# Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification

François Bachoc[a,b,*]

[a]*CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France*
[b]*Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII*
*Site Chevaleret, case 7012, 75205 Paris cedex 13*

**Abstract**

The Maximum Likelihood (ML) and Cross Validation (CV) methods for estimating covariance hyper-parameters are compared, in the context of Kriging with a misspecified covariance structure. A two-step approach is used. First, the case of the estimation of a single variance hyper-parameter is addressed, for which the fixed correlation function is misspecified. A predictive variance based quality criterion is introduced and a closed-form expression of this criterion is derived. It is shown that when the correlation function is misspecified, the CV does better compared to ML, while ML is optimal when the model is well-specified. In the second step, the results of the first step are extended to the case when the hyper-parameters of the correlation function are also estimated from data.

*Keywords:* Uncertainty quantification, metamodel, Kriging, hyper-parameter estimation, maximum likelihood, leave-one-out

## 1. Introduction

Kriging models (Matheron, 1970; Stein, 1999; Rasmussen and Williams, 2006) consist in interpolating the values of a Gaussian random field given observations at a finite set of observation points. They have become a popular method for a large range of applications, such as numerical code approximation (Sacks et al., 1989; Santner et al., 2003) and calibration (Paulo et al., 2012) or global optimization (Jones et al., 1998).

One of the main issues regarding Kriging is the choice of the covariance function. A Kriging model yields an unbiased predictor, with minimal variance and a correct predictive variance, if the covariance function used in the model coincides with that of the random field from which the observations stem. Significant results concerning the influence of a misspecified covariance function on Kriging predictions are available (Stein, 1988, 1990a,c). These results are based on the fixed-domain asymptotics (Stein, 1999, p.62), that is to say the case when the infinite sequence of observation points is dense in a bounded domain. In this setting, the results in (Stein, 1988, 1990a,c) state that if the true and assumed covariance functions yield equivalent Gaussian measures (see e.g Stein (1999, p.110) and Ibragimov and Rozanov (1978, p.63)), then there is asymptotically no loss using the incorrect covariance function. The asymptotic optimality holds for both the predictive means and variances.

Hence, in the fixed-domain asymptotics framework, it is sufficient to estimate a covariance function equivalent to the true one. Usually, it is assumed that the covariance function belongs to a given parametric family. In this case, the estimation boils down to estimating the corresponding parameters, that are called "hyper-parameters". Because of the above theoretical results, it is useful to distinguish between microergodic and non-microergodic hyper-parameters. Following the definition in Stein (1999), an hyper-parameter is microergodic if two covariance functions are orthogonal whenever they differ for it (as in Stein (1999), we say that two covariance functions are orthogonal if the two underlying Gaussian measures are orthogonal). Non-microergodic hyper-parameters cannot be consistently estimated but have no asymptotic influence on Kriging predictions, as shown in Zhang (2004) in the Matérn case. There is a fair amount of literature on the consistent estimation of microergodic hyper-parameters, particularly using the Maximum Likelihood (ML) method. Concerning the isotropic Matérn model, with fixed regularity parameter $\nu$ and free variance $\sigma^2$ and correlation length $\ell$, the microergodic ratio $\frac{\sigma^2}{\ell^{2\nu}}$ can be consistently estimated by ML for dimension $d \leq 3$ (Zhang, 2004), with asymptotic normality for $d = 1$ (Du et al., 2009). Both $\sigma^2$ and $\ell$ can be consistently estimated for $d > 4$ (Anderes, 2010). For the multiplicative Matérn model with $\nu = \frac{3}{2}$, both $\sigma^2$ and the $d$ correlation length hyper-parameters $\ell_1, ..., \ell_d$ are consistently estimated by ML for $d \geq 3$ (Loh, 2005). For the Gaussian model, the $d$ correlation lengths are consistently estimated by ML (Loh and Lam, 2000). Finally for the multiplicative exponential model, the microergodic ratio $\frac{\sigma^2}{\ell}$ is consistently estimated by ML with asymptotic normality for $d = 1$ (Ying, 1991). All hyper-parameters $\sigma^2, \ell_1, ..., \ell_d$ are consistently estimated by ML with asymptotic normality for

$d > 1$ (Ying, 1993).

We believe that the fixed-domain asymptotics does not solve completely the issue of the estimation of the covariance function. The first point is that the above theoretical results are asymptotic, while one could be interested in finite-sample studies, such as the numerical experiments performed in Stein (1999, ch.6.9), and the detailed study of the exponential covariance in Zhang and Zimmerman (2005). The second point is that one may not be able to asymptotically estimate a covariance function equivalent to the true one. Indeed, for instance, for two covariance functions of the isotropic Matérn class to be equivalent, it is necessary that their regularity parameters are equal (see the one-dimensional example in Stein (1999, p.136)). Yet, it is common practice, especially for the analysis of computer experiment data, to enforce the regularity parameter to an arbitrary value (see e.g Martin and Simpson (2004)). The interplay between the misspecification of the regularity parameter and the prediction mean square error is not trivial. The numerical experiments in Vazquez (2005, ch.5.3.3) show that a misspecified regularity parameter can have dramatic consequences.

The elements pointed out above justify addressing 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. In this context, we study the Cross Validation (CV) estimation method (Sundararajan and Keerthi, 2001; Zhang and Wang, 2010), and compare it with ML. This comparison has been an area of active research. Concerning theoretical results, Stein (1990b) showed that for the estimation of a signal-to-noise ratio parameter of a Brownian motion, CV has twice the asymptotic variance of ML in a well-specified case. Several numerical results are also available, coming either from Monte Carlo studies as in Santner et al. (2003, ch.3) or deterministic studies as in Martin and Simpson (2004). In both the above studies, the interpolated functions are smooth, and the covariance structures are adapted, being Gaussian in Martin and Simpson (2004) and having a free smoothness parameter in Santner et al. (2003).

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.

The paper is organized as follows. In Section 2, we detail the statistical framework for the estimation of a single variance hyper-parameter, 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 Section 3 we introduce the ML and Leave-One-Out (LOO) estimators of the variance hyper-parameter. In Section 4 we numerically apply the closed-form formulas of

Section 2 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, the CV does better compared to ML. Finally in Section 5 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.

## 2. Framework for the variance hyper-parameter estimation

We consider a Gaussian process $Y$, indexed by a set $\mathcal{X}$. $Y$ is 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

$$\mathcal{C} = \left\{ \sigma^2 R_2, \sigma^2 \in \mathbb{R}^+ \right\}, \tag{1}$$

with $R_2(x)$ a given stationary correlation function. Throughout this paper, $\mathbb{E}_i$, $var_i$, $cov_i$ and $\sim_i$, $i \in \{1,2\}$, denote means, variances, covariances and probability laws 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 $x_1, ..., x_n \in \mathcal{X}$. In this framework, the hyper-parameter $\sigma^2$ is estimated from the data $y = (y_1, ..., y_n)^t = (Y(x_1), ..., Y(x_n))^t$ using an estimator $\hat{\sigma}^2$. This estimation does not affect the Kriging prediction of $y_0 = Y(x_0)$, for a new point $x_0$, which we denote by $\hat{y}_0$:

$$\hat{y}_0 := \mathbb{E}_2(y_0|y) = \gamma_2^t \mathbf{\Gamma}_2^{-1} y, \tag{2}$$

where $(\gamma_i)_j = R_i(x_j - x_0)$ and $(\mathbf{\Gamma}_i)_{j,k} = R_i(x_j - x_k)$, $i \in \{1,2\}$, $1 \le j, k \le n$. The conditional mean square error of this non-optimal prediction is, after a simple calculation,

$$\mathbb{E}_1 \left[ (\hat{y}_0 - y_0)^2 | y \right] = (\gamma_1^t \mathbf{\Gamma}_1^{-1} y - \gamma_2^t \mathbf{\Gamma}_2^{-1} y)^2 + 1 - \gamma_1^t \mathbf{\Gamma}_1^{-1} \gamma_1. \tag{3}$$

However, using the covariance family $\mathcal{C}$, we use the classical Kriging predictive variance expression $\hat{\sigma}^2 c_{x_0}^2$, with

$$c_{x_0}^2 := var_2(y_0|y) = 1 - \gamma_2^t \mathbf{\Gamma}_2^{-1} \gamma_2. \tag{4}$$

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 2.1.** *For an estimator $\hat{\sigma}^2$ of $\sigma^2$, we call Risk at $x_0$ and denote by $\mathcal{R}_{\hat{\sigma}^2, x_0}$ the quantity*

$$\mathcal{R}_{\hat{\sigma}^2, x_0} = \mathbb{E}_1 \left[ \left( \mathbb{E}_1 \left[ (\hat{y}_0 - y_0)^2 | y \right] - \hat{\sigma}^2 c_{x_0}^2 \right)^2 \right].$$

If $\mathcal{R}_{\hat{\sigma}^2,x_0}$ is small, then this means that the predictive variance $\hat{\sigma}^2 c_{x_0}^2$ is a correct prediction of the conditional mean square error (3) of the prediction $\hat{y}_0$. Note that when $R_1 = R_2$ the minimizer of the Risk at every $x_0$ is $\hat{\sigma}^2 = 1$. When $R_1 \neq R_2$, an estimate of $\sigma^2$ 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 Section 3.

