is fixed, the calibration and prediction, and the quantification of the resulting uncertainty, are carried out explicitly by (7.31)-(7.38). In subsection 7.3.2, there is no explicit equivalent of (7.31)-(7.38). On the contrary, numerical optimization or MCMC methods are necessary.

The analytical test case of figures 2.6 and 2.7 revisited

The calibration and prediction equations (7.31)-(7.38) are those of a classical universal Kriging model, in which the calibrated code plays the role of the estimated mean function, and the inferred model error plays the role of the predicted deviation from the mean function. Hence, the qualitative conclusions we had drawn from figures 2.6 and 2.7 apply to them.

Let us summarize these conclusions from a computer model calibration and model error modeling point of view. First, in figures 2.6 and 2.7, the calibrated parameter $\boldsymbol{\beta}$ does not make the code function reproduce the experiments as best as possible. Indeed, the calibrated code alone does not constitute a predictive model of the physical system. When this calibrated code is completed by the inference of the model error, this results in a very accurate prediction of the physical system. Furthermore, in the framework of figures 2.6 and 2.7, the calibrated code predicts the physical system better in extrapolation (far from the experimental data) that a code function that would reproduce the experiments as best as possible.

Second, in extrapolation, the model error cannot be precisely inferred from the available observations and the inferred model error in (7.35) and (7.37) is hence very close to zero. Hence, in extrapolation, the prediction is made using the calibrated computer model only. This is expected, because when one cannot statistically improve the prediction of the computer model, a conservative choice is to rely only on physical knowledge.

General recommendations for the Gaussian process modeling with the linearization

The Gaussian process modeling of the model error, with the linear approximation (7.24) of the computer model with respect to the model parameters, is rather simple to carry out. We will

see, in the case of the thermal-hydraulic code FLICA 4 in chapter 8, that it has the potential to both improve the prediction capability of the computer model and correctly assess the resulting uncertainty. Before that, we conclude chapter 7 by giving general practical recommendations concerning the use of the Gaussian process modeling, with the linearization of the computer model with respect to the model parameters.

The first important point is that, in chapters 7 and 8, we do not address the complex field of code verification. As a consequence, discretization or numerical parameters, such as the length or volume of a cell in a numerical scheme, should not be considered as model parameters or treated by the present method without further study.

Let us now discuss the linear approximation (7.24). If the main objective is to achieve a precise enough prediction of the physical system, and not to calibrate the computer model, then it does not matter if the computer model is not linear with respect to its model parameters. Indeed, the linear approximation boils down to modeling the Gaussian process Z in (7.14) as the model error of the *linearized* computer model in (7.24). In the prediction formulas (7.35) and (7.37), we can see that the statistical correction can compensate for the linear approximation error of the code. This fact is confirmed in chapter 8 for the thermal-hydraulic code FLICA 4. Now, if calibration in itself is one of the main objectives, one should act with caution as regards to the linear approximation. In this case, we advise to run a sensitivity analysis first to check the linearity assumption (e.g the Morris method [Mor91]). If the linearity assumption is infirmed, then we recommend to proceed in two steps. First, a non-linear calibration should be carried out, like the non-linear Bayesian calibration of (7.19). Then, the model parameters should be fixed to their calibrated values, or a very narrow prior, centered around these values, should be used, before using the linearization of the computer model.

Concerning the computation of the derivatives with respect to the model parameters $\boldsymbol{\beta}$, two cases are possible. First the code can already provide them, by means of the Adjoint Sensitivity Method for instance [Cac03]. Similarly, automatic differentiation methods can be used on the source file of the code and yield a differentiated code [HP04]. If these kinds of methods are not available, finite differences are necessary to approximate the derivatives. Our main advice here is not to use a too small variation step. Indeed, on the one hand, if the code is approximately linear with respect to the model parameters, a too large variation step will provide a good estimate of the derivatives anyway, whereas a too small variation step can yield numerical errors. On the other hand, if the code is not approximately linear, the linear approximation should not be used for calibration. For prediction, the model error compensates for the linear approximation error as well as for the error in calculating the derivatives.

The fourth important point is that extrapolation is not recommended. This is a general advice for all Kriging models. The experimental results should be obtained in the prediction domain of interest. Hence, for example, Kriging methods are not advisable to address scaling issues, that intrinsically ask to extrapolate experimental results from one scale to another.

When dealing with more complex systems than that of the application case of chapter 8, such as system-thermal hydraulics, one may deal with high-dimensional problems, either with respect to the number of experimental conditions (dimension of $\boldsymbol{x}$) or to the number of model parameters (dimension of $\boldsymbol{\beta}$). The dimension of $\boldsymbol{x}$ is a potential problem. A common rule of

thumb for Kriging models is that one should have $n \geq 10dim(\boldsymbol{x})$. Note that screening methods exist and allow to select only the most impacting experimental conditions [MIDV08]. If the number of experiments is really too small compared to the dimension of $\boldsymbol{x}$, our opinion is that it is not possible to take into account the model error correctly, so that only the calibration should be carried out. If $\boldsymbol{\beta}$ is high-dimensional, we advise to use a Bayesian prior distribution both on $\boldsymbol{\beta}$ and on the covariance hyper-parameters for the model error, as we have discussed when presenting REML in (7.29). An alternative is to select only the most important model parameters (from physical expertise), and to fix the other model parameters at their nominal values. In this case the Gaussian process modeling of the model error also compensates for the error made by freezing these parameters.

Implementation in the gpLib library

The Gaussian process modeling of the model error, with the linear approximation (7.24) of the computer model with respect to the model parameters, has been implemented in the gpLib library [MMGB12]. The gpLib library, written in C language, provides the elementary functions for a universal Kriging model: prediction, conditional simulation, and criteria for estimation by Maximum Likelihood and Cross Validation. It has been integrated in the URANIE uncertainty platform, developed at CEA, in the objective of providing an autonomous and user-friendly Kriging framework, that can be used for a large variety of computer experiment problems.

Chapter 8

Calibration and improved prediction of the thermal-hydraulic code FLICA 4

This chapter is inspired by the article [BBGM]. We present an application case on the thermal-hydraulic code FLICA 4, for the Gaussian process modeling of the model error of subsection 7.3.3. The thermal-hydraulic code FLICA 4 is mainly dedicated to core thermal-hydraulic transient and steady state analysis [TBG$^+$00]. In the present context, FLICA 4 is used as a physical modeling of an experiment consisting in measuring the pressure drop in an ascending pressurized flow of liquid water through a tube that can be electrically heated. We focus on the frictional pressure drop (ΔP_{fric}) in a single phase flow. Several experimental results are available, giving the observed values of the variable of interest ΔP_{fric}, for different experimental conditions.

In section 8.1, we introduce the thermal-hydraulic code FLICA 4, and the associated experimental results. In section 8.2, we discuss the practical aspects of the Gaussian process modeling of the model error of subsection 7.3.3. We also introduce the Cross Validation procedure for the evaluation of the predictions obtained from the Gaussian process model. In section 8.3, we present and discuss the results of the Cross Validation procedure on the experimental results of section 8.1.

8.1 Presentation of FLICA 4 and of the experimental results

8.1.1 The thermal-hydraulic code FLICA 4

In this chapter 8, we focus on the single phase regime, meaning that all the water is in the liquid state during the experiment. The mathematical model for ΔP_{fric}, in the single phase regime,

is given by the local equation

$$\Delta P_{fric} = \frac{H}{2\rho D_h} G^2 f_{iso} f_h. \tag{8.1}$$

In (8.1), each quantity implicitly depends on space and time. (8.1) is hence numerically integrated in space and time by the thermal-hydraulic code FLICA 4. In (8.1), H is the friction height, ρ is the density, D_h is the hydraulic diameter, and G is the flowrate. f_{iso} and f_h are the friction coefficients respectively in the isothermal and heated flow regimes. The isothermal regime is defined by the temperature of the liquid being uniformly equal to the wall temperature. On the other hand, the heated flow regime is characterized by a heat flux imposed on the test section and thus a varying liquid temperature.

The friction coefficient in the isothermal regime is

$$f_{iso} = \begin{cases} \frac{a_l}{Re} & \text{if } Re < Re_l \\ \frac{a_t}{Re^{b_t}} & \text{if } Re_t < Re \\ \frac{a_l}{Re}\frac{Re_t - Re}{Re_t - Re_l} + \frac{a_t}{Re^{b_t}}\frac{Re - Re_l}{Re_t - Re_l} & \text{if } Re_l < Re < Re_t \end{cases} \tag{8.2}$$

where $R_e = \frac{GD_h}{\mu}$ is the Reynolds number and μ is the viscosity. The limiting values Re_l and Re_t for the Reynolds number are defined according to the literature and represent the limits of the transition regime between laminar and turbulent flows. a_l, a_t and b_t are parts of the model parameters of the thermal-hydraulic code FLICA 4. They are the three components of the vector $\boldsymbol{\beta}$ of model parameters in the isothermal regime.

The friction coefficient in the heated flow regime is a correction factor expressed as

$$f_h = 1 - \frac{P_h}{P_w} \frac{C_f (T_w - T_b)}{1 + d\left(\frac{T_w + T_b}{2T_0}\right)^n} \tag{8.3}$$

where P_h and P_w are the heated and wetted perimeters, T_w is the wall temperature, T_b is the bulk temperature, and $T_0 = 100°C$ is a normalization temperature. C_f, n and d are the three components of the vector $\boldsymbol{\beta}$ of model parameters in the heated flow case. Finally, note that tests with no heat flux (isothermal tests) result in $T_w = T_b$, therefore the correction factor f_h is equal to 1, as expected.

To summarize, in (8.1), the isothermal regime is defined by $f_h = 1$, the heated flow regime is defined by $f_h \neq 1$, and both regimes are subcases of the single phase regime.

Finally, the simulation time of the thermal-hydraulic code FLICA 4 is approximately one minute, to reproduce one experiment with one calibration parameter value. Hence, the matrix $\mathbf{H}$ of the derivatives of the code with respect to the model parameters, for all the experiments, in (7.25) can be computed by finite difference in a reasonable time.

8.1.2 The experimental results

Several experimental tests have been conducted in order to calibrate the FLICA 4 friction model. These tests have been used in previous calibration studies. The database is composed of n_i measurements in the isothermal regime, and n_h measurements in the heated flow regime. An experimental condition $\boldsymbol{x}$ consists in geometrical data (the channel width e, the hydraulic diameter D_h, and the friction height H_f) and in thermal-hydraulic conditions (the outlet pressure

P_o, the flowrate G_i, the wall heat flux ϕ_w, the inlet liquid enthalpy h_i^l, the thermodynamic title X_{th}^i, and the inlet temperature T_i). With respect to the nomenclature of section 7.1, the geometric data e, D_h, and H_f are control variables and the thermal-hydraulic conditions P_o, G_i, ϕ_w, h_i^l, X_{th}^i and T_i are environment variables. For each test the pressure drop due to friction ΔP_{fric} is measured.

8.2 Description of the procedure for the Gaussian process modeling

8.2.1 Objectives for the universal Kriging procedure

We carry out the Gaussian process modeling method, with the linear approximation, of subsection 7.3.3, on the thermal-hydraulic code FLICA 4 in the isothermal and heated flow regimes. We limit the calibration part of the study to the parameters a_t and b_t. That is to say, we enforce the parameter a_l of the isothermal model, and the parameters C_f, n and d of the heat correction model to their nominal values, computed in previous calibration studies. Indeed, the parameters a_t and b_t are the most influent parameters for the thermal-hydraulic code FLICA 4.

We work in the prior information case (calibration given by (7.33)). From previous calibration studies, we have $\boldsymbol{\beta}_{prior} = (0.22, 0.21)^t$. $\mathbf{Q}_{prior}$ corresponds to a 50% uncertainty and is chosen diagonal with diagonal vector $(0.11^2, 0.105^2)^t$. Hence, this prior is rather large, so that the calibration essentially depends on the experimental results.

An important point is that the two categories of experimental conditions (control and environment variables) are not equally represented in the experimental results. Indeed, the $n_i + n_h$ experiments are divided into eight campaigns. Within a campaign, the control variables remain constant, while the environment variables are varying. Hence, we only dispose of eight different control variables triplet. This means that, from the point of view of the prediction (7.37) given by the Gaussian process model, it is a very unlikely that the prediction of the calibrated code is significantly improved when considering new control variables. We experienced that, when predicting for new control variables, the Gaussian process method does not damage the predictions given by the nominal calibration of the thermal-hydraulic code FLICA 4 but it does not significantly improve it. However, as we see next, we can give significantly improved predictions for observed control variables and new environmental variables.

To summarize, this study follows the double objective of calibration and prediction, in the prior information case for the parameters a_t and b_t. Concerning the prediction, the objective is to predict for experienced control variables and new environment variables.

8.2.2 Exponential, Matérn and Gaussian covariance functions considered

The environment and control variables listed above are not independent. Hence, it would be redundant to incorporate all of them in the covariance function. One possible minimal set of environment and control variables is the set $(G_i, \phi_\omega, h_i^l, P_o, H_f, D_h)$. For this set, we will use the covariance function K_{mod} for the model error, with K_{mod} being built from one of the four

one-dimensional exponential, Matérn $\frac{3}{2}$, Matérn $\frac{5}{2}$ or Gaussian covariance functions of table 2.1. From the one-dimensional exponential covariance function, we use the tensorized version (2.7) to build the 6-dimensional covariance function K_{mod}. From the one-dimensional Matérn $\frac{3}{2}$, Matérn $\frac{5}{2}$ and Gaussian covariance functions, we use the isotropic version (2.6).

To summarize, we represent the experimental conditions of an experiment by

$$\boldsymbol{x} = (G_i, \phi_w, h_i^l, P_o, H_f, D_h).$$

The covariance function is $K_{mod}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \sigma^2 R_{mod}\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right)$, with R_{mod} being either, the tensorized exponential, the isotropic Matérn $\frac{3}{2}$, the isotropic Matérn $\frac{5}{2}$ or the Gaussian correlation function of table 2.1, (2.6) and (2.7). The hyper-parameters to be estimated are the variance σ^2 and the six correlation lengths $\ell_1, ..., \ell_6$. They are estimated by REML, as described in subsection 7.3.3.

Finally, we consider that the covariance matrix of the measure error vector $(\epsilon_1, ..., \epsilon_{n_i+n_h})$ is $\mathbf{K}_{mes} = \sigma_{mes}^2 \mathbf{I}_{n_i+n_h}$, with $\sigma_{mes} = 150Pa$ provided by the experimentalists.

8.2.3 K-folds Cross Validation for Kriging model validation

We have seen in subsection 2.2.4 that the quality of a Kriging model should not be evaluated on the data that helped to build it. Instead, Cross Validation is a very natural method to assess the predictive capability of a Kriging model.

We use a K-fold Cross Validation procedure, with $K = 10$, to evaluate the quality of the Gaussian process predictions (7.37) and (7.38). This Cross Validation procedure calculates the two following quality criteria.

$$RMSE^2 = \frac{1}{n} \sum_{i_c=1}^{n_c} \sum_{\boldsymbol{x} \in C_{i_c}} (\hat{y}_{\overline{C}_{i_c}}(\boldsymbol{x}) - Y_{obs}(\boldsymbol{x}))^2 \tag{8.4}$$

and

$$IC = \frac{1}{n} \sum_{i_c=1}^{n_c} \sum_{\boldsymbol{x} \in C_{i_c}} \mathbf{1}_{|\hat{y}_{\overline{C}_{i_c}}(\boldsymbol{x}) - Y_{obs}(\boldsymbol{x})| \leq 1.64 \hat{\sigma}_{\overline{C}_{i_c}}(\boldsymbol{x})}. \tag{8.5}$$

In (8.4) and (8.5), we partition the set of n experiments into $n_c = 10$ subsets $C_1, ..., C_{n_c}$, each subset being well distributed in the experimental domain. To build these subsets, we start from a numbering of the experiments for which two successive experiments are similar (for instance the experiments for the same control variables have successive indices). Then the subset 1 gather the experiments with indices $1, 11, ...$, the subset 2 those with indices $2, 12, ...$ and so on. For $\boldsymbol{x} = \boldsymbol{x}^{(i)} \in C_{i_c}$, we denote $Y_{obs}(\boldsymbol{x}) = y_{obs,i}$, with the notation (7.25). $\overline{C}_{i_c}$ is the set of experimental conditions and observations that is the union of the subsets $C_1, ..., C_{i_c-1}, C_{i_c+1}, ..., C_{n_c}$. $\hat{y}_{\overline{C}_{i_c}}(\boldsymbol{x})$ and $\hat{\sigma}_{\overline{C}_{i_c}}(\boldsymbol{x})$ are the posterior mean and standard deviation of the predicted output $Y_{obs}(\boldsymbol{x})$ given the experimental data in $\overline{C}_{i_c}$. $[\hat{y}_{\overline{C}_{i_c}}(\boldsymbol{x}) - 1.64\hat{\sigma}_{\overline{C}_{i_c}}(\boldsymbol{x}), \hat{y}_{\overline{C}_{i_c}}(\boldsymbol{x}) + 1.64\hat{\sigma}_{\overline{C}_{i_c}}(\boldsymbol{x})]$ corresponds to a 90% confidence interval. It is emphasized that at step i_c of the Cross Validation, the Gaussian process model is built without using the experimental results of the class C_{i_c}. Hence the important point is that, in the computation of the posterior mean and variance of the observed value $Y_{obs}(\boldsymbol{x})$ at $\boldsymbol{x}$, this observed value is unused, for the estimation of the hyper-parameters as well as for the prediction formula.

Remark 8.1. *In subsection 2.2.4, we have presented the virtual Cross Validation formulas in the case where the covariance function is not reeestimated at each step of the cross validation. In this chapter 8, we choose instead to reestimate the covariance hyper-parameters at each step of the Cross Validation. This is indeed more precise, since we observe in table 8.2 that the estimated values of these hyper-parameters vary among the different CV steps. Furthermore, the additional computational cost of the reestimation is not prohibitive, because of the moderate number of observation points.*

8.3 Results

8.3.1 Results in the isothermal regime

In a first step, we consider the results in the isothermal and turbulent flow regime only. That is to say, the regime when $f_h = 1$ in (8.1), and when $Re > Re_t$ in (8.2). We have $n_{it} < n_i$ experimental results.

The isothermal regime is characterized by no wall heat flux, $\phi_w = 0$. Hence, it is useless to include it in the covariance function, because it is uniformly zero for all the experimental conditions. So, we only have five correlation lengths out of six to estimate, which are ℓ_1, ℓ_3, ℓ_4, ℓ_5 and ℓ_6 corresponding to G_i, h_i^l, P_o, H_f and D_h.

On figure 8.1, we plot, for the 10-fold Cross Validation, the $n_c = 10$ posterior mean values of a_t and b_t for the four covariance functions of subsection 8.2.2. The conclusions are that the Gaussian process calibration does not significantly change the nominal values $a_t = 0.22$ and $b_t = 0.21$. Furthermore we do not notice significant differences concerning the choice of the covariance function for the calibration. Finally, we can observe a high correlation in the posterior means of a_t and b_t. This is confirmed in the n_c posterior covariance matrix, where the correlation coefficient is larger than 0.95.

Concerning the prediction, we first compute the $RMSE$ and IC criteria for the four covariance functions. Results are presented in table 8.1. The first comment is that the predictive variances of (7.38) are reliable, because they yield rather precise 90% confidence intervals. This is observed in a general way for Kriging, e.g in [LA12]. The second comment is that there is no significant difference between the different covariance functions. This may be due to the amplitude of the measurement error, which makes insignificant the problem of the regularity of the covariance function. It is shown in [Ste99] section 3.7 that, in a particular asymptotic context, even a small measurement error can have a significant effect on prediction errors.

We now present more detailed results for the Matérn $\frac{3}{2}$ covariance function. We first compare the Gaussian process predictions with the predictions given by the calibrated code alone. With the same Cross Validation procedure, the RMSE criterion for the calibrated code alone is $RMSE = 741Pa$. This is to be compared with a $RMSE$ around $300Pa$ for the Gaussian process method. Hence the inference of the model error process significantly improves the predictions of the code. We illustrate this in figure 8.2, where we plot, for each of the n_{it} observations, the predicted values and confidence intervals with the 10-fold Cross Validation method. The plots are done with respect to the experiment index. This index has physical meaning, because two experiments with successive indices are similar, as we discuss after (8.5). We first see that the

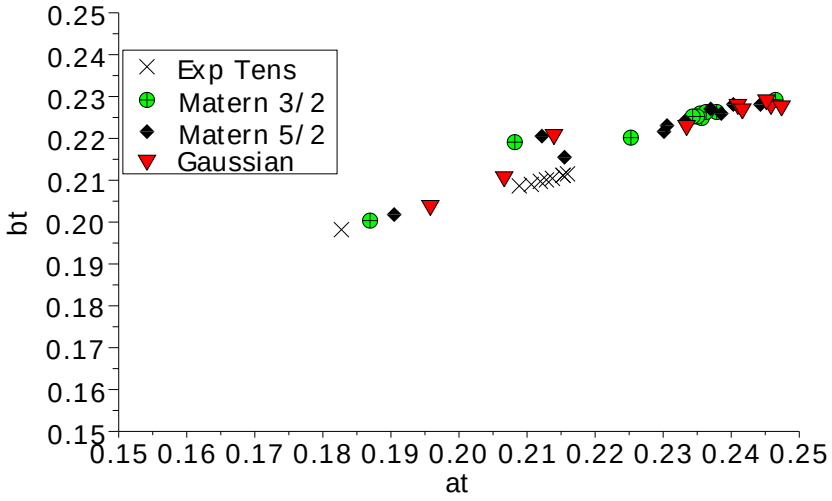

Figure 8.1: Calibration in the isothermal regime. 10-fold Cross Validation. Plot of the $n_c = 10$ posterior means (7.33) of a_t and b_t for the exponential, Matérn 3/2, Matérn 5/2 and Gaussian covariance functions of subsection 8.2.2.

Covariance function	$RMSE\ (Pa)$	IC
exponential	289.5	0.93
Matérn $\frac{3}{2}$	296.2	0.92
Matérn $\frac{5}{2}$	302.7	0.89
Gaussian	310.8	0.88

Table 8.1: Prediction results in the isothermal regime. RMSE and IC criteria of (8.4) and (8.5) obtained with a 10-fold Cross Validation procedure, for the covariance functions presented in subsection 8.2.2.

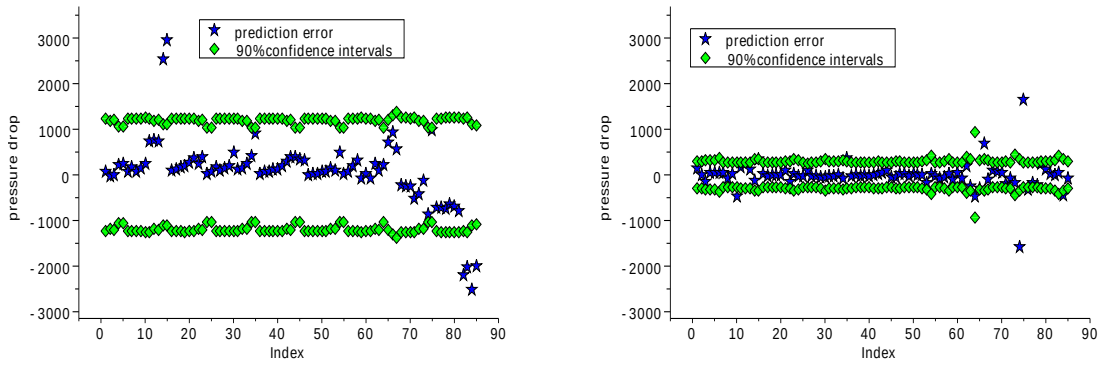

Figure 8.2: Prediction errors (observed values minus predicted values (7.37)) and 90% confidence intervals for these prediction errors, derived by the calibrated thermal-hydraulic code FLICA 4 (left), and the Gaussian process method (right). 90% confidence intervals are of the form $[-1.65\hat{\sigma}(\boldsymbol{x}), 1.65\hat{\sigma}(\boldsymbol{x})]$ with $\hat{\sigma}(\boldsymbol{x})$ given by (7.38). Plot with respect to the index of experiment.

Gaussian process modeling significantly reduces the prediction errors, and that the confidence intervals are reliable. Then, we observe a regularity in the plot of the prediction error for the calibrated code, especially for the largest indices. This regularity is not present anymore in the error of the Gaussian process method. The conclusion is that the Gaussian process method detects a regularity in the error of the calibrated code, and uses it to significantly improve its predictions.

Finally, in table 8.2, we show the $n_c = 10$ different estimations of $\sigma^2, \ell_1, \ell_3, \ell_4, \ell_5, \ell_6$, for the different steps of the Cross Validation. The first conclusion is the singularity at steps 5 and 6 of the Cross Validation. The explanation is that, among the n_{it} experimental results, there are two singular points that have very similar experimental conditions but substantially different values for the quantity of interest. These two points are in CV classes 5 and 6. Hence the estimation of the hyper-parameters in the CV steps 1, 2, 3, 4, 7, 8, 9, 10, where this singularity is present in the data used for the estimation, is different from the steps 5 and 6, where the singularity is absent. On figure 8.2, these two singular points yield the two largest prediction errors for the Gaussian process method. Indeed, when one of them is in the test group, the other one is in the learning group. As the Gaussian process modeling principle is to assume a correlated model error, the quantity of interest of the singular point of the test group is (up to the measurement error) predicted by the quantity of interest of the singular point of the learning group.

The correlation lengths in table 8.2 correspond to normalized experimental conditions varying between 0 and 1. Hence, the second conclusion is that the estimated correlation lengths are rather large, corresponding to rather large scales of variations of the model error, as discussed for figure 8.2. When an estimated correlation length is very large (larger than 10), it is equivalent to assuming that the model error is independent of the corresponding experimental condition. For instance, for all the CV steps, except step 6, the estimated correlation length ℓ_6, associated to the hydraulic diameter D_h, is very large. Thus, the hydraulic diameter is estimated as a non-influent input in nine CV steps out of ten. Similarly, the outlet pressure P_o, associated to

Cross Validation step	σ	ℓ_1	ℓ_3	ℓ_4	ℓ_5	ℓ_6
1	2220	2.3	4.0	100	0.40	53
2	2100	2.2	3.5	100	0.40	100
3	2088	2.1	3.8	100	0.39	100
4	2266	2.3	2.0	100	0.50	100
5	4491	3.4	100	24	1.36	100
6	1953	1.6	15	3.4	7.7	0.6
7	2385	2.4	4.6	100	0.44	100
8	2436	2.4	4.8	100	0.45	99
9	2331	2.4	4.2	100	0.43	100
10	2294	2.4	3.8	100	0.42	100

Table 8.2: Estimated hyper-parameters in the isothermal regime. Estimated correlation lengths for the Matérn $\frac{3}{2}$ covariance function of subsection 8.2.2, for the 10-fold Cross Validation procedure. For all the CV steps, except step 6, the estimated correlation length ℓ_6, associated to the hydraulic diameter D_h, is very large. Thus, the hydraulic diameter is estimated as a non-influent input in nine CV steps out of ten. Similarly, the outlet pressure P_o, associated to ℓ_4, is estimated as a non-influent input in eight CV steps out of ten.