**Proposition 2.2.** *Let $\hat{\sigma}^2$ be an estimator of $\sigma^2$ of the form $y^t \mathbf{M} y$ 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, $\mathbf{M}_0 = (\mathbf{\Gamma}_2^{-1}\gamma_2 - \mathbf{\Gamma}_1^{-1}\gamma_1)(\gamma_2^t \mathbf{\Gamma}_2^{-1} - \gamma_1^t \mathbf{\Gamma}_1^{-1})\mathbf{\Gamma}_1$, $\mathbf{M}_1 = \mathbf{M}\mathbf{\Gamma}_1$, $c_1 = 1 - \gamma_1^t \mathbf{\Gamma}_1^{-1}\gamma_1$ and $c_2 = 1 - \gamma_2^t \mathbf{\Gamma}_2^{-1}\gamma_2$, we have:*

$$
\begin{aligned}
\mathcal{R}_{\hat{\sigma}^2,x_0} &= f(\mathbf{M}_0, \mathbf{M}_0) + 2c_1 tr(\mathbf{M}_0) - 2c_2 f(\mathbf{M}_0, \mathbf{M}_1) \\
&\quad + c_1^2 - 2c_1 c_2 tr(\mathbf{M}_1) + c_2^2 f(\mathbf{M}_1, \mathbf{M}_1).
\end{aligned}
$$

It seems difficult at first sight to conclude from Proposition 2.2 whether one estimator is better than another given a correlation function error and a set of observation points. Therefore, in Section 4, we numerically analyze the Risk for the ML and CV estimators of the variance for several designs of experiments. Before that, we introduce the ML and CV estimators of $\sigma^2$.

### 3. ML and CV estimation of the variance hyper-parameter

In the framework of Section 2, the ML estimator $\hat{\sigma}^2_{ML}$ of $\sigma^2$ (see e.g Santner et al. (2003, p.66)) is

$$
\hat{\sigma}^2_{ML} = \frac{1}{n} y^t \mathbf{\Gamma}_2^{-1} y. \tag{5}
$$

Let us now consider the CV estimator of $\sigma^2$. The principle is that, given a value $\sigma^2$ specifying the covariance function used among the set $\mathcal{C}$, we can, for $1 \leq i \leq n$, compute $\hat{y}_{i,-i} := \mathbb{E}_2(y_i|y_1,...,y_{i-1},y_{i+1},...,y_n)$ and $\sigma^2 c_{i,-i}^2 := \sigma^2 var_2(y_i|y_1,...,y_{i-1},y_{i+1},...,y_n)$. The Cross Validation estimate of $\sigma^2$ is based on the criterion

$$
C_{LOO} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,-i})^2}{\sigma^2 c_{i,-i}^2}. \tag{6}
$$

It is noted in Cressie (1993, p.102) that if $\sigma^2 R_2$ is a correct estimate of the true covariance function, then we should expect this criterion to be close to 1. In Appendix C, we show that if the observation points expand in a regular way, then for most classical correlation functions (the Matérn family for example), if $\sigma^2 R_2$ is the true covariance function, then (6) converges toward 1 in the mean square sense.

Based on this rigorous result about the case of an infinite regular grid, we can conjecture that (6) should be close to 1 in a general context when the value

of $\sigma^2$ is correct. Hence the idea of the CV estimation of $\sigma^2$ is to seek the value of $\sigma^2$ so that this criterion is equal to 1, which yields the estimator

$$\hat{\sigma}^2_{CV} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,-i})^2}{c^2_{i,-i}}. \tag{7}$$

By means of the well-known virtual LOO formulas of Proposition 3.1 (see e.g Ripley (1981, ch.5.2)), we obtain the following vector-matrix closed-form expression of (7),

$$\hat{\sigma}^2_{CV} = \frac{1}{n} y^t \mathbf{\Gamma}_2^{-1} \left[ diag(\mathbf{\Gamma}_2^{-1}) \right]^{-1} \mathbf{\Gamma}_2^{-1} y,$$

with $diag(\mathbf{\Gamma}_2^{-1})$ the matrix obtained by setting to 0 all non-diagonal terms of $\mathbf{\Gamma}_2^{-1}$. In the supplementary material, we give a short reminder, with proofs, of the virtual LOO formulas given in proposition 3.1.

**Proposition 3.1.** *For $1 \leq i \leq n$:*

$$c^2_{i,-i} = \frac{1}{(\mathbf{\Gamma}_2^{-1})_{i,i}}$$

*and*

$$y_i - \hat{y}_{i,-i} = \frac{1}{(\mathbf{\Gamma}_2^{-1})_{i,i}} (\mathbf{\Gamma}_2^{-1} y)_i.$$

When $R_1 = R_2$, we see that ML is more efficient than CV. Indeed

$$\mathcal{R}_{\hat{\sigma}^2, x_0} = \mathbb{E}_1 \left( (1 - \gamma_1^t \mathbf{\Gamma}_1^{-1} \gamma_1) - \hat{\sigma}^2 (1 - \gamma_1^t \mathbf{\Gamma}_1^{-1} \gamma_1) \right)^2 = (1 - \gamma_1^t \mathbf{\Gamma}_1^{-1} \gamma_1)^2 \mathbb{E}_1((\hat{\sigma}^2 - 1)^2), \tag{8}$$

so that the Risk of definition 2.1 is proportional to the quadratic error in estimating the true $\sigma^2 = 1$. We calculate $\mathbb{E}_1(\hat{\sigma}^2_{ML}) = \mathbb{E}_1(\hat{\sigma}^2_{CV}) = 1$, hence both estimators are unbiased. Concerning their variances, on the one hand we recall in Appendix B that $var_1(\hat{\sigma}^2_{ML}) = \frac{2}{n}$, the Cramér-Rao bound for the estimation of $\sigma^2$, in the case when the true $\sigma^2$ is 1 and the model for $Y$ is $\mathcal{C}$. On the other hand:

$$\begin{aligned} var_1(\hat{\sigma}^2_{CV}) &= \frac{2}{n^2} tr(\mathbf{\Gamma}_1^{-1} \left[ diag(\mathbf{\Gamma}_1^{-1}) \right]^{-1} \mathbf{\Gamma}_1^{-1} \left[ diag(\mathbf{\Gamma}_1^{-1}) \right]^{-1}) \\ &= \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(\mathbf{\Gamma}_1^{-1})^2_{i,j}}{(\mathbf{\Gamma}_1^{-1})_{i,i} (\mathbf{\Gamma}_1^{-1})_{j,j}}. \end{aligned}$$

Hence $var_1(\hat{\sigma}^2_{CV}) \geq \frac{2}{n^2} \sum_{i=1}^{n} \frac{(\mathbf{\Gamma}_1^{-1})^2_{i,i}}{(\mathbf{\Gamma}_1^{-1})_{i,i} (\mathbf{\Gamma}_1^{-1})_{i,i}} = \frac{2}{n}$, the Cramér-Rao bound.

However $var_1(\hat{\sigma}^2_{CV})$ is only upper-bounded by 2 (because $\mathbf{\Gamma}_1^{-1}$ is a covariance matrix). Furthermore $var_1(\hat{\sigma}^2_{CV})$ can be arbitrarily close to 2. Indeed, with

$$\mathbf{\Gamma}_1 = \mathbf{\Gamma}_2 = \frac{n - 1 + \epsilon}{n - 1} \mathbf{I} - \frac{\epsilon}{n - 1} \mathbf{J},$$

where $\mathbf{J}$ is the $n \times n$ matrix with all coefficients being 1, we obtain

$$var_1(\hat{\sigma}^2_{CV}) = \frac{2}{n} + \frac{2(n-1)}{n} \frac{\epsilon^2}{(\epsilon + (n-1)(1-\epsilon))^2}.$$

Hence, when $R_1 = R_2$, ML is more efficient to estimate the variance parameter. The object of the next section is to study the case $R_1 \neq R_2$ numerically.