ℓ_4, is estimated as a non-influent input in eight CV steps out of ten. The third conclusion is that the estimations of the hyper-parameters can vary moderately among the Cross Validation steps. This is an argument in favor of reestimating the hyper-parameters at each step of the Cross Validation, because this takes these variations into account. Finally let us notice that, for the Gaussian process model to be used for new experimental conditions, the hyper-parameters are to be reestimated with all the observations.

8.3.2 Results in the single-phase regime

We now use all the experiments of the single phase regime (isothermal and heated flow regimes), that is to say $n = n_i + n_h$ experiments. Hence, we estimate six correlation lengths for the six environment and control variables G_i, ϕ_w, h_i^l, P_o, H_f and D_h.

Concerning the prediction, we first compute the $RMSE$ and IC criteria for the four covariance functions. Results are presented in table 8.3. As in the isothermal case, we see that the predictive variances are reliable and that there is no significant difference between the four covariance functions. As for the isothermal regime, we present the results for the the Matérn $\frac{3}{2}$ covariance function in more details.

With the same Cross Validation procedure, the RMSE criterion for the calibrated code alone is $RMSE = 567Pa$. This is to be compared with a $RMSE$ around $200Pa$ of the Gaussian process method. Hence, the inference of the model error process significantly improves the predictions of the code, in the same way as in the isothermal regime. We illustrate this in figure 8.3, where we plot the same quantities as in figure 8.2. We obtain the same conclusion: the Gaussian process model detects a regularity in the error of the calibrated code, and uses it to

Covariance function	$RMSE\ (Pa)$	IC
exponential	202.2	0.95
Matérn $\frac{3}{2}$	196.2	0.95
Matérn $\frac{5}{2}$	196.9	0.95
Gaussian	199.5	0.94

Table 8.3: Prediction results in the single phase regime. Same setting as in table 8.1.

improve its predictions.

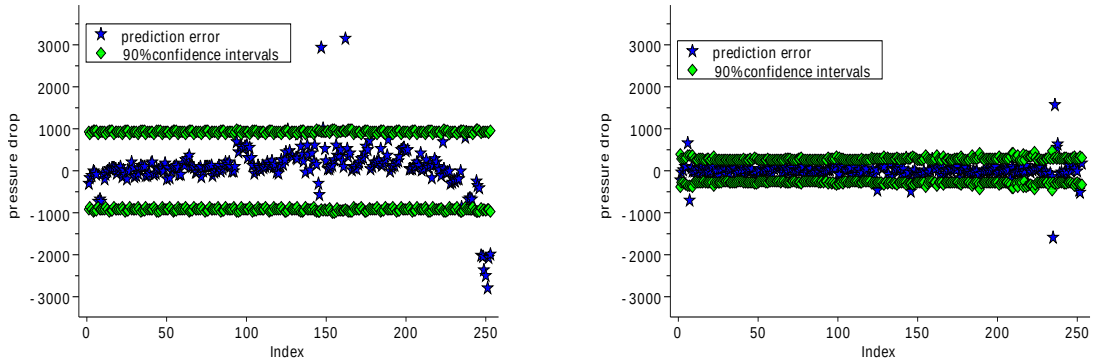

Figure 8.3: Same settings as in figure 8.2 but in the single phase regime.

8.3.3 Influence of the linear approximation

All the results above are obtained by using the linear approximation of the thermal-hydraulic code FLICA 4 with respect to a_t and b_t (subsection 7.3.3). We have implemented the calibration and prediction methods of subsection 7.3.2, when the thermal-hydraulic code FLICA 4 is not considered linear with respect to a_t and b_t. Integrals in the a_t, b_t domain were calculated on a 5×5 grid, which, to avoid bias, was also used when the linear approximation of the thermal-hydraulic code FLICA 4 was used.

More precisely, let $a_{t,1}, ..., a_{t,5}$ and $b_{t,1}, ..., b_{t,5}$ define the 5×5 regular integration grid. The posterior mean of $\boldsymbol{\beta}$ is approximated in the non-linear case, from (7.19), by

$$\frac{\sum_{i=1}^{5}\sum_{j=1}^{5}(a_{t,i}, b_{t,j})^t p(\boldsymbol{y}_{obs}|a_{t,i}, b_{t,j}) p(a_{t,i}, b_{t,j})}{\sum_{i=1}^{5}\sum_{j=1}^{5} p(\boldsymbol{y}_{obs}|a_{t,i}, b_{t,j}) p(a_{t,i}, b_{t,j})}, \tag{8.6}$$

with

$$p(a_{t,i}, b_{t,j}) = \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{|\mathbf{Q}_{prior}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}^{i,j} - \boldsymbol{\beta}_{prior})^t \mathbf{Q}_{prior}^{-1}(\boldsymbol{\beta}^{i,j} - \boldsymbol{\beta}_{prior})\right),$$

where $\boldsymbol{\beta}^{i,j} = (a_{t,i}, b_{t,j})^t$, and with

$$p(\boldsymbol{y}_{obs}|a_{t,i}, b_{t,j}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\mathbf{K}|}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{a_{t,i},b_{t,j}})^t \mathbf{K}^{-1} (\boldsymbol{y}_{obs} - \boldsymbol{m}^{a_{t,i},b_{t,j}})\right),$$

where $m_k^{a_{t,i},b_{t,j}} = f_{mod}(\boldsymbol{x}^{(k)}, a_{t,i}, b_{t,j})$ is the result of the FLICA 4 calculation, parameterized by $a_{t,i}, b_{t,j}$, for the experimental condition $\boldsymbol{x}^{(k)}$. The matrix $\mathbf{K}$ is the covariance matrix of the measure and error process at the experimental conditions $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, that is $K_{i,j} = Cov(Z(\boldsymbol{x}^{(i)}) + \epsilon_i, Z(\boldsymbol{x}^{(j)}) + \epsilon_j)$.

Remark 8.2. *In* (8.6)*, the* 25 *summation terms* $p(\boldsymbol{y}_{obs}|a_{t,i}, b_{t,j})p(a_{t,i}, b_{t,j})$*, in the numerator and the denominator, can be very small. For numerical reasons, we recommend to calculate their logarithms separately, and to subtract the largest of the logarithms to each of the logarithms. The equation* (8.6) *can then be computed with the* 25 *new summation terms, of which at least one is equal to* 1*. We make the same remark for* (8.7) *below.*

In the non-linear case, the conditional mean of the physical system, at a new point $\boldsymbol{x}^{(new)}$, is obtained from

$$\mathbb{E}\left(Y_{real}(\boldsymbol{x}^{(new)})|\boldsymbol{y}_{obs}\right) = \mathbb{E}\left(\mathbb{E}\left(Y_{real}(\boldsymbol{x}^{(new)})|\boldsymbol{y}_{obs}, \boldsymbol{\beta}\right)|\boldsymbol{y}_{obs}\right),$$

and is approximated by

$$\frac{\sum_{i=1}^{5}\sum_{j=1}^{5} \mathbb{E}\left(Y_{real}(\boldsymbol{x}^{(new)})|\boldsymbol{y}_{obs}, a_{t,i}, b_{t,j}\right) p(\boldsymbol{y}_{obs}|a_{t,i}, b_{t,j})p(a_{t,i}, b_{t,j})}{\sum_{i=1}^{5}\sum_{j=1}^{5} p(\boldsymbol{y}_{obs}|a_{t,i}, b_{t,j})p(a_{t,i}, b_{t,j})}, \tag{8.7}$$

with

$$\mathbb{E}\left(Y_{real}(\boldsymbol{x}^{(new)})|\boldsymbol{y}_{obs}, a_{t,i}, b_{t,j}\right) =$$
$$f_{mod}(\boldsymbol{x}^{(new)}, a_{t,i}, b_{t,j}) + \boldsymbol{k}(\boldsymbol{x}^{(new)})^t \mathbf{K}^{-1}(\boldsymbol{y}_{obs} - \boldsymbol{m}^{a_{t,i},b_{t,j}}),$$

where $\boldsymbol{k}(\boldsymbol{x}^{(new)})$ is the correlation vector of the model error process, between $\boldsymbol{x}^{(new)}$ and $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, that is $k_i(\boldsymbol{x}^{(new)}) = Cov(Z(\boldsymbol{x}^{(new)}), Z(\boldsymbol{x}^{(i)}))$. When we say that we use the same 5×5 regular grid, when FLICA 4 is linearized, we mean that we calculate the posterior mean of $(a_t, b_t)^t$ and the prediction of $Y_{real}(\boldsymbol{x}^{(new)})$ by using (8.6) and (8.7), where the FLICA 4 code is replaced by its linear approximation (7.24).

We consider the single phase regime, and we use the same 10-fold CV procedure as before for the formulas (8.6) and (8.7), in the linear and non-linear cases. We obtain $RMSE = 197.8$ with the linear approximation and $RMSE = 196.9$ without the linear approximation (less than 1% relative difference). The posterior means of a_t and b_t, along the different CV steps, have a Root Mean Square Difference of 0.025 (more than 10% relative difference), between the cases where the linear approximation was made or not. Hence, this is an illustration of the general remark in the recommendations of subsection 7.3.3: if the computer model is non-linear with respect to its calibration parameters, it is the model error with respect to the linearized computer model that is inferred. Thus, the predictions of the physical system are similar, whether or not the linear approximation is made.

Chapter 9

Kriging meta-modeling of the GERMINAL computer model

This chapter corresponds to an application case on the GERMINAL computer code, carried out in collaboration with Karim Ammar, PhD student at the Service d' Etudes des Réacteurs et de Mathématiques Appliquées, at CEA Saclay.

9.1 Introduction

This chapter aims at using Kriging as a metamodel (or response surface) of a complex computer code. A metamodel of a computer code is a function which has the same inputs and outputs as the code, which is much cheaper to use, and which is aimed to be a precise enough approximation of the code. We refer e.g. to [BD87] for an introduction to metamodels.

In this chapter 9, we consider metamodels that do not use any knowledge of the computer code (it is considered as a black box function). The construction of the metamodel only uses a sample of input points and of corresponding code values. Two classical examples of black-box metamodels are Kriging, as we have said, and artificial neural networks (see e.g. chapter 4 of [Mit97] for an introduction to artificial neural networks).

The goal of this chapter is to illustrate the good properties of Kriging metamodels. We will see that they give a precise approximation of the computer code, and that they also give a reliable prediction of the approximation error. The illustration is done with the GERMINAL thermal-mechanical code [MRPT92]. The GERMINAL code studies fuel pin thermal-mechanical behavior during steady-state and incidental conditions. Its utilization is part of a multi-physic and multi-objective optimal design problem of a reactor core. We work in this general framework, and focus on the metamodelization of the GERMINAL code. In this context, artificial neural networks have been used first, which enables us to compare the Kriging metamodelization results with the artificial neural network metamodelization results. We conclude that the Kriging prediction results are good compared to those of the artificial neural networks.

Furthermore, thanks to the Kriging predictive variance, Kriging models enable to give an expected order of magnitude for the prediction errors. By using this predictive variance, we

are able to automatically select the values, computed by the GERMINAL code, for which the Kriging prediction squared error is significantly larger than the predictive variance. By manually investigating these GERMINAL computations, the physicists are able to confirm that they do correspond to computation failures. We thus illustrate the strong interest of the probabilistic modeling underlying the Kriging metamodel, from the point of view of automatic outlier detection.

Chapter 9 is organized as follows. In section 9.2 we present the Kriging metamodeling of the GERMINAL computer model. In subsection 9.2.1 we introduce the nuclear core optimal design context underlying the utilization of the GERMINAL code. In subsection 9.2.2 we detail the inputs and outputs that we consider for the metamodeling of the GERMINAL code. In subsection 9.2.3, we present the settings we use for the Kriging model.

The results are presented in section 9.3. In subsection 9.3.1, we discuss the results of the Maximum Likelihood estimation of the covariance hyper-parameters of the Kriging model. In subsection 9.3.2, we consider the prediction results of the Kriging and artificial neural network metamodels. In subsection 9.3.3, we present the Kriging Leave-One-Out detection of GERMINAL output values that correspond to computation failures.

9.2 Presentation and context for the GERMINAL computer model

9.2.1 A nuclear reactor core design problem

The GERMINAL computer model [MRPT92] is a thermal-mechanical code, which studies fuel pin thermal-mechanical behavior during steady-state and incidental conditions. In a few words, a fuel pin consists of a hollow fuel cylinder, surrounded by a protective clad. A gas-filled gap exists between the clad and the fuel cylinder. In the primary circuit of a nuclear reactor core, a large number of fuel pins are embedded in fuel assemblies. Fuel assemblies are themselves aggregated in the reactor core. On figure 9.1, we give a schematic representation of a fuel pin and of a fuel assembly. In a reactor core, the coolant (sodium in figure 9.1) circulates in a fuel assembly, in between the fuel pins.

The GERMINAL CODE aims at studying the thermal-mechanical impact of the nuclear flux and power on a fuel pin. The aim is to answer the question: will the fuel pin resist the irradiation? The typical result of a GERMINAL calculation is a series of spatio-temporal functions giving the values of variables of interests in the fuel pin, during the simulated time period.

In the context that motivated this chapter 9, the GERMINAL code is used in a more general context of a nuclear core (multi-objective) optimal design. Thus, the GERMINAL elementary calculations for fuel pins are aggregated and coupled with other computer models addressing different physical problems.