## 4. Numerical results for the variance hyper-parameter estimation

All the numerical experiments are carried out with the numerical software Scilab (Gomez et al., 1999). 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.

### 4.1. 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 $x_0$.

The first criterion is the Risk on Target Ratio (RTR),

$$RTR(x_0) = \frac{\sqrt{\mathcal{R}_{\hat{\sigma}^2, x_0}}}{\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right]}, \tag{9}$$

with $\hat{\sigma}^2$ being either $\hat{\sigma}^2_{ML}$ or $\hat{\sigma}^2_{CV}$.

From Definition 2.1 we obtain

$$RTR(x_0) = \frac{\sqrt{\mathbb{E}_1\left[\left(\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right] - \hat{\sigma}^2 c^2_{x_0}\right)^2\right]}}{\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right]}. \tag{10}$$

The numerator of (10) is the mean square error in predicting the random quantity $\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right]$ (the target in the RTR acronym) with the predictor $\hat{\sigma}^2 c^2_{x_0}$. The denominator of (10) is, by the law of total expectation, the mean of the predictand $\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right]$. Hence, the RTR in (9) is a relative prediction error, which is easily interpreted.

We have the following bias-variance decomposition of the Risk,

$$\mathcal{R}_{\hat{\sigma}^2, x_0} = \left(\underbrace{\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right] - \mathbb{E}_1\left[\hat{\sigma}^2 c^2_{x_0}\right]}_{\text{bias}}\right)^2 + \underbrace{var_1\left(\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right] - \hat{\sigma}^2 c^2_{x_0}\right)}_{\text{variance}}. \tag{11}$$

Hence the second criterion is the Bias on Target Ratio (BTR) and is the relative bias

$$BTR(x_0) = \frac{|\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right] - \mathbb{E}_1\left(\hat{\sigma}^2 c^2_{x_0}\right)|}{\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right]}. \tag{12}$$

7

The following equation summarizes the link between RTR and BTR:

$$\left(\underbrace{RTR}_{\text{relative error}}\right)^2 = \left(\underbrace{BTR}_{\text{relative bias}}\right)^2 + \underbrace{\frac{var_1\left(\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right] - \hat{\sigma}^2 c_{x_0}^2\right)}{\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right]^2}}_{\text{relative variance}}.$$

(13)

*Pointwise criteria when $R_1 = R_2$.* When $R_1 = R_2$, $\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right]$ does not depend on $y$. Therefore, the RTR and BTR simplify into $RTR(x_0) = \sqrt{\mathbb{E}_1\left[(\hat{\sigma}^2 - 1)^2\right]}$ and $BTR(x_0) = |1 - \mathbb{E}_1(\hat{\sigma}^2)|$. 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.* In this paragraph, we define the two integrated versions of RTR and BTR over the prediction space $\mathcal{X}$. Assume $\mathcal{X}$ is equipped with a probability measure $\mu$. Then we define

$$IRTR = \sqrt{\int_{\mathcal{X}} RTR^2(x_0)d\mu(x_0)}$$

(14)

and

$$IBTR = \sqrt{\int_{\mathcal{X}} BTR^2(x_0)d\mu(x_0)}.$$

(15)

Hence we have the equivalent of (13) for IRTR and IBTR:

$$IRTR^2 = IBTR^2 + \int_{\mathcal{X}} \frac{var_1\left(\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2|y\right] - \hat{\sigma}^2 c_{x_0}^2\right)}{\mathbb{E}_1\left[(\hat{y}_0 - y_0)^2\right]^2}d\mu(x_0).$$

(16)

*4.2. Designs of experiments studied*

We consider three different kinds of Designs Of Experiments (DOEs) of $n$ observation points on the prediction space $\mathcal{X} = [0, 1]^d$.

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 for 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 Stein (1999, ch.6.9) that such points can dramatically improve the estimation of the covariance hyper-parameters.

The second DOE is the Latin Hypercube Sampling Maximin (LHS-Maximin) design (see e.g Santner et al. (2003)). This design is one of the most widespread non-iterative designs in Kriging. To generate a LHS-Maximin design, we generate 1000 LHS designs, and keep the one that maximizes the min criterion (i.e. the minimum distance between two different observation points). Let us notice

that this is the method used by the Matlab function *lhsdesign(...,'maximin',k)* which generates $k$ LHS designs with default $k = 5$.

The third DOE is a deterministic sparse regular grid. It is built according to the Smolyak (1963) sparse tensorization construction of the one-dimensional regular grid $G = \{\frac{1}{2^l}, ..., \frac{2^l-1}{2^l}\}$ of level $l$.

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).

*4.3. Families of correlation functions studied*

We consider two classical families of stationary correlation functions.

- The power-exponential correlation function family, parameterized by the vector of correlation lengths $\ell = (\ell_1, ..., \ell_d)$ and the power $p$. $R$ is power-exponential $(\ell, p)$ when

$$R(h) = \exp\left(-\sum_{i=1}^{d} \left(\frac{|h_i|}{\ell_i}\right)^p\right). \tag{17}$$

- The Matérn correlation function family, parameterized by the vector of correlation lengths $\ell = (\ell_1, ..., \ell_d)$ and the regularity parameter $\nu$. $R$ is Matérn $(\ell, \nu)$ when

$$R(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu}|h|_\ell\right)^\nu K_\nu \left(2\sqrt{\nu}|h|_\ell\right), \tag{18}$$

with $|h|_\ell = \sqrt{\sum_{i=1}^{d} \frac{h_i^2}{\ell_i^2}}$, $\Gamma$ the Gamma function and $K_\nu$ the modified Bessel function of second order. See e.g Stein (1999, p.31) for a presentation of the Matérn correlation function.

*4.4. 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 (see Section 4.2). We compare the empirical means of the two integrated criteria IRTR and IBTR (see Section 4.1) 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 $\nu_2$. Finally, $R_1$ is Matérn $((1.2, ..., 1.2), 1.5)$ and $R_2$ is Matérn $((\ell_2, ..., \ell_2), 1.5)$ with varying $\ell_2$.

On Figure 1, 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
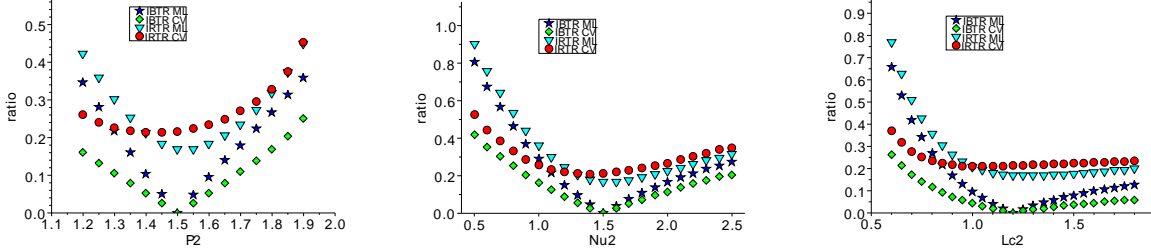
Figure 1: Influence of the model error for the SRS DOE (see Section 4.2). Plot of the IRTR and IBTR integrated criteria (see Section 4.1) for ML and CV. Left: power-exponential correlation function with error on the exponent, the true exponent is $p_1 = 1.5$ and the model exponent $p_2$ varies in $[1.2, 1.9]$. Middle: Matérn correlation function with error on the regularity parameter, the true regularity parameter is $\nu_1 = 1.5$ and the model regularity parameter $\nu_2$ varies in $[0.5, 2.5]$. Right: Matérn correlation function with error on the correlation length, the true correlation length is $\ell_1 = 1.2$ and the model correlation length $\ell_2$ varies in $[0.6, 1.8]$. ML is optimal when there is no model error while CV is more robust to model misspecifications.
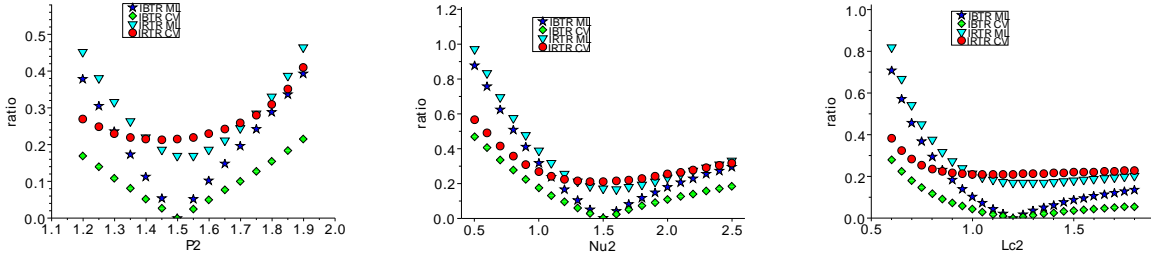


Figure 2: Same setting as in Figure 1, but with the LHS-Maximin DOE (see Section 4.2). ML is optimal when there is no model error while CV is more robust to model misspecifications.

of IRTR. Looking at (16), 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.