The optimization problem requires a large number of computer model evaluations. That is why, in this general context, it has been decided to build metamodels for the computer models involved. Specifically, artificial neural networks (see e.g. chapter 4 of [Mit97]) have been used extensively and we have investigated Kriging models later. The principle for address-

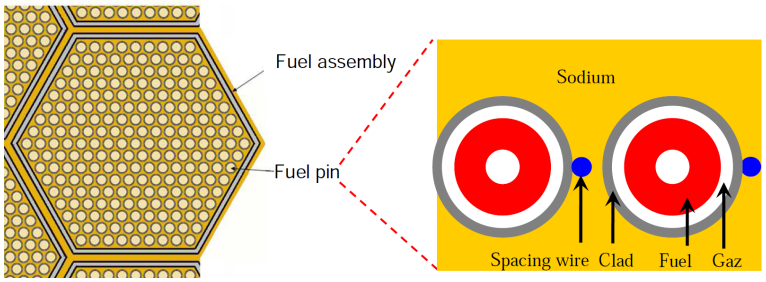

Figure 9.1: A schematic representation (from above) of a nuclear fuel pin and of a fuel assembly. A fuel assembly is composed of an aggregation of fuel pins, which consist in hollow fuel cylinders, surrounded by protective clads. A gas-filled gap exists between a clad and a fuel cylinder. Notice that the fuel pins are separated by spacing wires in a fuel assembly.

ing the multi-objective optimization problem is thus to carry out multi-objective optimization algorithms (notably genetic algorithms) on the metamodel functions, since it would be computationally prohibitive to do so on the computer models. The metamodels are then improved iteratively, in the potentially interesting input areas obtained from the genetic algorithms, by carrying out more code evaluations. The reader may refer to [HGA+10] for more details on this general core design optimization context.

In this chapter 9, we focus on the practical problem of building a Kriging model of the computer model GERMINAL. We are not specifically oriented toward optimization; instead the Kriging model follows the general objective of a small mean prediction error, over the domain of interest for the inputs of the GERMINAL code.

9.2.2 Inputs and outputs considered

The GERMINAL code has here 12 inputs, that we denote $x_1, ..., x_{12}$, and that are as follows.

- x_1 and x_2 concern the time aspect of the exploitation of the fuel pin. x_1, also denoted "l_cycle" is the cycle length. It is the time period between two maintenances of the fuel pin. x_2, also denoted "nb_c" is the number of cycle for the GERMINAL simulation.

- $x_3, ..., x_9$ concern the nature of the fuel pin. x_3 ("Pu") is the plutonium concentration. x_4 ("DiamHoll_mm") is the diameter of the shadow of the fuel pin. x_5("DiamExtClad_mm") is the diameter of the protective clad of the fuel pin. x_6 ("T_gap_mm") is the thickness of the gap between the fuel and the protective clad. x_7 ("T_Clad_mm") is the thickness of the protective clad. We refer to figure 9.1 for a visualization of $x_4, ..., x_7$. Finally, x_8 ("h_fiss_cm") is the height of the fuel pin.

- x_9 ("Plmean_W_cm"), x_{10} ("Fz") and x_{11} ("ampl_var") characterize the power map in the fuel pin. Notice that, in the multi-physics coupling presented in subsection 9.2.1, x_9, x_{10} and x_{11} are not fixed by the user. Instead, they are the output of calculations obtained from other computer models.

- x_{12}, also denoted "VD_cm3" is the disposal volume for the fission gas produced in the fuel pin (this disposal volume is located at the two extremities of the fuel pin).

The first output, denoted Y_1 or "T_Core_0.5D", and called the initial temperature, is the maximum temperature in the fuel pin at the initial time of the calculation. Because the temporal aspect of the simulation is absent in the computation of Y_1, the dependence of Y_1 with respect to $x_1, ..., x_{12}$ is rather simple, which will result, as we will see, in particularly good metamodel prediction results.

The second output, denoted Y_2 or "Fusion_Margin", and called the fusion margin, is the difference between the fusion temperature of the fuel and the maximum temperature, in space and time, of the fuel during the simulation. We make the following comments on this second output.

- Fusion is a highly undesirable phenomenon for the fuel pin. Therefore, a positive Y_2 indicates that, at least from the point of view of fusion, the fuel pin had a normal behavior during the simulation. On the contrary, a negative Y_2 is considered as an accident. We can hence notice that, in the general multi-objective optimization problem mentioned in subsection 9.2.1, Y_2 is a criterion that should be maximized.

- Y_2 is an output that uses all the temporal aspects of a GERMINAL simulation. Furthermore, since it is defined as a maximum in time, it is not expected to be very regular with respect to the inputs $x_1, ..., x_{12}$. Thus, the metamodeling task is more difficult for the output Y_2 than for the output Y_1.

- The GERMINAL computer model is not designed to simulate phenomena where a significant proportion of the fuel melts down. Furthermore, the protective clad is not meant to be impacted. Hence, little credit should be given to the values of strongly negative Y_2 obtained from GERMINAL. Indeed, a strongly negative fusion margin implies that the fuel temperature exceeds the fusion temperature for a significant proportion of the fuel. Furthermore, the clad may be impacted. Hence, the pin has reached a state that the GERMINAL code is not meant to simulate. This is a strong additional difficulty for the metamodelization problem. Hence, it was decided, in the general multi-objective optimization context of subsection 9.2.1, to filter out the input points yielding strongly negative Y_2 in the data bases. Similarly, the obtained metamodel is expected to be precise only for new input points that do not yield strongly negative Y_2. In this chapter 9, we follow this approach, and we work with learning and validation samples that have been filtered.

9.2.3 Setting for the Kriging model

We consider a simple Kriging framework (see chapter 2). Indeed, we also investigated a universal Kriging framework, with an affine mean function, and essentially obtained the same results as in the simple Kriging framework.

Parameterization of the covariance function

We have noticed that there are some numerical instabilities in the GERMINAL code. These instabilities create pairs of inputs that are very close to one another, but that yield non-negligibly distant outputs. We address this instability by introducing a nugget effect in the Kriging model. More precisely, we model the GERMINAL output function Y (Y_1 or Y_2) by a Gaussian process of the form

$$Y = Y_c + Y_n. \tag{9.1}$$

In (9.1), Y_c is the continuous component of Y. It is a centered Gaussian process with isotropic Matérn $\frac{3}{2}$ covariance function of the form $\sigma^2 R_{\boldsymbol{\ell}}$, with $\sigma^2 R_{\boldsymbol{\ell}}$ defined by table 2.1 and (2.6). σ^2 and $\boldsymbol{\ell}$ are the variance and correlation length hyper-parameters that are estimated from data. In (9.1), Y_n is the nugget component. It is a centered Gaussian process with covariance function

$$K_n(\boldsymbol{x}, \boldsymbol{y}) = \sigma_n^2 \mathbf{1}_{\boldsymbol{x}=\boldsymbol{y}}.$$

The incorporation of the nugget component does not contradict the fact that the GERMINAL code is deterministic, and explains its very small scale numerical discontinuities. σ_n^2 is also a hyper-parameter that is estimated from data.

We carry out the estimation by Maximum Likelihood (3.6). From a practical point of view, it is interesting to use an alternative parameterization of the covariance function of Y in (9.1). Denoting $\alpha = \frac{\sigma_n^2}{\sigma^2}$, the covariance function of Y is

$$K_{\sigma^2,\boldsymbol{\ell},\alpha}(\boldsymbol{x}, \boldsymbol{y}) = \sigma^2 \left(R_{\boldsymbol{\ell}}(\boldsymbol{x}, \boldsymbol{y}) + \alpha \mathbf{1}_{\boldsymbol{x}=\boldsymbol{y}} \right). \tag{9.2}$$

(9.2) enables to use proposition 3.21 and thus to gain one dimension in the numerical optimization problem.

Learning and test bases

Let Y denote one of the two outputs Y_1, Y_2. We possess a learning base $\boldsymbol{x}^{(l,1)}, y_{l,1}, ..., \boldsymbol{x}^{(l,n_l)}, y_{l,n_l}$ and a test base $\boldsymbol{x}^{(t,1)}, y_{t,1}, ..., \boldsymbol{x}^{(t,n_t)}, y_{t,n_t}$, where $y_{l,i} = Y(\boldsymbol{x}^{(l,i)})$ and $y_{t,i} = Y(\boldsymbol{x}^{(t,i)})$. We carry out the Maximum Likelihood estimation on the learning base, and the points of the learning base are also the observation points from which the Kriging model is built in (2.9) and (2.10) (the support points).

We consider two criteria on the test base. The first one is the Root Mean Square Error, with $\hat{y}(\boldsymbol{x})$ the Kriging prediction at $\boldsymbol{x}$,

$$RMSE^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} \left(\hat{y}(\boldsymbol{x}^{(t,i)}) - Y(\boldsymbol{x}^{(t,i)}) \right)^2. \tag{9.3}$$

The second one is the 90% Confidence Intervals Ratio and is, with $\hat{\sigma}^2(\boldsymbol{x})$ the Kriging predictive variance at $\boldsymbol{x}$,

$$CIR = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{1}_{|\hat{y}(\boldsymbol{x}^{(t,i)}) - Y(\boldsymbol{x}^{(t,i)})| \leq 1.64 \hat{\sigma}(\boldsymbol{x}^{(t,i)})}. \tag{9.4}$$

The CIR criterion should be close to 0.9.

The third criterion is the Mean Square Normalized Error and is

$$MSNE = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{\left(\hat{y}(\boldsymbol{x}^{(t,i)}) - Y(\boldsymbol{x}^{(t,i)})\right)^2}{\hat{\sigma}^2(\boldsymbol{x}^{(t,i)})}. \tag{9.5}$$

The MSNE criterion should be close to 1.

Remark 9.1. *A GERMINAL calculation takes approximately one minute. Thus, an important point is that, in the GERMINAL application, the sizes of the learning and test bases are large. More precisely, for the output Y_1, the learning and test bases have* 15722 *and* 6521 *elements. For the output Y_2, they have* 3807 *and* 1613 *elements (because of the filtering). These kinds of base sizes start to be computationally problematic for ML. In our case, we have used a random subsample of the learning base, of size* 1000*, to compute the ML estimator. This method is not optimal, and in fact there exists several methods in the literature to address ML for Kriging with large data sets. We refer, e.g., to the references [ACW12], [SCA12] and [SCA13] that both provide competitive methods for addressing very large data sets, and a short review of other existing methods. In a private communication with us, Michael Stein also recommends the utilization of a simple Likelihood approximation ([Vec88]), consisting for example in partitioning the observations into contiguous blocks of size, say,* 1000*, and in minimizing the sum of the different likelihood criteria of proposition 3.21, corresponding to the different blocks. Indeed, this solution is almost as simple to implement as the solution we used, and the computation time is only* 16 *times larger, when it uses e.g. all the* 15722 *available observations for estimation.*

Concerning the Kriging prediction, which still requires to carry out a matrix inversion, we have used 7000 *support points for Y_1 (including the* 1000 *points that are used for the ML estimation) and* 3807 *support points for Y_2.*

Remark 9.2. *We normalize linearly each of the* 12 *inputs between* 0 *and* 1*. Therefore, in subsection 9.3.1, the orders of magnitude of the estimated correlation lengths should be compared with* 1*.*

Normalized Leave-One-Out errors

For the case of the fusion margin output Y_2, we have seen in subsection 9.2.2 that the learning and test bases are filtered, because a GERMINAL calculation can result in a computation failure when addressing negative fusion margin phenomena. Despite this filtering, the learning base may still contain some observation points that actually correspond to computation failures. These possible computation failures can not be studied manually for all the observation points of the learning base. We show here that the computation of the Leave-One-Out errors and predictive variances can be an automatic method to exhibit observation points that are likely to be computation failures. Hence, a small number of observation points with high LOO errors, compared to the LOO predictive variances, can have their GERMINAL calculations verified manually.

We study the normalized LOO criterion, for all the observation points $\boldsymbol{x}^{(l,1)}, y_{l,1}, ..., \boldsymbol{x}^{(l,n_l)}, y_{l,n_l}$, that is

$$\epsilon_{n,LOO,i} = \frac{\hat{y}_{l,i} - y_{l,i}}{\hat{\sigma}_{l,i}}, \tag{9.6}$$

$\sigma(^\circ)$	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5	ℓ_6	ℓ_7	ℓ_8	ℓ_9	ℓ_{10}	ℓ_{11}	ℓ_{12}	α
890	100	100	100	6.1	9.9	17	100	35	5.0	13	13	100	7.6×10^{-5}

Table 9.1: Estimated hyper-parameters $(\sigma, \boldsymbol{\ell}, \alpha)$ for the output Y_1 ("T_Core_0.5D") of the GERMINAL code.

where $\hat{y}_{l,i}$ and $\hat{\sigma}^2_{l,i}$ are the Kriging LOO prediction and predictive variances of $y_{l,i}$ given $y_{l,1}, ..., y_{l,i-1}, y_{l,i+1}, ..., y_{l,n}$ (see subsection 2.2.4). The hyper-parameters $\sigma^2, \boldsymbol{\ell}, \alpha$ are kept to their ML estimate of table 9.2. Notice that the computation of all the LOO errors and predictive variances in (9.6) is fast, thanks to proposition 2.35.

The $\epsilon_{n,LOO,i}$ follow standard Gaussian distributions under the Kriging model (notice, though, that they are not independent). We sort the $|\epsilon_{n,LOO,i}|$, $1 \leq i \leq l_n$ by decreasing order, and the principle is to manually investigate the observation points corresponding to the largest values.

9.3 Results of the Kriging model

9.3.1 Interpretation of the estimated covariance hyper-parameters

For the initial temperature output Y_1 ("T_Core_0.5D"), the estimated hyper-parameters are given in table 9.1. Let us first consider the nugget effect. The standard deviation of the nugget process in (9.1) is $\sqrt{894^2 \times 7.61 \times 10^{-5}} = 7.8^\circ$. This value is coherent with the numerical behavior of the computer model GERMINAL. Furthermore, we will see below that the RMSE criterion is around 9° for Y_1. Thus, this RMSE is essentially composed of the standard deviation of the nugget process. This is intuitive, because the output Y_1 has a rather simple relationship with respect to the inputs, so that when the number of learning points is large, as it is the case here, the continuous component $Y_{c,1}$ of Y_1 in (9.1) is almost perfectly predicted, but the nugget component $Y_{n,2}$ cannot be predicted.

Let us now discuss the correlation lengths. They are found by the experts to make physical sense. For example, it is known that, for our learning base, the input x_2 ("nb_c") is actually truncated by the GERMINAL code to the same integer value for all the learning and test points, so that it has a zero impact on the outputs Y_1 and Y_2. We confirm this fact with the Kriging model, because the estimated correlation length is $\ell_2 = 100$ (for normalized inputs in $[0, 1]$). The smallest estimated correlation length, which intuitively corresponds to a very influent input, is ℓ_9, for "Plmean_W_cm". This is also a fact that is anticipated by the physicists. Indeed, "Plmean_W_cm" has a strong direct influence on the power map in the fuel pin, which is basically related to the temperature in the fuel pin and so to Y_1. Similarly, the inputs x_{10} and x_{11} impact the power map, so that their estimated correlation lengths are not large. The inputs x_4 ("DiamHoll_mm") and x_5 ("DiamExtClad_mm"), characterizing the geometry of the fuel pin are also known to have a strong impact on Y_1.

On table 9.2, we show the equivalent of table 9.1, but for the fusion margin output Y_2. The standard deviation of the nugget process is $\sqrt{1469^2 \times 3.73 \times 10^{-4}} = 28^\circ$. This value also makes sense from a numerical point of view, since the computation of Y_2 involves the temporal aspect, and is thus less stable than for Y_1. Concerning the estimated correlation lengths, we still have

$\sigma(^\circ)$	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5	ℓ_6	ℓ_7	ℓ_8	ℓ_9	ℓ_{10}	ℓ_{11}	ℓ_{12}	α
1470	25	100	68	18	5.1	17	100	55	2.4	7.4	6.2	100	3.7×10^{-4}

Table 9.2: Estimated hyper-parameters $(\sigma, \boldsymbol{\ell}, \alpha)$ for the output Y_2 ("Fusion_Margin") of the GERMINAL code.

$\ell_2 = 100$, for the input x_2 ("nb_c") that has a zero influence. The input x_9 ("Plmean_W_cm") remains the most influent. Overall, the hierarchy of the influences of the different inputs remains the same between Y_1 and Y_2. Finally, we observe that the estimated correlation lengths are globally smaller for Y_2 than for Y_1. This is intuitive, because smaller correlation lengths correspond to Gaussian processes that are predicted with more difficulty, which is the case for Y_2 compared to Y_1 as we have discussed in subsection 9.2.2.

We now illustrate the estimated hyper-parameters, and especially the nugget effect of value 28°. We choose two 12-dimensional input points $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$, for which the output Y_2 is positive and negative. We then evaluate the GERMINAL code on 97 points in the segment joining the two points $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$. This enables us to consider a one-dimensional subfunction of the 12-dimensional GERMINAL code, which is useful for plotting and interpreting the Kriging predictions.

On figure 9.2, we plot the 97 observation points of the segment, and the Kriging prediction and 90% confidence intervals (for a Kriging model using only the 97 observation points as support point). The estimated hyper-parameters of the Kriging model are those of table 9.2. We observe that there is indeed a numerical instability, which can be represented by a nugget effect with standard deviation 28°. We also observe that the Kriging model appears to be appropriate. More specifically, it interpolates the observations in the areas where there is no numerical instability, and it does not interpolate the observations in the areas where there is a numerical instability. The confidence intervals appear to be of the right order of magnitude, and their size is almost constant, because of the considerable value of the standard deviation of the nugget process in (9.1). Finally, the numerical instability is stronger when Y_2 is negative, especially there is an outlier observation for which $Y_2 = -500$. Because of the nugget effect, and because of the relatively large correlation lengths, the Kriging prediction is not too much impacted by this outlier point.

This is as previously discussed in subsection 9.2.2.

9.3.2 Prediction results

The prediction results for the output Y_1 ("T_Core_0.5D") are given in table 9.3. The standard deviation of the output on the test base is 344°, and the RMSE (9.3) criterion for the Kriging prediction is 9.03°. Thus, the Kriging prediction has a 3% relative error, which confirms, as mentioned in subsection 9.2.2 that the output Y_1 is a rather simple function of the inputs. We recall, from table 9.1, that the estimated nugget standard deviation σ_n is (9.1) is 7.8°. Hence, we see that the most part of the prediction error comes from the numerical instability of the GERMINAL calculations.

In table 9.3, we also see that the Kriging predictive variances have appropriate orders of

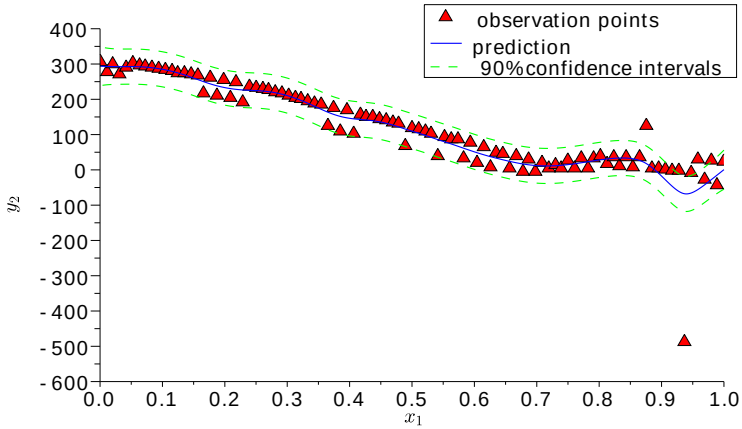

Figure 9.2: One-dimensional plot of the Kriging prediction for the output Y_2 of the GERMINAL code. 97 observation points are taken on a segment joining two 12-dimensional observation points on the input space of the $x_1, ..., x_{12}$. We index the points on the segment by their x_1 component (x-axis in the plot). The Kriging model is built with the estimated hyper-parameters of table 9.2 and its support points are the 97 observation points. We observe a general numerical instability which justifies the presence of the nugget effect. We also observe an outlier GERMINAL calculation point. Because of the nugget effect, and because of the relatively large correlation lengths, the Kriging prediction is not too much impacted by this outlier point.

	$RMSE$	CIR	$MSNE$
Kriging	9.03°	0.92	0.84
Neural networks	12.2°		

Table 9.3: Prediction results for the output Y_1 ("T_Core_0.5D") of the GERMINAL code. The standard deviation of the output on the test base is 344°.

magnitude. Indeed, the criterion CIR is relatively close to 0.9 and the criterion $MSNE$ is relatively close to 1.

We have compared the Kriging RMSE with the RMSE obtained from an artificial neural network. The artificial neural network method is the one used as a metamodel method in the general optimization problem of subsection 9.2.1 and of [HGA+10]. The artificial neural network method is implemented in the URANIE uncertainty platform, developed at CEA. In a few words, the artificial neural networks have one hidden layer and the activation function is a hyperbolic tangent. For a given number of hidden layer neurons, the weights of the artificial neural network are selected by using an early stopping algorithm. More precisely, this early stopping algorithm splits the learning base into two subbases. It carries out a gradient descent method for optimizing the weights of the artificial neural network, based on the data for the first subbase, but stops the gradient descent method earlier than at its convergence, when the obtained prediction error on the second subbase starts to increase. Note that this method has a random component, due to a random split of the learning base and a random initialization of the weights in the gradient descent method.

When the RMSE value obtained from the artificial neural network method is presented in table 9.3, a loop is actually carried out over the number of hidden layer neurons (from 15 to 30, by a step of 3). For each number of hidden layer neurons, the weights are optimized twice on the learning base, with the (random) early stopping method presented above. The RMSE value presented is then that of the artificial neural network, characterized by the number of hidden neurons and the weights, maximizing a score, on the test base, involving several prediction error criteria, including RMSE, the mean absolute error and the maximum absolute error. Hence, notice that the artificial neural network for which the RMSE value is presented in table 9.3 has actually been computed using knowledge on the test base. This is an advantage given to the artificial neural network method, in this comparison, because the Kriging model is only built from the learning base.

On table 9.3, we observe that, despite this advantage given to the artificial neural network method, the Kriging RMSE is smaller than that of the artificial neural network method (9.03° compared to 12.2°).

The prediction results for the output Y_2 ("Fusion_Margin") are given in table 9.4. The standard deviation of the output on the test base is 342°, and the RMSE (9.3) criterion for the Kriging prediction is 35.9°. The Kriging relative error is around 10%. For Y_1 this relative error is 3%. Hence, we have a confirmation that the output Y_2 is a more complex function of the inputs than the output Y_1. The reasons are given in subsection 9.2.2: the output Y_1 only involves the initial state of a GERMINAL simulation, while Y_2 involves the simulation during

	$RMSE$	CIR	$MSNE$
Kriging	35.9°	0.89	1.03
Neural networks	39.7°		

Table 9.4: Prediction results for the output Y_2 ("Fusion_Margin") of the GERMINAL code. The standard deviation of the output on the test base is 342°.

the whole time period. We recall, from table 9.2, that the estimated nugget standard deviation σ_n in (9.1) is 28°. Hence, similarly to Y_1, we see that an important part of the prediction error comes from this nugget effect.

Concerning the Kriging predictive variances, from CIR and MSNE in table 9.4 we see that their order of magnitudes are appropriate, similarly to table 9.3 for Y_1.

Finally, similarly to table 9.3, the RMSE of the artificial neural network method is larger than that of Kriging (39.7° compared to 35.9°). The artificial neural network is built by using the same method as in 9.3.

As a conclusion, we have seen that a standard Kriging model (stationary Matérn $\frac{3}{2}$ covariance function) gives good prediction results, compared to artificial neural network methods.

Notice that we have also confirmed other general facts, in the comparison between Kriging and artificial neural network methods, that result from the intrinsic difference between the two methods. This difference is that a Kriging metamodel function explicitly uses the data points each time it is called for a new point, while the artificial neural network function, after being built from the data base, does not use it when being called for new points. As a result, the inline computation time may be larger for Kriging. By inline computation time, we mean the computation time required for using the metamodel for a large number of new points, after it has been built from the learning base. For Kriging, the standard prediction (2.9) at a new point $\boldsymbol{x}^{(new)}$ requires to loop over all the learning points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$. Thus, this standard Kriging prediction method has a $O(n)$ computational cost. Notice that there exists alternative to this $O(n)$ prediction method, such as screening methods. On the contrary, once the neural network metamodel is built from the learning base, its complexity only scales with the number of hidden layer neurons, which is generally much smaller than the number of data points. In the case of the GERMINAL computer model, we have a confirmation that the inline computation time is significantly larger for the Kriging metamodel than for the artificial neural network metamodel. On the other hand, because the Kriging metamodel function explicitly uses the data points, it is ensured that its prediction error on the data base is small and only caused by the nugget effect. This has been a comforting fact, from the code user point of view, in the GERMINAL application case.

For the Kriging model, the estimation of the nugget effect is important in the case of the GERMINAL model, because we see that the value of this effect explains a large part of the prediction error. In figure 9.2, we also see that it would make no sense to interpolate the observed values exactly. Furthermore, the Kriging provides a prediction of its prediction error, that has been shown to be accurate. As an example of an utilization of this anticipation of the prediction error, we now see an example of automatic detection of computation failures in

subsection 9.3.3.

9.3.3 Detection of computation failures for the "Fusion_Margin" output

We consider the detection of computation failures for the output Y_2, since we have seen that its computation is the most subject to numerical instability. The 10 largest normalized LOO errors of (9.6) are

$$\begin{pmatrix} 14.2 \\ 7.8 \\ -4.4 \\ 4.4 \\ 4.4 \\ 3.9 \\ 3.8 \\ 3.7 \\ -3.6 \\ 3.6 \end{pmatrix}. \tag{9.7}$$

In (9.7), we see the two largest normalized LOO errors 14.2 and 7.8 as particularly large, compared to a standard Gaussian distribution and compared to the eight remaining ones. We then investigate them particularly. The next errors are large as well, but their investigation is less of a priority.

Let us also notice that abnormally large LOO errors are not necessarily caused by a computation failure. They can be the consequence of a Kriging model that is not perfectly adapted. For example, we have classically used a stationary covariance function. This is a rather strong assumption, and can result in overoptimistic predictive variances, in areas of the input space where the variations of the output are much more important than in the rest of the input space.

As a potential confirmation of the limits of the Kriging model treated here, we have also investigated the third largest LOO error in (9.7). Contrary to what we will see below for the two largest ones, the physicists have not found indicators of computational problem in the associated GERMINAL simulation. It can only be noted that the input point for this output is rather marginal in the input domain, so that it can correspond to an area of the input domain that has not been sufficiently explored. Nevertheless, the third largest LOO error in (9.7) is, unless shown otherwise, an observation point that is badly predicted by the Kriging model.

When investigating the two largest normalized LOO errors in (9.7), we see that their GERMINAL fusion margin values are 217° and 211°. This means that the two GERMINAL simulations predict that there is no fusion phenomenon for these two points. Instead, the two Kriging predictions are −304° and −182°, so that the Kriging model predicts a fusion phenomenon.

As a standard component of physical research process, the GERMINAL code has been updated since we carried out the study of this chapter 9. Note that some aspects of these updates were directly motivated by the numerical instabilities that we exhibited in figure 9.2, and by the points with high normalized LOO errors that we pointed out. As a consequence, a later version of the GERMINAL code predicts the fusion margin output Y_2, for the two largest normalized

LOO errors in (9.7), at -251 and -171. Hence the differences between the Kriging predictions and the new GERMINAL values are now explained by the predictive variances. Furthermore, the new GERMINAL calculations do confirm that fusion phenomena took place.