On Figure 2 we plot the results for the LHS-Maximin DOE. The results are similar to these 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.

On Figure 3, 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 (compare the min, between ML and CV, of IRTR for the SRS and LHS-Maximin designs against the regular
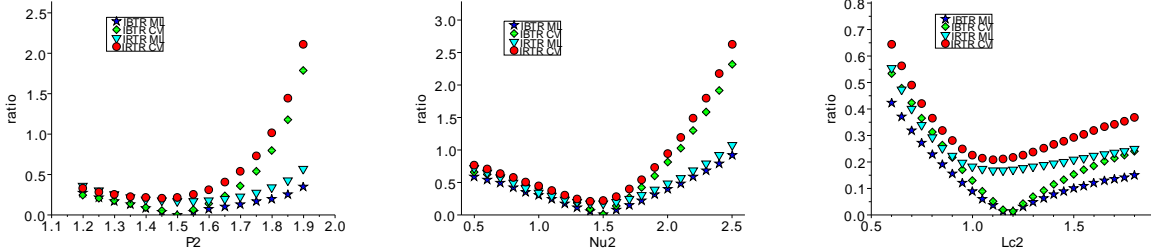
10

Figure 3: Same setting as in Figure 1 but with the regular sparse grid DOE (see Section 4.2). 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.

grid). This is especially true for misspecifications on the exponent for the power-exponential correlation function and on the regularity 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 $\sigma^2$ 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 $\sigma^2$ and not a $\sigma^2$ for the purpose of assessing predictive variances at particular points. Furthermore, it is noted in Iooss et al. (2010) 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.

### 4.5. Influence of the number of points

Using the same procedure as in Section 4.4, 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
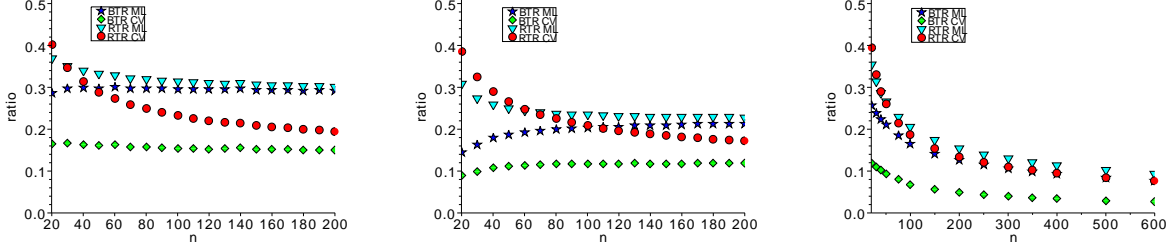
Figure 4: Influence of the number $n$ of observation points for the SRS DOE (see Section 4.2). Plot of the RTR and BTR criteria (see Section 4.1) for prediction at the center of the domain and for ML and CV. Left: power-exponential correlation function with error on the exponent, the true exponent is $p_1 = 1.5$ and the model exponent is $p_2 = 1.7$. Middle: Matérn correlation function with error on the regularity parameter, the true regularity parameter is $\nu_1 = 1.5$ and the model regularity parameter is $\nu_2 = 1.8$. Right: 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 $\ell_2 = 1.8$.

of interest and focus on the prediction on the point $x_0$ having all its components set to $\frac{1}{2}$ (center of domain).

On Figure 4 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 (13) that RTR is the sum of BTR and of a relative 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).

On Figure 5 we plot the results for the LHS-Maximin DOE. The results are similar to these of the SRS DOE. The RTR of CV is smaller than the RTR of ML for a slightly smaller $n$. This confirms the results of Section 4.4 where the model error for which the IRTR of ML reaches the IRTR of CV is smaller for LHS-Maximin than for SRS.
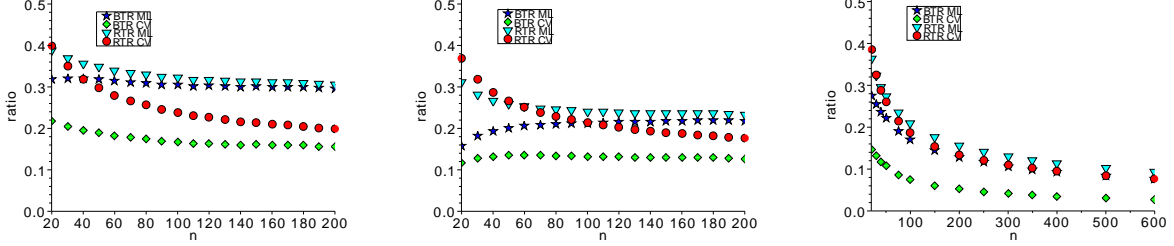
12

Figure 5: Same setting as in Figure 4, but with the LHS-Maximin DOE (see Section 4.2).

## 5. Study on analytical functions

In this section, we compare ML and CV on analytical functions, instead of realizations of Gaussian processes, as was the case in Section 4. The first goal is to illustrate the results of Section 4 on the estimation of the variance hyper-parameter. Indeed, the study of Section 4 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 is to generalize Section 4 to the case where correlation hyper-parameters are estimated from data.

### 5.1. ML and CV estimations of covariance hyper-parameters

We consider a set of observations $(x_1, y_1), ..., (x_n, y_n)$ as in Section 2, and the family $\left\{ \sigma^2 R_\theta, \sigma^2 > 0, \theta \in \Theta \right\}$ of stationary covariance functions, with $R_\theta$ a stationary correlation function, and $\Theta$ a finite-dimensional set. We denote by $\mathbb{E}_\theta$ and $var_\theta$ the means and variances with respect to the distribution of a stationary Gaussian process with mean zero, variance one and correlation function $R_\theta$.

The ML estimate of $(\sigma^2, \theta)$ is $\hat{\theta}_{ML} \in \text{argmin}_{\theta \in \Theta} |\mathbf{\Gamma}_\theta|^{1/n} \hat{\sigma}^2_{ML}(\mathbf{\Gamma}_\theta)$ (see e.g Marrel et al. (2008)), with $\mathbf{\Gamma}_\theta$ the correlation matrix of the training sample with correlation function $R_\theta$, and $\hat{\sigma}^2_{ML}(\mathbf{\Gamma}_\theta)$ as in (5). $\hat{\sigma}^2_{ML}(\mathbf{\Gamma}_{\hat{\theta}_{ML}})$ is the estimate of $\sigma^2$.

For CV we choose to use the following estimator for the hyper-parameter $\theta$:

$$\hat{\theta}_{CV} \in \underset{\theta \in \Theta}{\text{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_{i,\theta}(y_{-i}))^2, \tag{19}$$

with $\hat{y}_{i,\theta}(y_{-i}) = \mathbb{E}_\theta(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$. This is the Leave-One-Out mean square error criterion that is used, for instance, in Santner et al. (2003) when the ML and CV estimations of $\theta$ are compared. Given $\hat{\theta}_{CV}$ and hence $\mathbf{\Gamma}_{\hat{\theta}_{CV}}$, $\sigma^2$ is estimated by (7).