To summarize, a Kriging model has been carried out on a data base corresponding to a given version of the GERMINAL code. The normalized LOO errors have been sorted by decreasing order, and two of them are significantly larger than the other ones. The interest of this automatic outlier selection is that it is prohibitive to investigate each GERMINAL calculation manually. It has been shown that the two LOO outliers do correspond to computational failures. Furthermore, a later version of the GERMINAL code yield two new values for them, that are close to the Kriging predictions.

Chapter 10

Conclusion and perspectives

On the interest of Gaussian process models for the analysis of computer experiments

In this thesis, we have confirmed the strong interest of Gaussian process models for the analysis of computer experiments. Indeed, from their intrinsic ability to approximate a deterministic function and to associate a probability distribution to the resulting error, they can be used to address a large variety of problems. In chapter 7, we have considered two rather different frameworks for the analysis of the discrepancies between a computer model and a set of experimental results. In the first framework, the discrepancies are explained by an intrinsic variability of the computer model, while in the second one, they are explained by a model error function. This function is represented by a realization of a Gaussian process. In these two frameworks, the treatment is possible when the computer model function is not expensive to run, or when a linear approximation of it is carried out. In the remaining case, the treatment can notably be made possible by building a Gaussian process model of the computer code.

We have focused on the Gaussian process modeling of the model error, in the case where a linear approximation of the code with respect to its model parameters is done. This method has the advantage of being simple, and we have seen in chapter 8 that the resulting prediction of the physical system is similar to that of the non-linear method, even if the computer model is actually non-linear. Indeed, in this case, the model error function is defined with respect to the linearized model. We have also seen in chapter 8 that the prediction is composed of the calibrated computer model, completed by a Gaussian process inference of the model error. This complementarity between the physical model and the statistical model yields a prediction that is significantly more precise than the one of the calibrated code only.

We have also highlighted, in chapter 9 on the GERMINAL thermo-mechanical code, the accuracy of the Gaussian process meta-modeling of a computer model, even in a relatively high-dimensional case (12 input variables). The predictive variance is an additional benefit of Gaussian process models, that has been illustrated in the case of automatic computation failure detection.

An analysis of Maximum Likelihood and Cross Validation for covariance hyper-parameter estimation

Another central point of the thesis is the comparison of Maximum Likelihood and Cross Validation, for the estimation of the covariance function of a Gaussian process.

We have confirmed in chapter 5 that Maximum Likelihood is preferable over Cross Validation, in the well-specified case where the true covariance function of the Gaussian process does belong to the parametric family used for estimation. This conclusion holds in an expansion-domain asymptotic context, and is independent of the design of experiments. We have also shown that Cross Validation has the same rate of convergence as Maximum Likelihood, so that, in the well-specified case, using it instead of Maximum Likelihood is sub-optimal but not too prejudicial.

In chapter 6, we have addressed the misspecified case, that is the case where the true covariance function of the Gaussian process does not belong to the parametric family of functions used for estimation. We have shown that, when the Design Of Experiments is not too regular, Cross Validation is preferable to Maximum Likelihood. Indeed, it is more robust to the misspecification of the covariance function family, in the sense that it has a smaller bias than Maximum Likelihood. We interpret this by the fact that, in the misspecified case, it is not possible to estimate an hyper-parameter yielding Kriging conditional distributions that are good in all aspects (mean, variance, quantiles). The Maximum Likelihood estimator tries, in nature, to do so. On the contrary, the Cross Validation estimator is goal-oriented, and only addresses the punctual conditional means and the predictions of the associated prediction errors. This enables it to obtain better results, in terms of mean square prediction error and of predictive variance reliability, than Maximum Likelihood.

A joint conclusion of chapters 5 and 6 is that we have found that covariance function estimation generally benefits from an irregular sampling. In the well-specified case, where Maximum Likelihood is preferable, the asymptotic variance of the Maximum Likelihood estimator is smaller when using an irregular sampling. In the misspecified case, for Maximum Likelihood, we have not found a significant difference between an irregular or regular sampling. However, for Cross Validation, there is a significant degradation when using a regular sampling. Indeed, when having observation points that are on a regular grid, Cross Validation estimates covariance hyper-parameters adapted only to predictions on this regular grid. Because of the covariance function family misspecification, this does not generalize at all to predictions outside the regular grid. This results in a large bias for Cross Validation, in the misspecified case and when using a regular grid of observation points, as we have shown in chapter 6.

The fact that an irregular sampling is profitable to covariance function estimation has been noted in the literature ([Ste99], chapter 6.9, [ZZ06], [JDLI08]). This is opposed to the case of prediction with known covariance function, where regularly-spaced samplings appear as more efficient, as we confirm in chapter 5. The references [ZZ06] and [PM12] notice that using space-filling samplings, augmented with closely spaced observation points, yield efficient samplings for Kriging prediction with estimated hyper-parameters. The results of chapter 5 are in agreement with this conclusion.

Finally, we have observed in both the misspecified and well-specified cases that Cross Vali-

dation has a larger variance than Maximum Likelihood. This conclusion holds for all the kinds of samplings that we have investigated.

Other Cross Validation criteria in the literature

A natural perspective arises from our last conclusion: Cross Validation has a larger variance than Maximum Likelihood. Furthermore, the variance of Cross Validation can increase when the sampling becomes irregular. To interpret this fact, let us rewrite the Cross Validation criterion for clarity,

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{y}_{i,\boldsymbol{\theta}}\}^2, \tag{10.1}$$

where, for $1 \leq i \leq n$, $\hat{y}_{i,\boldsymbol{\theta}}$ is the prediction of y_i according to $y_1, ..., y_{i-1}, y_{i+1}, ..., y_n$, and according to the covariance function $\sigma^2 R_{\boldsymbol{\theta}}$. In (10.1), the LOO errors have heterogeneous variances when the sampling is irregular. More specifically, the LOO errors for observation points that are isolated have larger variances. This increases the variance of the Cross Validation estimator minimizing (10.1).

Another criterion, that could avoid this heterogeneity problem is ([RW06], chapter 5, [ZW10], [SK01]) the LOO log predictive probability,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\ln\left(\sigma^2\hat{c}^2_{i,\boldsymbol{\theta}}\right) + \frac{(y_i - \hat{y}_{i,\boldsymbol{\theta}})^2}{\sigma^2\hat{c}^2_{i,\boldsymbol{\theta}}}\right\}, \tag{10.2}$$

that is minimized jointly w.r.t σ^2 and $\boldsymbol{\theta}$. In (10.2), $\sigma^2\hat{c}^2_{i,\boldsymbol{\theta}}$ is the Kriging predictive variance of y_i according to $y_1, ..., y_{i-1}, y_{i+1}, ..., y_n$ and given the covariance function $\sigma^2 R_{\boldsymbol{\theta}}$. In (10.2), the n terms should have homogeneous variances, because the prediction error is divided by the predictive variance. Hence, the Cross Validation estimator corresponding to the criterion (10.2) could benefit from an asymptotic study, like the one of chapter 5. It is possible that, in the well-specified case, it yields a smaller asymptotic variance than the criterion (10.1). However, it is also possible that, in the framework of chapter 6, the criterion (10.2) be more sensitive to the covariance function misspecification. Indeed, similarly to Maximum Likelihood, it considers the full conditional distribution, and not only the conditional mean.

Designing new Cross Validation criteria

It may be interesting to design new Cross Validation criteria, in light of the trade-off we pointed out, between robustness to misspecification and small variance in the well-specified case. One possibility that arises, for numerical reasons, in chapter 6, is the penalization of large estimated variance $\hat{\sigma}^2_{LOO}$. Indeed, we have observed that this penalization indirectly prevents Cross Validation from estimating too large correlation lengths, which has been noticed as problematic for the Cross Validation method [MS04]. It can hence be an interesting direction of research to study a Cross Validation criterion of the form

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{y}_{i,\boldsymbol{\theta}}\}^2 + f_p\left(\hat{\sigma}^2_{LOO}(\boldsymbol{\theta})\right), \tag{10.3}$$

with

$$\hat{\sigma}^2_{LOO}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,\boldsymbol{\theta}})^2}{\hat{c}^2_{i,\boldsymbol{\theta}}},$$

and where f_p is an increasing penalty function. This function could also depend on data.

Another possibility for improving the Cross Validation criterion is to normalize the Cross Validation errors of (10.1), but with normalization parameters that are fixed in the optimization problem, contrary to (10.2). The correlation hyper-parameter $\boldsymbol{\theta}$ can hence be estimated by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,\boldsymbol{\theta}}(y_{-i}))^2}{v_i^2}, \tag{10.4}$$

where the v_i^2 are normalization terms, independent of $\boldsymbol{\theta}$. One possibility is to use a first-step estimation $\hat{\boldsymbol{\theta}}_1$, for example by Maximum Likelihood, to set the v_i^2 as functions of the Kriging LOO conditional variances, with hyper-parameter $\hat{\boldsymbol{\theta}}_1$, and then estimate $\boldsymbol{\theta}$ by minimizing (10.4). Notice that this normalization can make the variance of the Cross Validation estimator less sensitive to an irregular sampling, but it may not improve it in the framework of the regular grid in chapter 5. Indeed, the LOO errors are already asymptotically homogeneous in this framework. Thus, the principle of (10.4) can be extended, for instance by decorrelating the LOO errors, before minimizing their square mean.

Fixed-domain asymptotic analysis of Cross Validation

In this thesis, we have addressed the expansion-domain asymptotic framework when addressing the Cross Validation estimator in chapter 5. Indeed, as we have discussed, expansion-domain asymptotics enables us to state a general asymptotic normality result for Cross Validation, similarly to Maximum Likelihood [MM84]. Notably, these expansion-domain asymptotic results for Maximum Likelihood and Cross Validation hold for a large class of covariance function families. In the literature, the fixed-domain asymptotic results for Maximum Likelihood address particular families of covariance functions (for instance the exponential family in [Yin91] and [Yin93], or the isotropic Matérn family with fixed regularity parameter in [Zha04]). Furthermore, the fixed-domain asymptotic analysis yields a large variety of results, notably according to microergodicity (chapter 4, [Ste99], chapter 6.2) or non-microergodicity of the covariance hyper-parameters.

Hence, the similar process for the Cross Validation estimator minimizing (10.1), would be to address its fixed-domain asymptotic properties on particular covariance function families, now that a general expansion-asymptotic result is available in chapter 5. Historically, the first covariance structure for which fixed-domain asymptotic results were obtained for Maximum Likelihood is the exponential covariance structure ([Yin91], [Yin93]), due to its Markovian properties. It is hence a good candidate to start investigating the fixed-domain asymptotic properties of Cross Validation.

Appendix A

Notation

In general, scalars are written in italic, vectors are written in bold italic, and matrices are written in simple bold. In general, there is no font distinction between deterministic quantities, realizations of random quantities and random quantities.

Mathematical symbols

$\hat{f}$	The Fourier transform of a multidimensional function f
$\mathbf{I}_n$	The identity matrix of size n
$\mathbf{J}_n$	The matrix of size n whose coefficients are all 1
$Tr(\mathbf{M})$	The trace of a matrix $\mathbf{M}$
$\lvert\mathbf{M}\rvert$	The determinant of a matrix $\mathbf{M}$
$\lvert\lvert\mathbf{M}\rvert\rvert_2$	For a $n \times n$ matrix $\mathbf{M}$: $\sqrt{\frac{1}{n}\sum_{i,j=1}^{n} M_{i,j}^2}$
$\lvert\lvert\mathbf{M}\rvert\rvert$	The largest singular value of a matrix $\mathbf{M}$
$Diag(\mathbf{M})$	For a matrix $\mathbf{M}$: $(Diag(\mathbf{M}))_{i,j} = M_{i,j}\mathbf{1}_{i=j}$
$\lvert\boldsymbol{v}\rvert$	The Euclidean norm of a vector $\boldsymbol{v}$
$\lvert\boldsymbol{v}\rvert_\infty$	For a vector $\boldsymbol{v}$, $\lvert\boldsymbol{v}\rvert_\infty = \max_i \lvert v_i\rvert$
$\mathcal{N}(\boldsymbol{m}, \mathbf{K})$	Gaussian distribution with mean vector $\boldsymbol{m}$ and covariance matrix $\mathbf{K}$
$GP(m, K)$	Gaussian process with mean function m and covariance function K
$\mathcal{X}^2(n)$	The $\mathcal{X}^2$ distribution with n degrees of freedom
$\boldsymbol{e}^{(k)}$	The k-th base vector
Φ_{m,σ^2}	The cumulative distribution function of the Gaussian distribution with mean m and variance σ^2

Ordinal variables

n	Number of observation points
d	Dimension of the input space of a Gaussian process
	Number of experimental conditions for a physical system
p	Number of covariance hyper-parameters of a parametric covariance function family
m	Number of regression functions for an universal Kriging model
	Number of computer model parameters

Kriging

$\boldsymbol{x} = (x_1, ..., x_d)$	d-dimensional input of a Gaussian process, or of a physical system
Y	A real-valued Gaussian process
$\mathcal{D}$	The definition domain of a Gaussian process ($\mathcal{D} \subset \mathbb{R}^d$)
$\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$	n observation points in $\mathcal{D}$
$\mathcal{X}$	The observation set $\left\{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}\right\}$
$\mathbf{H}$	The $n \times m$ matrix of the regression functions at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
R	The correlation function of Y
$\mathbf{R}$	The $n \times n$ correlation matrix of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
K	The covariance function of Y
$\mathbf{K}$	The $n \times n$ covariance matrix of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
$\boldsymbol{y}$	The random vector of Y at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
$\boldsymbol{x}$	A prediction point in $\mathcal{D}$
$\hat{y}(\boldsymbol{x})$	The prediction (BLUP or conditional expectation) of Y at $\boldsymbol{x}$
$\hat{\sigma}^2(\boldsymbol{x})$	The predictive variance of Y at $\boldsymbol{x}$
$\boldsymbol{h}(\boldsymbol{x})$	The m-dimensional vector of the regression functions at $\boldsymbol{x}$
$\boldsymbol{r}(\boldsymbol{x})$	The n-dimensional correlation vector of Y between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$
$\boldsymbol{k}(\boldsymbol{x})$	The n-dimensional covariance vector of Y between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$

Misspecification of covariance function

R_1	The true correlation function of Y
K_1	The true covariance function of Y
R_2	The assumed correlation function of Y
K_2	The assumed covariance function of Y

For model i, $i = 1, 2$:

$\mathbf{R}_i$	The $n \times n$ correlation matrix at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ with R_i
$\mathbf{K}_i$	The $n \times n$ covariance matrix at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ with K_i
$\boldsymbol{r}_i(\boldsymbol{x})$	The $n \times 1$ correlation vector between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$ with R_i
$\boldsymbol{k}_i(\boldsymbol{x})$	The $n \times 1$ covariance vector between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$ with K_i
$\hat{y}_i(\boldsymbol{x})$	The prediction of Y at $\boldsymbol{x}$ with R_i
$\hat{\sigma}_i^2(\boldsymbol{x})$	The predictive variance of Y at $\boldsymbol{x}$ with K_i
$\mathbb{E}_i$, Var_i, Cov_i, $\sim_i$	Mean value, variance, covariance and distribution of a function of Y, when the covariance function of Y is K_i

Cross Validation

$\mathcal{S}$	A subset of $\mathcal{X}$
$\hat{y}_{\mathcal{S}}(\boldsymbol{x})$	The prediction of Y at $\boldsymbol{x}$ according to the observation data in $\mathcal{S}$
$\hat{\sigma}^2_{\mathcal{S}}(\boldsymbol{x})$	The predictive variance of Y at $\boldsymbol{x}$ according to the observation data in $\mathcal{S}$
$\hat{y}_{i}$	The prediction of Y at $\boldsymbol{x}^{(i)}$ according to the observation data in $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(i-1)}, \boldsymbol{x}^{(i+1)}, ..., \boldsymbol{x}^{(n)}$
$\hat{\sigma}^2_{i}$	The predictive variance of Y at $\boldsymbol{x}^{(i)}$ according to the observation data in $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(i-1)}, \boldsymbol{x}^{(i+1)}, ..., \boldsymbol{x}^{(n)}$

Parametric families of covariance functions

σ^2	A variance hyper-parameter
$\boldsymbol{\theta}$	A correlation hyper-parameter
$R_{\boldsymbol{\theta}}$	A correlation function
$\sigma^2 R_{\boldsymbol{\theta}}$	A covariance function
$\mathcal{R} = \{R_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$	A parametric set of correlation functions
$\mathcal{K} = \left\{\sigma^2 R_{\boldsymbol{\theta}}, \sigma^2 > 0, \boldsymbol{\theta} \in \Theta\right\}$	A parametric set of covariance functions
$\mathbf{R}_{\boldsymbol{\theta}}$	The $n \times n$ correlation matrix of $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ with correlation function $R_{\boldsymbol{\theta}}$
$\sigma^2 \mathbf{R}_{\boldsymbol{\theta}}$	The $n \times n$ covariance matrix of $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ with covariance function $\sigma^2 R_{\boldsymbol{\theta}}$
$\boldsymbol{\psi}$	$\boldsymbol{\psi} := (\sigma^2, \boldsymbol{\theta})$
$K_{\boldsymbol{\psi}}$	$K_{\boldsymbol{\psi}} := \sigma^2 R_{\boldsymbol{\theta}}$
$\mathcal{K} = \{K_{\boldsymbol{\psi}}, \boldsymbol{\psi} \in \Psi\}$	A parametric set of covariance functions
$\mathbf{K}_{\boldsymbol{\psi}}$	$\mathbf{K}_{\boldsymbol{\psi}} := \sigma^2 \mathbf{R}_{\boldsymbol{\theta}}$

Criteria for estimation by Maximum Likelihood

$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$	The modified opposite log-likelihood of the observations at $(\sigma^2, \boldsymbol{\theta})$
$\mathcal{L}(\boldsymbol{\theta})$	The marginal modified opposite log-likelihood $\min_{\boldsymbol{\beta},\sigma^2} L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$
$L_R(\sigma^2, \boldsymbol{\theta})$	The modified opposite restricted log-likelihood of the observations at $(\sigma^2, \boldsymbol{\theta})$
$\mathcal{L}_R(\boldsymbol{\theta})$	The marginal modified opposite restricted log-likelihood $\min_{\sigma^2} L_R(\sigma^2, \boldsymbol{\theta})$
$L(\boldsymbol{\psi})$	The modified opposite log-likelihood of the observations at $\boldsymbol{\psi}$
$L_R(\boldsymbol{\psi})$	The modified opposite restricted log-likelihood of the observations at $\boldsymbol{\psi}$

Maximum Likelihood estimators

$\hat{\boldsymbol{\theta}}_{ML}$	The Maximum Likelihood estimator of $\boldsymbol{\theta}$
$\hat{\sigma}^2_{ML}$	The Maximum Likelihood estimator of σ^2
$\hat{\boldsymbol{\psi}}_{ML}$	The Maximum Likelihood estimator of $\boldsymbol{\psi}$
$\hat{\boldsymbol{\theta}}_{REML}$	The REstricted Maximum Likelihood estimator of $\boldsymbol{\theta}$
$\hat{\sigma}^2_{REML}$	The REstricted Maximum Likelihood estimator of σ^2
$\hat{\boldsymbol{\psi}}_{REML}$	The REstricted Maximum Likelihood estimator of $\boldsymbol{\psi}$

Estimation by Cross Validation

$\hat{y}_{i,\boldsymbol{\theta}}$	The prediction of Y at $\boldsymbol{x}^{(i)}$ according to the observation data in $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(i-1)}, \boldsymbol{x}^{(i+1)}, ..., \boldsymbol{x}^{(n)}$, with correlation hyper-parameters $\boldsymbol{\theta}$
$\sigma^2 \hat{c}^2_{i,\boldsymbol{\theta}}$	The predictive variance of Y at $\boldsymbol{x}^{(i)}$ according to the observation data in $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(i-1)}, \boldsymbol{x}^{(i+1)}, ..., \boldsymbol{x}^{(n)}$, with covariance hyper-parameters $(\sigma^2, \boldsymbol{\theta})$
$LOO(\boldsymbol{\theta})$	The Leave-One-Out Mean Square Error of the observations at $\boldsymbol{\theta}$
$\hat{\sigma}^2_{LOO}(\boldsymbol{\theta})$	The Leave-One-Out estimation of σ^2 given $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}_{LOO}$	The Leave-One-Out estimation of $\boldsymbol{\theta}$
$\hat{\sigma}^2_{LOO}$	The LOO estimation of σ^2: $\hat{\sigma}^2_{LOO} := \hat{\sigma}^2_{LOO}(\hat{\boldsymbol{\theta}}_{LOO})$

Computer model and experiments

$\boldsymbol{x}$	d-dimensional vector of experimental conditions / inputs of the physical system
$\boldsymbol{\beta}$	m-dimensional vector of computer model parameters
$f_{mod}(\boldsymbol{x}, \boldsymbol{\beta})$	Computer model at $\boldsymbol{x}$ and $\boldsymbol{\beta}$
$f_{obs}(\boldsymbol{x})$	Physical system observed at $\boldsymbol{x}$
$f_{real}(\boldsymbol{x})$	Physical system at $\boldsymbol{x}$
$\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$	n observation points for the physical system
$\boldsymbol{\epsilon}$	n-dimensional vector of measure errors at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
$\boldsymbol{y}_{obs}$	n-dimensional vector of observations of the physical system at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
$\mathbf{H}$	$n \times m$ matrix of the derivatives of f_{mod} with respect to $\boldsymbol{\beta}$ at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
Y_{obs}	Gaussian process representation of the physical system
Z	Gaussian process of the model error
R_{mod}	The correlation function of Z
K_{mod}	The covariance function of Z
$\mathbf{R}_{mod}$	The $n \times n$ correlation matrix of Z at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
$\mathbf{K}_{mod}$	The $n \times n$ covariance matrix of Z at $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$
$\mathbf{K}_{mes}$	The $n \times n$ covariance matrix of $\boldsymbol{\epsilon}$
$\mathbf{K}$	$\mathbf{K} := \mathbf{K}_{mod} + \mathbf{K}_{mes}$
$\boldsymbol{r}(\boldsymbol{x})$	The n-dimensional correlation vector of Z between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$
$\boldsymbol{k}(\boldsymbol{x})$	The n-dimensional covariance vector of Z between $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ and $\boldsymbol{x}$
$\boldsymbol{h}(\boldsymbol{x})$	The m-dimensional vector of the derivatives of the computer model at $\boldsymbol{x}$

Abbreviations

BLUP	Best Linear Unbiased Predictor
GP	Gaussian Process
ML	Maximum Likelihood
REML	REstricted Maximum Likelihood
CV	Cross Validation
LOO	Leave-One-Out
MSE	Mean Square Error
RMSE	Root Mean Square Error
pdf	probability density function

Appendix B

Reference

Bibliography

[Abr97] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical report, Norwegian computing center, 1997.

[Abt99] M. Abt. Estimating the prediction mean squared error in gaussian stochastic processes with exponential correlation structure. Scandinavian Journal of Statistics, 26:563–578, 1999.

[AC12] I. Andrianakis and P. G. Challenor. The effect of the nugget on Gaussian process emulators of computer models. Computational Statistics and Data Analysis, 56:4215–4228, 2012.

[ACW12] M. Anitescu, J. Chen, and L. Wang. A matrix-free approach for solving the gaussian process maximum likelihood problem. SIAM Journal on Scientific Computing, 34(1):A240–A262, 2012.

[Adl81] R.J. Adler. The Geometry of Random Fields. Wiley, New York, 1981.

[Adl90] R.J. Adler. An introduction to continuity, extrema, and related topics for general Gaussian processes. Hayward, CA: Institute of mathematical statistics, 1990.

[And10] E. Anderes. On the consistent separation of scale and variance for Gaussian random fields. The Annals of Statistics, 38:870–893, 2010.

[AS65] M. Abramowitz and I. Stegun. Handbook of mathematical functions. Dover, New York, 1965. ninth ed.

[Aub00] J.P. Aubin. Applied Functional Analysis. Wiley-Interscience, New York, 2000.

[Bac] F. Bachoc. Asymptotic analysis of the role of the spatial sampling for hyperparameter estimation of Gaussian processes. Submitted to the Journal of Multivariate Analysis.

[Bac13] F. Bachoc. Cross validation and maximum likelihood estimations of hyperparameters of Gaussian processes with model mispecification. Computational Statistics and Data Analysis, 66:55–69, 2013.

[Bar90] S. Barnett. Matrices, Methods and Applications. Oxford Applied Mathematics and Computing Sciences Series, Clarendon Press, Oxford, 1990.

[Bar10] P. Barbillon. Méthodes d'interpolation à noyaux pour l'approximation de fonctions type boîte noire coûteuses. PhD thesis, Université Paris-Sud 11, 2010. Available at http://tel.archives-ouvertes.fr/tel-00559502/.

[BBGM] F. Bachoc, G. Bois, J. Garnier, and J.M Martinez. Calibration and improved prediction of computer models by universal Kriging. Nuclear Science and Engineering. In press.

[BBP+07] M. J. Bayarri, J.O. Berger, R. Paulo, J. Sacks, J.A. Cafeo, J. Cavendish, C.H. Lin, and J. Tu. A framework for validation of computer models. Technometrics, 49(2):138–154, 2007.

[BBV11] R. Benassi, J. Bect, and E. Vazquez. Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion. In LION5, online proceedings, Roma, Italy, 2011.

[BD62] D. Blackwell and L.E. Dubins. Merging of opinions with increasing information. Annals of Mathematical Statistics, 33:882–886, 1962.

[BD87] G. Box and N. Draper. Empirical Model Building and Response Surfaces. Wiley Series in Probability and Mathematical Statistics, 1987.

[Bet09] R. Bettinger. Inversion d'un système par Krigeage. Application à la synthèse de catalyseurs à haut débits. PhD thesis, Université de Nice-Sophia Antipolis, 2009. Available at http://hal.archives-ouvertes.fr/tel-00460162/.

[BGL+12] J Bect, D Ginsbourger, L Li, V Picheny, and E Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. Statistics and Computing, 22:773–793, 2012.

[Bil12] P. Billingsley. Probability and measure. Wiley, New York, 2012.

[BO08] L.S. Bastos and T. O'Hagan. Diagnostics for gaussian process emulators. Technometrics, 51:425–438, 2008.

[Cac03] D.G. Cacuci. Sensitivity and uncertainty analysis. Theory. Chapman & Hall/CRC, Boca Raton, FL, 2003.

[CD99] J.-P. Chilès and P. Delfiner. Geostatistics : Modeling Spatial Uncertainty. Wiley, New York, 1999.

[CG13a] C. Chevalier and D. Ginsbourger. Fast computation of the multipoint expected improvement with applications in batch selection. In Proceedings of the LION7 conference, Lecture Notes in Computer Science, 2013.

[CG13b] Clément Chevalier and David Ginsbourger. Corrected kriging update formulae for batch-sequential data assimilation. In To be presented at IAMG Madrid 2013, Session 4: Data assimilation in Geosciences, 2013.

[CL67] H. Cramér and M.R. Leadbetter. Stationary and Related Stochastic Processes - Sample Function Properties and Their Applications. Wiley, New-York, 1967.

[CL93] N. Cressie and S.N Lahiri. The asymptotic distribution of REML estimators. Journal of Multivariate Analysis, 45:217–233, 1993.

[Cre93] N. Cressie. Statistics for Spatial Data. Wiley, New York, 1993.

[CT06] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? Information Theory, IEEE Transactions on, 52(12):5406 –5425, 2006.