13

After elementary transformations for the estimator $\hat{\theta}_{ML}$ and the use of Proposition 3.1 for the estimator $\hat{\theta}_{CV}$, the functions to minimize are

$$f_{ML}(\theta) = \frac{1}{n} \log det(\mathbf{\Gamma}_\theta) + \log \left( y^t \mathbf{\Gamma}_\theta^{-1} y \right) \qquad (20)$$

and

$$f_{CV}(\theta) = y^t \mathbf{\Gamma}_\theta^{-1} diag(\mathbf{\Gamma}_\theta^{-1})^{-2} \mathbf{\Gamma}_\theta^{-1} y. \qquad (21)$$

In the supplementary material, we recall the closed-form expressions of the gradients of $f_{ML}$ and $f_{CV}$, as functions of the first-order derivatives of the correlation function. These expressions are available in the literature, see e.g Mardia and Marshall (1984) for ML and Sundararajan and Keerthi (2001) for CV. 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, implemented in the Scilab function *optim*. Since the functions $f_{ML}$ and $f_{CV}$ 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 Martin and Simpson (2004). An important point is that, when $\theta$ 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 uniformly the space of correlation lengths, as is the case with a LHS design, using directly 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 regularity parameter). To tackle this issue we recommend to use the Nugget effect. More specifically, for a given correlation matrix $\mathbf{\Gamma}_\theta$, we actually compute $\mathbf{\Gamma}_\theta + \tau^2 I_n$, 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 Andrianakis and Challenor (2012). However, for the CV estimation of $\theta$, when the correlation function belongs to the Gaussian family, or the Matérn family with large regularity parameter, another structural problem appears. For $\hat{\sigma}_{CV}^2$ very large, as the overall predictive variance term $\hat{\sigma}_{CV}^2 (1 - \gamma_\theta \mathbf{\Gamma}_\theta^{-1} \gamma_\theta)$ has the same order of magnitude as the observations, the term $1 - \gamma_\theta \mathbf{\Gamma}_\theta^{-1} \gamma_\theta$ is very small. Hence, a fixed numerical error on the inversion of $\mathbf{\Gamma}_\theta$, however small it is, may cause the term $1 - \gamma_\theta \mathbf{\Gamma}_\theta^{-1} \gamma_\theta$ 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 $\hat{\sigma}_{CV}^2$, which yields small $(1 - \gamma_\theta \mathbf{\Gamma}_\theta^{-1} \gamma_\theta)$, so possibly negative ones. Notice however that the

relative errors of the Kriging prediction terms $\gamma_\theta^t \mathbf{\Gamma}_\theta^{-1} y$ are correct. It is noted in Martin and Simpson (2004, 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 $\hat{\sigma}_{CV}^2$ in the minimization of $f_{CV}$. 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 $\theta$ estimation, depends on the DOE substantially. In practice, we use a penalty for $\hat{\sigma}_{CV}^2$ starting at 1000 times the empirical variance $\frac{1}{n} y^t y$. This penalty is needed only for CV when the correlation function is Gaussian or Matérn with free regularity parameter.

*5.2. Procedure*

We consider a deterministic function $f$ on $[0,1]^d$. We generate $n_p = 100$ LHS-Maximin training samples of the form $x_{a,1}, f(x_{a,1}), ..., x_{a,n}, f(x_{a,n})$. We denote $y_{a,i} = f(x_{a,i})$. From each training sample, we estimate $\sigma^2$ and $\theta$ with the ML and CV methods presented in Section 5.1.

We are interested in two criteria on a large Monte Carlo test sample $x_{t,1}, f(x_{t,1}), ..., x_{t,n_t}, f(x_{t,n_t})$ on $[0,1]^d$ ($n_t = 10000$). We denote $y_{t,i} = f(x_{t,i})$, $\hat{y}_{t,i}(y_a) = \mathbb{E}_{\hat{\theta}}(y_{t,i}|y_a)$ and $\hat{\sigma}^2 c_{t,i}^2(y_a) = \sigma^2 var_{\hat{\theta}}(y_{t,i}|y_a)$, where $\hat{\theta}$ comes from either the ML or CV method.

The first criterion is the Mean Square Error (MSE) and evaluates the prediction capability of the estimated correlation function $R_{\hat{\theta}}$:

$$\frac{1}{n_t} \sum_{i=1}^{n_t} (y_{t,i} - \hat{y}_{t,i}(y_a))^2. \tag{22}$$

The second criterion is the Predictive Variance Adequation (PVA):

$$\left| \log \left( \frac{1}{n_t} \sum_{i=1}^{n_t} \frac{(y_{t,i} - \hat{y}_{t,i}(y_a))^2}{\hat{\sigma}^2 c_{t,i}^2(y_a)} \right) \right|. \tag{23}$$

This criterion evaluates the quality of the predictive variances given by the estimated covariance hyper-parameters $\hat{\sigma}^2, \hat{\theta}$. 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.

*5.3. Analytical functions studied*

We study the two following analytical functions. The first one, for $d = 3$, is the Ishigami function:

$$f(x_1, x_2, x_3) = \sin(-\pi + 2\pi x_1) + 7\sin((-\pi + 2\pi x_2))^2 + 0.1(-\pi + 2\pi x_3)^4 \sin(-\pi + 2\pi x_1). \tag{24}$$

15

The second one, for $d = 10$, is a simplified version of the Morris function (Morris, 1991),

$$f(x) = \sum_{i=1}^{10} w_i(x) + \sum_{1 \leq i < j \leq 6} w_i(x)w_j(x) + \sum_{1 \leq i < j < k \leq 5} w_i(x)w_j(x)w_k(x)$$
$$+ w_1(x)w_2(x)w_3(x)w_4(x),$$

with $\quad w_i(x) = \begin{cases} 2\left(\frac{1.1x_i}{x_i+0.1} - 0.5\right), & \text{if } i = 3, 5, 7 \\ 2(x_i - 0.5) & \text{otherwise.} \end{cases}$

### 5.4. Results with enforced correlation lengths

We work with the Ishigami function, with $n = 100$ observation points. For the correlation function family, we study the exponential and Gaussian families (power-exponential family of (17) with enforced $p = 1$ for exponential and $p = 2$ for Gaussian).

For each of these two correlation models, we enforce three vectors $\ell$ 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 in Section 5.1. 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 1. The first comment is that, comparing line 1, 2 and 3 against line 4, 5 and 6, 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). The second comment is that the benefit obtained by replacing an isotropic correlation length by different correlation lengths is smaller for the exponential class than for the Gaussian one. 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 Section 4: The ML estimation of $\sigma^2$ is more appropriate when the correlation function is well-specified while the CV estimation is more appropriate when the correlation function is misspecified.

| 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 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.

*5.5. Results with estimated correlation lengths*

We work with the Ishigami and Morris functions, with $n = 100$ observation points. We use the exponential and Gaussian models as in Section 5.4, as well as the Matérn model of (18). We distinguish two subcases for the vector $\ell$ 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.

In Table 2, we present the results. 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 regularity hyper-parameters are large, so that the MSEs and the PVAs for the Matérn model are similar to the ones 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 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 that the exponential correlation function yields 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 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 | 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 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 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 2, the Gaussian correlation model does not yield smaller MSEs than the exponential one.

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 4: ML is optimal for the best adapted correlation models, while CV is more robust to cases of model misspecification.

### 5.6. 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 can equally be studied, for which this process is considered to have a mean at location $x$ of the form $\sum_{i=1}^{p} \beta_i g_i(x)$, with known functions $g_i$ and unknown coefficients $\beta_i$. For instance a closed-form formula similar to that of Proposition 2.2 can be obtained in the same fashion, and virtual LOO formulas are also available (Dubrule, 1983). 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

| 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 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 2.

covariance function choice. Furthermore it is shown in Stein (1999, 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, in Table 4 we study, for the Ishigami function, the influence of using a universal Kriging model with either a constant or affine mean function. The process is the same as for Table 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 Martin and Simpson (2004). The second overall conclusion is that the main finding of Section 4 and of Table 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 most adapted and yields the smallest MSE.

## 6. Conclusion

In this paper, 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 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 when the model is misspecified, CV performs better than ML. Second, we have studied the general case when 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.

19

Because CV is more robust to model misspecification, it is less likely than ML to yield substantially incorrect predictions. However, ML is more likely to yield the best predictions, under the condition that the correlation function family is ideally specified. Hence, CV is a suitable method for problems when one gives priority to robustness over best possible performance. Studying data-driven methods, to choose between ML and CV according to a specific objective, may motivate further research.

We also pointed out in Section 4.4 that some DOEs that are too regular are not suitable for CV. Investigating this further and studying numerical criteria for quantifying the suitability of a DOE for CV could motivate further research. One could be also interested in CV-oriented adaptive improvement of a DOE. Finally, we emphasized the need, in some cases, for a penalization of large correlation lengths for CV. We enforced this penalization via a penalization of too large estimated global variance hyper-parameters. The motivation here was purely numerical, but there could be a statistical motivation for doing so. The statistical analysis of penalization methods, similar to the one proposed here, for CV may motivate further research. In the ML case, there exist penalizations of correlation hyper-parameters that both reduce the small sample variance, and recover the asymptotic distribution of the non-penalized case, as shown in Li and Sudjianto (2005).

**Acknowledgments**

## Appendix A. Proof of Proposition 2.2

Using the definition of the Risk, the expression of $\hat{\sigma}^2$, (3) and (4), we get:

$$
\begin{aligned}
\mathcal{R}_{\hat{\sigma}^2,x_0} &= \mathbb{E}_1\big[(\gamma_1^t\boldsymbol{\Gamma}_1^{-1}y - \gamma_2^t\boldsymbol{\Gamma}_2^{-1}y)^2 + 1 - \gamma_1^t\boldsymbol{\Gamma}_1^{-1}\gamma_1 \\
&\quad - y^t\mathbf{M}y(1 - \gamma_2^t\boldsymbol{\Gamma}_2^{-1}\gamma_2)\big]^2 \\
&= \mathbb{E}_1\big[y^t(\boldsymbol{\Gamma}_2^{-1}\gamma_2 - \boldsymbol{\Gamma}_1^{-1}\gamma_1)(\gamma_2^t\boldsymbol{\Gamma}_2^{-1} - \gamma_1^t\boldsymbol{\Gamma}_1^{-1})y \\
&\quad + 1 - \gamma_1^t\boldsymbol{\Gamma}_1^{-1}\gamma_1 - y^t\mathbf{M}y(1 - \gamma_2^t\boldsymbol{\Gamma}_2^{-1}\gamma_2)\big]^2.
\end{aligned}
$$

Then, writing $y = \boldsymbol{\Gamma}_1^{\frac{1}{2}}z$ with $z \sim_1 \mathcal{N}(0, \mathbf{I}_n)$, we get:

$$
\mathcal{R}_{\hat{\sigma}^2,x_0} = \mathbb{E}_1\left(z^t\tilde{\mathbf{M}}_0 z + c_1 - c_2 z^t\tilde{\mathbf{M}}_1 z\right)^2, \tag{A.1}
$$

with

$$
\tilde{\mathbf{M}}_0 = \boldsymbol{\Gamma}_1^{\frac{1}{2}}(\boldsymbol{\Gamma}_2^{-1}\gamma_2 - \boldsymbol{\Gamma}_1^{-1}\gamma_1)(\gamma_2^t\boldsymbol{\Gamma}_2^{-1} - \gamma_1^t\boldsymbol{\Gamma}_1^{-1})\boldsymbol{\Gamma}_1^{\frac{1}{2}}
$$

and

$$
\tilde{\mathbf{M}}_1 = \boldsymbol{\Gamma}_1^{\frac{1}{2}}\mathbf{M}\boldsymbol{\Gamma}_1^{\frac{1}{2}}.
$$

To compute this expression, we use the following lemma. The proof relies only on $4^{\text{th}}$ moment calculation for Gaussian variables and is omitted.

**Lemma Appendix A.1.** *Let $z \sim_1 \mathcal{N}(0, \mathbf{I}_n)$, and $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ real symmetric matrices. Then:*

$$
\mathbb{E}_1(z^t\mathbf{A}z z^t\mathbf{B}z) = f(\mathbf{A}, \mathbf{B}).
$$

Using the lemma and expanding (A.1) yields

$$
\begin{aligned}
\mathcal{R}_{\hat{\sigma}^2,x_0} &= f(\tilde{\mathbf{M}}_0, \tilde{\mathbf{M}}_0) + 2c_1 tr(\tilde{\mathbf{M}}_0) - 2c_2 f(\tilde{\mathbf{M}}_0, \tilde{\mathbf{M}}_1) \\
&\quad + c_1^2 - 2c_1 c_2 tr(\tilde{\mathbf{M}}_1) + c_2^2 f(\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_1).
\end{aligned} \tag{A.2}
$$

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 (A.2), which completes the proof.

## Appendix B. On the optimality of $\hat{\sigma}^2_{ML}$ when $R_1 = R_2$

Here, we consider the case $R_2 = R_1$. We first recall the Cramér-Rao inequality, which states that when $\sigma^2 = 1$, for an unbiased estimator $\hat{\sigma}^2$ of $\sigma^2$:

$$
var_1(\hat{\sigma}^2) \geq \mathbb{E}_1^{-1}\left[\left(\frac{\partial}{\partial\sigma^2}\left(\ln L(y, \sigma^2)\right)_{\sigma^2=1}\right)^2\right],
$$

with, $L(y, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{y^t \mathbf{\Gamma}_1^{-1} y}{2\sigma^2}\right)$, the likelihood of the observations. We then calculate the Cramér-Rao bound:

$$
\begin{aligned}
\mathbb{E}_1^{-1}\left[\left(\frac{\partial}{\partial \sigma^2}\left(\ln L(y, \sigma^2)\right)_{\sigma^2=1}\right)^2\right] &= \mathbb{E}_1^{-1}\left[\left(\frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\ln\sigma^2 - \frac{y^t \mathbf{\Gamma}_1^{-1} y}{2\sigma^2}\right)_{\sigma^2=1}\right)^2\right] \\
&= \mathbb{E}_1^{-1}\left[\frac{n^2}{4} + \frac{1}{4}(y^t\mathbf{\Gamma}_1^{-1}y)^2 - \frac{n}{2}y^t\mathbf{\Gamma}_1^{-1}y\right] \\
&= \left(\frac{n^2}{4} + \frac{n^2+2n}{4} - \frac{n^2}{2}\right)^{-1} \\
&= \frac{2}{n},
\end{aligned}
$$

where we used Lemma Appendix A.1 with $\mathbf{A} = \mathbf{B} = \mathbf{I}_n$ to show $\mathbb{E}_1\left((y^t\mathbf{\Gamma}_1^{-1}y)^2\right) = n^2 + 2n$. Hence, the Cramér-Rao bound of the statistical model $\mathcal{C}$ is $\frac{2}{n}$ when $\sigma^2 = 1$.

Finally, $var_1(\hat{\sigma}_{ML}^2) = var_1(\frac{1}{n}y^t\mathbf{\Gamma}_1^{-1}y) = \frac{1}{n^2}var_1(\sum_{i=1}^n z_i^2) = \frac{2}{n}$, with $z = \mathbf{\Gamma}_1^{-\frac{1}{2}}y \sim_1 \mathcal{N}(0, \mathbf{I}_n)$.

## Appendix C. Convergence of (6)

**Proposition Appendix C.1.** *Assume $\mathcal{X} = \mathbb{R}^d$ and that the observation points are distinct and located on the infinite regular grid $\{(i_1\delta, ..., i_d\delta), (i_1, ..., i_d) \in \mathbb{Z}^d\}$ for fixed $\delta > 0$. When $\sigma^2 R_2$ is the covariance function of the Gaussian process $Y$, is summable on the regular grid, and has a strictly positive continuous Fourier transform, then the criterion $C_{LOO}$ has mean one and converges in mean square to one as $n \to +\infty$.*

*Proof.* Consider, for simplicity, $\sigma^2 = 1$. Using the formulas of Proposition 3.1 and introducing $z = \mathbf{\Gamma}_2^{-\frac{1}{2}}y \sim_2 \mathcal{N}(0, \mathbf{I}_n)$ yields:

$$
C_{LOO} = \frac{1}{n}\sum_{i=1}^n \frac{(y_i - \hat{y}_{i,-i})^2}{c_{i,-i}^2} = \frac{1}{n}z^t\mathbf{\Gamma}_2^{-\frac{1}{2}}\left[diag(\mathbf{\Gamma}_2^{-1})\right]^{-1}\mathbf{\Gamma}_2^{-\frac{1}{2}}z,
$$

with $diag(\mathbf{\Gamma}_2^{-1})$ the matrix obtained by setting to 0 all non-diagonal terms of $\mathbf{\Gamma}_2^{-1}$. Then,

$$
\mathbb{E}_2(C_{LOO}) = \frac{1}{n}tr\left(\mathbf{\Gamma}_2^{-\frac{1}{2}}\left[diag(\mathbf{\Gamma}_2^{-1})\right]^{-1}\mathbf{\Gamma}_2^{-\frac{1}{2}}\right) = \frac{1}{n}\sum_{i=1}^n \frac{(\mathbf{\Gamma}_2^{-1})_{i,i}}{(\mathbf{\Gamma}_2^{-1})_{i,i}} = 1.
$$

Furthermore,

$$
var_2(C_{LOO}) = \frac{2}{n^2}tr\left(\mathbf{\Gamma}_2^{-1}\left[diag(\mathbf{\Gamma}_2^{-1})\right]^{-1}\mathbf{\Gamma}_2^{-1}\left[diag(\mathbf{\Gamma}_2^{-1})\right]^{-1}\right). \tag{C.1}
$$