[Daq10] W. Daqing. Fixed Domain Asymptotics and Consistent Estimation for Gaussian Random Field Models in Spatial Statistics and Computer Experiments. PhD thesis, National University of Singapore, 2010.

[dC96] A. de Crécy. Determination of the uncertainties of the constitutive relationships in the cathare 2 code. In Proceedings of the 4th ASME/JSME International Conference on Nuclear Engineering, 1996.

[dC01] A. de Crécy. Determination of the uncertainties of the constitutive relationships of the cathare 2 code. In M&C 2001 Salt Lake City, Utah, USA, 2001.

[Die97] C. R. Dietrich. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. SIAM J. SCI. COMPUT., 18:1088–1107, 1997.

[Doo53] J. L. Doob. Stochastic Processes. Wiley, New-York, 1953.

[dRDT08] E. de Rocquigny, N. Devictor, and S. Tarantola. Uncertainty in industrial practice. Wiley, 2008.

[Dub83] O. Dubrule. Cross validation of Kriging in a unique neighborhood. Mathematical Geology, 15:687–699, 1983.

[dV96] A.W. Van der Vaart. Maximum likelihood estimation under a spatial sampling scheme. The Annals of Statistics, 24(5):2049–2057, 1996.

[DZM09] J. Du, H. Zhang, and V.S. Mandrekar. Fixed domain asymptotics properties of tapered maximum likelihood estimators. The Annals of Statistics, 37:3330–3361, 2009.

[Fu12] S. Fu. Inversion probabiliste bayésienne en analyse d'incertitude. PhD thesis, Université Paris-Sud 11, 2012. Available at http://tel.archives-ouvertes.fr/tel-00766341/.

[GBC+99] C. Gomez, C. Bunks, J.P. Chancelier, F. Delebecque, M. Goursat, R. Nikoukhah, and S. Steer. Engineering and Scientific Computing with Scilab. Birkhaüser, Boston, 1999.

[GG98] T. Gerstner and M. Griebel. Numerical integration using sparse grids. Numerical algorithms, 18:209–232, 1998.

[GG12] L. Le Gratiet and J. Garnier. Regularity dependence of the rate of convergence of the learning curve for gaussian process regression. 2012. Preprint.

[GL96] G.H. Golub and C.F. Van Loan. Matrix computations. Johns Hopkins Studies in Mathematical Sciences, Baltimore, 1996. 3rd edition.

[Gra01] R.M. Gray. Toeplitz and circulant matrices: A review. Technical report, 2001.

[GS74] I.I. Gihman and A.V. Skorohod. The theory of stochastic processes. Springer-Verlag, Berlin, 1974. vol.1.

[Har74] D.A. Harville. Bayesian inference for variant components using only error contrasts. Biometrika, 61:383–385, 1974.

[HGA+10] E. Hourcade, F. Gaudier, G. Arnaud, D. Funtowiez, and K. Ammar. A supercomputing application for reactors core design and optimization. In Joint International Conference on Supercomputing in Nuclear Applications and Monte Carlo 2010 (SNA + MC2010), Tokyo, Japan, October 17-21, 2010.

[HKC+04] D. Higdon, M. Kennedy, J.C. Cavendish, J.A. Cafeo, and R.D. Ryne. Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing, 26:448–466, 2004.

[HP04] L. Hascoët and V. Pascual. Tapenade 2.1 user's guide. Technical Report 0300, INRIA, 2004.

[HTF08] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. Springer, New York, 2008.

[IBFM10] B Iooss, L Boussouf, V Feuillard, and A Marrel. Numerical studies of the metamodel fitting and validation processes. International Journal of Advances in Systems and Measurements, 3:11–21, 2010.

[IR78] I.A. Ibragimov and Y.A. Rozanov. Gaussian Random Processes. Springer-Verlag, New York, 1978.

[JDLI08] N. Jeannée, Y. Desnoyers, F. Lamadie, and B. Iooss. Geostatistical sampling optimization of contaminated premises. In DEM 2008 - Decommissionning Challenges: an industrial reality?, Avignon, France, September 2008, 2008.

[JSW98] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black box functions. Journal of Global Optimization, 13:455–492, 1998.

[KHF+06] T. Kawano, K.M. Hanson, S. Frankle, P. Talou, M.B. Chadwick, and R.C. Little. Evaluation and propagation of the ^{239}Pu fission cross-section uncertainties using a Monte Carlo technique. Nuclear Science and Engineering, 153:1–7, 2006.

[KO01] M. Kennedy and A. O'Hagan. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63:425–464, 2001.

[Kou03] S. C. Kou. On the efficiency of selection criteria in spline regression. Probability Theory and Related Fields, 127:153–176, 2003.

[LA12] B. A. Lockwood and M. Anitescu. Gradient-enhanced universal kriging for uncertainty propagation. Nuclear Science and Engineering, 170:168–195, 2012.

[Lah03] S. N. Lahiri. Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. Sankhyã: The Indian Journal of Statistics, 65:356–388, 2003.

[LL00] W.L. Loh and T.K. Lam. Estimating structured correlation matrices in smooth Gaussian random field models. The Annals of Statistics, 28:880–904, 2000.

[LM04] S. N. Lahiri and K. Mukherjee. Asymptotic distributions of M-estimators in a spatial regression model under some fixed and stochastic spatial sampling designs. Annals of the Institute of Statistical Mathematics, 56:225–250, 2004.

[Loh05] W.L. Loh. Fixed domain asymptotics for a subclass of Matérn type Gaussian random fields. The Annals of Statistics, 33:2344–2394, 2005.

[LS05] R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in Gaussian Kriging model. Technometrics, 47:111–120, 2005.

[Mat70] G Matheron. La Théorie des Variables Régionalisées et ses Applications. Fasicule 5 in Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris, 1970.

[MIDV08] A. Marrel, B. Iooss, F. Van Dorpe, and E. Volkova. An efficient methodology for modeling complex computer codes with Gaussian processes. Computational Statistics and Data Analysis, 52:4731–4744, 2008.

[Mit97] T. M. Mitchell. Machine Learning. McGraw-Hill, New York, 1997.

[MM84] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71:135–146, 1984.

[MMGB12] J.M. Martinez, A. Marrel, N. Gilardi, and F. Bachoc. Krigeage par processus gaussiens Librairie gpLib. Technical report, CEA,DEN/DANS/DM2S/STMF/LGLS/RT/12-026/A, 2012.

[Mon05] D. C. Montgomery. Design and Analysis of Experiments. Wiley, New York, 2005. 6th edition.

[Mor91] M. D. Morris. Factorial sampling plans for preliminary computational experiments. Technometrics, pages 161–174, 1991.

[MRPT92] J.C. Melis, L. Roche, J.P. Piron, and J. Truffert. Germinal - a computer code for predicting fuel pin behaviour. Journal of Nuclear Materials, 188:303–307, 1992.

[MS04] J.D. Martin and T.W. Simpson. On the use of Kriging models to approximate deterministic computer models. In DETC'04 ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference Salt Lake City, Utah USA, September 28 - October 2, 2004.

[Nie92] H. Niederreiter. Random Number Generation and Quasi-Monte Carlo Methods. Series SIAM CBMS-NSF, SIAM, Philadelphia, 1992.

[Nou09] A. Nouy. Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. Archives of Computational Methods in Engineering, 16:251–285, 2009.

[NR96] E. Novak and K. Ritter. High dimensional integration of smooth functions over cubes. Numerisch Mathematik, 75:79–97, 1996.

[NW06] J. Nocedal and S.J. Wright. Numerical Optimization. Springer-Verlag, Berlin, 2006. 2nd edition.

[PCD08] A. Petruzzi, D.G. Cacuci, and F. D'auria. Best-estimate model calibration and prediction through experimental data assimilation-II: Application to a blowdown benchmark experiment. Nuclear Science and Engineering, 165:45–100, 2008.

[PM12] L. Pronzato and W. G. Müller. Design of computer experiments: space filling and beyond. Statistics and Computing, 22(3):681–701, 2012.

[PTVF07] W. H. Press, S. A. Teukolsky, W.T. Vetterling, and B.P. Flannery. Numerical recipes: The art of Scientific computing. Cambridge university press, 2007. 3rd edition.

[PY01] H. Putter and A. Young. On the effect of covariance function estimation on the accuracy of kriging predictors. Bernoulli, 7(3):421–438, 2001.

[RC99] C. Robert and G. Casella. Monte Carlo statistical methods. Springer-Verlag, New York, 1999.

[RGD12] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. Journal of Statistical Software, 51:1–55, 2012.

[Rip81] B.D Ripley. Spatial Statistics. Wiley, New York, 1981.

[Rob01] C. Robert. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer, New York, 2001.

[RW06] C.E. Rasmussen and C.K.I. Williams. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, 2006.

[SCA12] M.L. Stein, J. Chen, and M. Anitescu. Difference filter preconditioning for large covariance matrices. SIAM Journal on Matrix Analysis and Applications, 33(1):52–72, 2012.

[SCA13] M.L. Stein, J. Chen, and M. Anitescu. Stochastic approximation of score functions for gaussian processes. Annals of Applied Statistics, 7(2):1162–1191, 2013.

[SK01] S. Sundararajan and S.S. Keerthi. Predictive approaches for choosing hyperparameters in Gaussian processes. Neural Computation, 13:1103–1118, 2001.

[Smo63] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. Soviet Math. Dokl, 4:240–243, 1963.

[SS02] B. Schölkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization,Optimization, and Beyond. MIT Press, Cambridge, MA, 2002.

[Ste88] M.L. Stein. Asymptotically efficient prediction of a random field with a misspecified covariance function. The Annals of Statistics, 16:55–63, 1988.

[Ste90a] M.L. Stein. Bounds on the efficiency of linear predictions using an incorrect covariance function. The Annals of Statistics, 18:1116–1138, 1990.

[Ste90b] M.L. Stein. A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. The Annals of Statistics, 18:1139–1157, 1990.

[Ste90c] M.L. Stein. Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. The Annals of Statistics, 18:850–872, 1990.

[Ste93] M. L. Stein. Spline smoothing with an estimated order parameter. Annals of Statistics, 21:1522–1544, 1993.

[Ste99] M.L Stein. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York, 1999.

[STV04] B Schölkopf, K Tsuda, and J-P Vert. Kernel Methods in Computational Biology. MIT Press, Cambridge, MA, USA, 2004.

[Swe80] T.J. Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. The Annals of Statistics, 8:1375–1381, 1980.

[SWMW89] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. Statistical Science, 4:409–423, 1989.

[SWN03] T.J Santner, B.J Williams, and W.I Notz. The Design and Analysis of Computer Experiments. Springer, New York, 2003.

[Tar07] L. Tartar. An Introduction to Sobolev Spaces and Interpolation Spaces. Lecture Notes of the Unione Matematica Italiana, 2007.

[TBG+00] I. Toumia, A. Bergeron, D. Gallo, E. Royer, and D. Caruge. Flica-4: a three-dimensional two-phase flow computer code with advanced numerical methods for nuclear applications. Nuclear Engineering and Design, 200:139–155, 2000.

[Tyr96] E. E. Tyrtyshnikov. A unifying approach to some old and new theorems on distribution and clustering. Linear Algebra and its Applications, 232:1–43, 1996.

[Van98] A.W. Van der Vaart. Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[Vaz05] E. Vazquez. Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. PhD thesis, Université Paris XI Orsay, 2005. Available at http://tel.archives-ouvertes.fr/tel-00010199/en.

[VB10] E. Vazquez and J. Bect. Pointwise consistency of the Kriging predictor with known mean and covariance functions. In mODa 9 - Advances in Model-Oriented Design and Analysis. 14th-19th June 2010, Bertinoro, Italy, 2010.

[Vec88] A. V. Vecchia. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society , Ser. B, 50:297–312, 1988.

[VM12] S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. Annales de la faculté des sciences de Toulouse, 21(3):529–555, 2012.

[VVB10] F.A.C. Viana, G. Venter, and V. Balabanov. An algorithm for fast optimal latin hypercube design of experiments. International Journal for Numerical Methods in Engineering, 82:135–156, 2010.

[Wah90] G. Wahba. Spline Models for Observational Data. SIAM, Philadelphia, 1990.

[WCT09] S. Wang, W. Chen, and K-L. Tsui. Bayesian validation of computer models. Technometrics, 51:439–451, 2009.

[Whi82] H. White. Maximum likelihood estimation of misspecified models. Econometrica, 50(1):1–25, 1982.

[WLX13] W.Y. Wu, C.Y. Lim, and Y. Xiao. Tail estimation of the spectral density for a stationary gaussian random field. Journal of Multivariate Analysis, 116:74–91, 2013.

[Yin91] Z. Ying. Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. Journal of Multivariate Analysis, 36:280–296, 1991.

[Yin93] Z. Ying. Maximum likelihood estimation of parameters under a spatial sampling scheme. The Annals of Statistics, 21:1567–1590, 1993.

[YS85] S. J. Yakowitz and F. Szidarovszky. A comparison of Kriging with nonparametric regression methods. Journal of Multivariate Analysis, 16:21–53, 1985.

[ZC92] D.L. Zimmerman and N. Cressie. Mean squared prediction error in the spatial linear model with estimated covariance parameters. Ann. Inst. Statist. Math., 44:27–43, 1992.

[Zem65] A. H. Zemanian. Distribution Theory and Transform Analysis. McGraw-Hill, New York, 1965.

[Zha04] H Zhang. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. Journal of the American Statistical Association, 99:250–261, 2004.

[ZW10] H. Zhang and Y. Wang. Kriging and cross validation for massive spatial data. Environmetrics, 21:290–304, 2010.

[ZZ05] H. Zhang and D.L. Zimmerman. Toward reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92:921–936, 2005.

[ZZ06] Z. Zhu and H. Zhang. Spatial sampling design under the infill asymptotic framework. Environmetrics, 17:323–337, 2006.