22

Then, with $\lambda_{min}(\mathbf{\Gamma})$ and $\lambda_{max}(\mathbf{\Gamma})$ the smallest and largest eigenvalues of a symmetric, strictly positive, matrix $\mathbf{\Gamma}$,

$$\lambda_{max}\big(\big[diag(\mathbf{\Gamma}_2^{-1})\big]^{-1}\big) = \max_{1 \le i \le n} \frac{1}{(\mathbf{\Gamma}_2^{-1})_{i,i}} \le \frac{1}{\lambda_{min}(\mathbf{\Gamma}_2^{-1})} = \lambda_{max}(\mathbf{\Gamma}_2). \qquad (C.2)$$

Hence, from (C.1) and (C.2),

$$\frac{2}{n}\left(\frac{\lambda_{min}(\mathbf{\Gamma}_2)}{\lambda_{max}(\mathbf{\Gamma}_2)}\right)^2 \le var_2(C_{LOO}) \le \frac{2}{n}\left(\frac{\lambda_{max}(\mathbf{\Gamma}_2)}{\lambda_{min}(\mathbf{\Gamma}_2)}\right)^2.$$

Hence $0 < \inf_n \lambda_{min}(\mathbf{\Gamma}_2) \le \sup_n \lambda_{max}(\mathbf{\Gamma}_2) < +\infty$ implies the proposition.

First, for all $n$, $\lambda_{max}(\mathbf{\Gamma}_2)$ is smaller than the largest $l^1$ norm of the rows of $\mathbf{\Gamma}_2$ and hence is smaller than the absolute sum of $R_2$ over the infinite regular grid. Hence we have $\sup_n \lambda_{max}(\mathbf{\Gamma}_2) < +\infty$.

Then, let $\hat{R}_2$ be the Fourier transform of $R_2$, and let $R_{min} > 0$ be its infimum over $[-\frac{\pi}{\delta}, \frac{\pi}{\delta}]^d$. For all $n$, $x_1, ..., x_n$ in the infinite regular grid, and for all $\epsilon \in \mathbb{R}^n$:

$$\epsilon^t \mathbf{\Gamma}_2 \epsilon = \sum_{k,l=1}^n R_2(x_k - x_l)\epsilon_k \epsilon_l = \sum_{k,l=1}^n \epsilon_k \epsilon_l \int_{\mathbb{R}^d} e^{it(x_k - x_l)} \hat{R}_2(t)dt = \int_{\mathbb{R}^d} \hat{R}_2(t) \left|\sum_{k=1}^n \epsilon_k e^{itx_k}\right|^2 dt$$

$$\ge R_{min} \int_{[-\frac{\pi}{\delta}, \frac{\pi}{\delta}]^d} \left|\sum_{k=1}^n \epsilon_k e^{itx_k}\right|^2 dt = R_{min} \int_{[-\frac{\pi}{\delta}, \frac{\pi}{\delta}]^d} \sum_{k,l=1}^n \epsilon_k \epsilon_l e^{it(x_k - x_l)} dt.$$

Then, because $\int_{[-\frac{\pi}{\delta}, \frac{\pi}{\delta}]} e^{itp\delta}dt = 0$ for a non-zero integer $p$, and because the $x_k$, $1 \le k \le n$, are distinct and located on the regular grid, we have $\int_{[-\frac{\pi}{\delta}, \frac{\pi}{\delta}]^d} e^{it(x_k - x_l)}dt = 0$ for $k \ne l$. Hence:

$$\epsilon^t \mathbf{\Gamma}_2 \epsilon \ge R_{min} \int_{[-\frac{\pi}{\delta}, \frac{\pi}{\delta}]^d} \sum_{k=1}^n \epsilon_k^2 dt = R_{min} \left(\frac{2\pi}{\delta}\right)^d ||\epsilon||^2,$$

which in turn implies $0 < \inf_n \lambda_{min}(\mathbf{\Gamma}_2)$ and thus completes the proof.

# References

Anderes, E., 2010. On the consistent separation of scale and variance for Gaussian random fields. The Annals of Statistics 38, 870–893.

Andrianakis, I., Challenor, P.G., 2012. The effect of the nugget on Gaussian process emulators of computer models. Computational Statistics and Data Analysis 56, 4215–4228.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.

Du, J., Zhang, H., Mandrekar, V., 2009. Fixed domain asymptotics properties of tapered maximum likelihood estimators. The Annals of Statistics 37, 3330–3361.

Dubrule, O., 1983. Cross validation of Kriging in a unique neighborhood. Mathematical Geology 15, 687–699.

Gomez, C., Bunks, C., Chancelier, J., Delebecque, F., Goursat, M., Nikoukhah, R., Steer, S., 1999. Engineering and Scientific Computing with Scilab. Birkhaüser, Boston.

Ibragimov, I., Rozanov, Y., 1978. Gaussian Random Processes. Springer-Verlag, New York.

Iooss, B., Boussouf, L., Feuillard, V., Marrel, A., 2010. Numerical studies of the metamodel fitting and validation processes. International Journal of Advances in Systems and Measurements 3, 11–21.

Jones, D., Schonlau, M., Welch, W., 1998. Efficient global optimization of expensive black box functions. Journal of Global Optimization 13, 455–492.

Li, R., Sudjianto, A., 2005. Analysis of computer experiments using penalized likelihood in Gaussian Kriging model. Technometrics 47, 111–120.

Loh, W., 2005. Fixed domain asymptotics for a subclass of Matérn type Gaussian random fields. The Annals of Statistics 33, 2344–2394.

Loh, W., Lam, T., 2000. Estimating structured correlation matrices in smooth Gaussian random field models. The Annals of Statistics 28, 880–904.

Mardia, K., Marshall, R., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71, 135–146.

Marrel, A., Iooss, B., Dorpe, F.V., Volkova, E., 2008. An efficient methodology for modeling complex computer codes with Gaussian processes. Computational Statistics and Data Analysis 52, 4731–4744.

Martin, J., Simpson, T., 2004. 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.

Matheron, G., 1970. 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.

Morris, M.D., 1991. Factorial sampling plans for preliminary compututational experiments. Technometrics 33, 161–174.

Paulo, R., Garcia-Donato, G., Palomo, J., 2012. Calibration of computer models with multivariate output. Computational Statistics and Data Analysis 56, 3959–3974.

Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. The MIT Press, Cambridge.

Ripley, B., 1981. Spatial Statistics. Wiley, New York.

Sacks, J., Welch, W., Mitchell, T., Wynn, H., 1989. Design and analysis of computer experiments. Statistical Science 4, 409–423.

Santner, T., Williams, B., Notz, W., 2003. The Design and Analysis of Computer Experiments. Springer, New York.

Smolyak, S., 1963. Quadrature and interpolation formulas for tensor products of certain classes of functions. Soviet Math. Dokl 4, 240–243.

Stein, M., 1988. Asymptotically efficient prediction of a random field with a misspecified covariance function. The Annals of Statistics 16, 55–63.

Stein, M., 1990a. Bounds on the efficiency of linear predictions using an incorrect covariance function. The Annals of Statistics 18, 1116–1138.

Stein, M., 1990b. 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.

Stein, M., 1990c. Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. The Annals of Statistics 18, 850–872.

Stein, M., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Sundararajan, S., Keerthi, S., 2001. Predictive approaches for choosing hyper-parameters in Gaussian processes. Neural Computation 13, 1103–1118.

Vazquez, E., 2005. Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. Ph.D. thesis. Université Paris XI Orsay. Available at http://tel.archives-ouvertes.fr/tel-00010199/en.

Ying, Z., 1991. Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. Journal of Multivariate Analysis 36, 280–296.

Ying, Z., 1993. Maximum likelihood estimation of parameters under a spatial sampling scheme. The Annals of Statistics 21, 1567–1590.

Zhang, H., 2004. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. Journal of the American Statistical Association 99, 250–261.

Zhang, H., Wang, Y., 2010. Kriging and cross validation for massive spatial data. Environmetrics 21, 290–304.

Zhang, H., Zimmerman, D., 2005. Toward reconciling two asymptotic frameworks in spatial statistics. Biometrika 92, 921–936.

# Supplementary material for
# Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification

François Bachoc[a,b,*]

[a] CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
[b] Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII
Site Chevaleret, case 7012, 75205 Paris cedex 13

---

**Abstract**

In this supplementary material, a proof is given for the virtual Leave-One-Out formulas. The expression of the gradient of the Cross Validation criterion is also given, together with a proof.

---

In all the supplementary material, the notations are the same as in the paper.

**Proposition 0.1.** *For* $1 \leq i \leq n$:

$$c_{i,-i}^2 = \frac{1}{(\boldsymbol{\Gamma}_2^{-1})_{i,i}}$$

*and*

$$y_i - \hat{y}_{i,-i} = \frac{1}{(\boldsymbol{\Gamma}_2^{-1})_{i,i}}(\boldsymbol{\Gamma}_2^{-1}y)_i.$$

*Proof.* We first show that if the formulas of the proposition are true for $i = 1$ then they are true for all $i$. Let $1 \leq i \leq n$. Let $\tilde{y}$ be the vector taken from $y$ by interchanging the $i$th and the 1st components. Then $cov_2(\tilde{y}) = \mathbf{M}^{1i}\boldsymbol{\Gamma}_2\mathbf{M}^{1i}$ with $(\mathbf{M}^{1i})_{k,l}$ being 1 if $k = l$ and $\{k\} \notin \{1, i\}$, 1 if $\{k, l\} = \{1, i\}$ and 0 elsewhere. $\mathbf{M}^{1i}$ is the matrix interchanging components 1 and $i$ of a vector. Then, using the formulas of the proposition for $\tilde{y}$ with $i = 1$:

$$var_2(y_i|y_{-i}) = var_2(\tilde{y}_1|\tilde{y}_{-1}) = \frac{1}{((\mathbf{M}^{1i}\boldsymbol{\Gamma}_2\mathbf{M}^{1i})^{-1})_{1,1}} = \frac{1}{(\mathbf{M}^{1i}\boldsymbol{\Gamma}_2^{-1}\mathbf{M}^{1i})_{1,1}} = \frac{1}{((\boldsymbol{\Gamma}_2)^{-1})_{i,i}}.$$

---

*Corresponding author: François Bachoc
CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
Phone: +33 (0) 1 69 08 97 91
Email: francois.bachoc@cea.fr

Furthermore:

$$y_i - \mathbb{E}_2(y_i|y_{-i}) = \tilde{y}_1 - \mathbb{E}_2(\tilde{y}_1|\tilde{y}_{-1}) = \frac{1}{((\mathbf{M}^{1i}\mathbf{\Gamma}_2\mathbf{M}^{1i})^{-1})_{1,1}}((\mathbf{M}^{1i}\mathbf{\Gamma}_2\mathbf{M}^{1i})^{-1}\tilde{y})_1$$

$$= \frac{1}{((\mathbf{\Gamma}_2)^{-1})_{i,i}}(\mathbf{M}^{1i}\mathbf{\Gamma}_2^{-1}\mathbf{M}^{1i}\tilde{y})_1 = \frac{1}{((\mathbf{\Gamma}_2)^{-1})_{i,i}}(\mathbf{M}^{1i}(\mathbf{\Gamma}_2^{-1}y))_1 = \frac{1}{((\mathbf{\Gamma}_2)^{-1})_{i,i}}(\mathbf{\Gamma}_2^{-1}y)_i.$$

Hence it is sufficient to show the formulas for the case $i = 1$, which yields the simplest expressions. Using a formula for the inverse of a partitioned matrix (Rao, 1991, p.33), we write:

$$\mathbf{\Gamma}_2^{-1} = \begin{pmatrix} 1 & \gamma_{2,1}^t \\ \gamma_{2,1} & \mathbf{\Gamma}_{2,-1} \end{pmatrix}^{-1} \tag{1}$$

$$= \frac{1}{1 - \gamma_{2,1}^t(\mathbf{\Gamma}_{2,-1})^{-1}\gamma_{2,1}}\begin{pmatrix} 1 & -\gamma_{2,1}^t(\mathbf{\Gamma}_{2,-1})^{-1} \\ -(\mathbf{\Gamma}_{2,-1})^{-1}\gamma_{2,1} & (\mathbf{\Gamma}_{2,-1})^{-1}\gamma_{2,1}\gamma_{2,1}^t(\mathbf{\Gamma}_{2,-1})^{-1} + (\mathbf{\Gamma}_{2,-1})^{-1} \end{pmatrix}.$$

From (1), we first get

$$(\mathbf{\Gamma}_2^{-1})_{1,1} = \frac{1}{1 - \gamma_{2,1}^t(\mathbf{\Gamma}_{2,-1})^{-1}\gamma_{2,1}},$$

which, together with

$$c_{1,-1}^2 = 1 - \gamma_{2,1}^t(\mathbf{\Gamma}_{2,-1})^{-1}\gamma_{2,1},$$

proves the first one of the formulas.

Right multiplying (1) with vector $y$ and taking the first component yields

$$(\mathbf{\Gamma}_2^{-1})_{1,1}(y_1 - \gamma_{2,1}^t(\mathbf{\Gamma}_{2,-1})^{-1}y_{-1}) = (\mathbf{\Gamma}_2^{-1}y)_1,$$

which, together with

$$\hat{y}_{1,-1} = \gamma_{2,1}^t\mathbf{\Gamma}_{2,-1}y_{-1},$$

proves the second one of the formulas.

Let us notice that we have presented here the virtual LOO formulas in the case of simple Kriging, that is to say the case when the mean of the stationary process is zero. Indeed, it is the case of interest in the present paper. Similar virtual LOO formulas have been proved in Dubrule (1983) in the more general case of universal Kriging, for which the mean at $x$ is of the form $\sum_{i=1}^p \beta_i g_i(x)$ with known functions $g_i$ and unknown coefficients $\beta_i$. It is also shown in Dubrule (1983) how to generalize these formulas to the case of $K$-fold cross validation. $\square$

**Proposition 0.2.** *Let $\theta_i$ be a component of $\theta$ and $\left(\frac{\partial}{\partial\theta_i}\mathbf{\Gamma}_\theta\right)$ be the derivative of $\mathbf{\Gamma}_\theta$ with respect to $\theta_i$. Then:*

$$\frac{\partial}{\partial\theta_i}f_{ML}(\theta) = \frac{1}{n}tr\left(\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial\theta_i}\mathbf{\Gamma}_\theta\right)\right) - \frac{1}{y^t\mathbf{\Gamma}_\theta^{-1}y}y^t\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial\theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}y$$

*and*

$$\frac{\partial}{\partial \theta_i} f_{CV}(\theta) = 2y^t \mathbf{\Gamma}_\theta^{-1} diag(\mathbf{\Gamma}_\theta^{-1})^{-2} diag\left(\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}\right) diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\mathbf{\Gamma}_\theta^{-1}y$$

$$- 2y^t \mathbf{\Gamma}_\theta^{-1} diag(\mathbf{\Gamma}_\theta^{-1})^{-2}\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}y.$$

*Proof.* The expression of $\frac{\partial}{\partial \theta_i} f_{ML}(\theta)$ is directly obtained from the matrix-derivative formulas

$$\frac{\partial}{\partial \theta_i} \log(det(\mathbf{\Gamma}_\theta)) = tr\left(\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\right) \tag{2}$$

and

$$\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta^{-1} = -\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}. \tag{3}$$

For $f_{CV}$, we have, with $\epsilon = diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\mathbf{\Gamma}_\theta^{-1}y$, and making an extensive use of (3),

$$\frac{\partial}{\partial \theta_i}\epsilon = \left(\frac{\partial}{\partial \theta_i}diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\right)\mathbf{\Gamma}_\theta^{-1}y + diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta^{-1}\right)y$$

$$= -diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\left(\frac{\partial}{\partial \theta_i}diag(\mathbf{\Gamma}_\theta^{-1})\right)diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\mathbf{\Gamma}_\theta^{-1}y - diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}y$$

$$= diag(\mathbf{\Gamma}_\theta^{-1})^{-1}diag\left(\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}\right)diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\mathbf{\Gamma}_\theta^{-1}y$$

$$- diag(\mathbf{\Gamma}_\theta^{-1})^{-1}\mathbf{\Gamma}_\theta^{-1}\left(\frac{\partial}{\partial \theta_i}\mathbf{\Gamma}_\theta\right)\mathbf{\Gamma}_\theta^{-1}y,$$

which completes the proof with $\frac{\partial}{\partial \theta_i}(\epsilon^t\epsilon) = 2\epsilon^t\left(\frac{\partial}{\partial \theta_i}\epsilon\right).$

$\square$

## References

Dubrule, O., 1983. Cross validation of Kriging in a unique neighborhood. Mathematical Geology 15, 687–699.

Rao, C., 1991. Linear Statistical Inference and its Applications. Wiley, New York.