Maximum de vraisemblance et Validation Croisée pour l'estimation des hyper-paramètres de covariance pour le Krigeage

François Bachoc [1], Josselin Garnier [2] & Jean-Marc Martinez [3]

[1] *CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France. Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VII.*
francois.bachoc@cea.fr
[2] *UFR Mathematiques, site Chevaleret, case 7012 Universite Paris VII 75205 Paris Cedex 13, France.*
garnier@math.univ-paris-diderot.fr
[3] *CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France.*
jean-marc.martinez@cea.fr

Résumé. Dans le cadre de l'estimation de la fonction de covariance pour un modèle de Krigeage, nous étudions les estimateurs du Maximum de Vraisemblance (MV) et de la Validation Croisée (VC). Dans un premier temps, nous montrons que, lorsque la famille paramétrique de fonctions de covariance est mal spécifiée, la VC est plus robuste que le MV. Nous étudions d'abord analytiquement le cas de l'estimation d'un unique paramètre de variance, puis nous étudions numériquement le cas général. Dans un second temps, nous considérons, dans le cas où la famille paramétrique est bien spécifiée, l'impact asymptotique de la régularité du plan d'expériences sur les estimateurs par MV et VC. Le plan d'expériences est une grille régulière parfaite aléatoirement perturbée, le degré de perturbation étant fonction d'un unique paramètre scalaire de régularité. Nous prouvons alors la consistance et la normalité asymptotique, pour les estimateurs du MV et de la VC. Les matrices de covariance asymptotique sont des fonctions déterministes du paramètre de régularité. Par une étude exhaustive de ces matrices, nous montrons que l'irrégularité du plan d'expériences est souvent un avantage pour l'estimation, mais nous identifions des cas pour lesquels cela est faux. Ainsi, nous répondons par la négative à l'assertion selon laquelle un plan d'expériences irrégulier est toujours meilleur qu'un plan d'expériences régulier pour l'estimation des hyper-paramètres de covariance.

Mots-clés. Quantification des incertitudes, Krigeage, processus Gaussiens, estimation de la fonction de covariance, Validation Croisée, Maximum de vraisemblance,

Abstract. The choice of the covariance function is an important issue in Kriging. For this issue, we study the Maximum Likelihood (ML) and Cross Validation (CV) estimators. In a first step, we show that the CV estimator is relevant, by showing that it is more robust than ML, when the parametric set of covariance functions for the estimation is misspecified. First, we study analytically the case of the estimation of a single variance

hyper-parameter, when the correlation function is fixed and misspecified. Second, we study the general case numerically. In a second step, we study, in the case where the set of covariance functions is well-specified, the impact of the regularity of the spatial sampling, in an assymptotic context. The spatial sampling is a randomly perturbed regular grid and its deviation from the perfect regular grid is controlled by a single scalar regularity parameter. Consistency and asymptotic normality are proved for the Maximum Likelihood and Cross Validation estimators of the hyper-parameters. The asymptotic covariance matrices of the hyper-parameter estimators are deterministic functions of the regularity parameter. By means of an exhaustive study of the asymptotic covariance matrices, it is shown that irregular sampling is generally an advantage to estimation, but we identify cases where it is not the case. Therefore, a negative answer is given to the claim that irregular sampling is always better for hyper-parameter estimation than regular sampling.

Keywords. Uncertainty quantification, Kriging, Gaussian process, Covariance function estimation, Cross Validation, Maximum Likelihood

1 Le problème de l'estimation paramétrique de la fonction de covariance pour le Krigeage

La méthode de Krigeage consiste à modéliser une fonction déterministe par la réalisation d'un processus Gaussien Y, indexé par un domaine $\mathcal{X}$ de $\mathbb{R}^d$. Nous considérons ce processus centré et stationnaire. La fonction de covariance de Y est notée R_1. Lorsque cette fonction de covariance est connue, la résolution complète du modèle de Krigeage (calcul de la loi conditionnelle du processus Gaussien, sachant un ensemble d'observations de celui ci) est effectuée par des équations d'algébre linéaire (voir par exemple Santner et al. (2003)).

Néanmoins, la fonction de covariance R_1 est inconnue dans la grande majorité des cas d'application. La pratique la plus répandue consiste à estimer cette fonction, à partir d'un ensemble d'observations du processus Gaussien, et de la supposer ensuite connue et égale à son estimation (méthode *plug in*, Stein (1999), ch.6.8).

Nous traitons ici le cas le plus classique dans lequel la fonction de covariance est estimée parmi les fonctions de l'ensemble paramétrique $\mathcal{R} = \{\sigma^2 R_\theta, \sigma^2 > 0, \theta \in \Theta\}$, où Θ est un domaine compact de $\mathbb{R}^p$, et R_θ est une fonction de corrélation stationnaire.

Lorsque le processus Gaussien est observé sur les points d'observation $x_1, ..., x_n \in \mathcal{X}$, nous notons $y_i = Y(x_i)$. Nous étudions deux estimateurs pour les hyper-paramètres σ^2, θ.

Le premier est l'estimateur par Maximum de Vraisemblance (MV) (Mardia et Marshall, 1984), que nous notons $\hat{\sigma}^2_{ML}$ et $\hat{\theta}_{ML}$.

Le second est l'estimateur par Validation Croisée (VC) (Rasmussen et Williams (2003),

ch.5). Cet estimateur est défini par

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

et

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

avec $\hat{y}_{i,\theta}(y_{-i}) := \mathbb{E}_\theta(y_i|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$ et $c^2_{i,-i,\theta} := var_\theta(y_i|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$. Nous notons $\mathbb{E}_\theta$ et var_θ les moyennes et variances lorsque Y suit la distribution donnée par la fonction de covariance R_θ avec les hyper-paramètres de covariance $\sigma^2 = 1$ et θ. En résumé, l'estimateur de la VC estime d'abord θ en minimisant l'Erreur Quadratique Moyenne (EQM) empirique par *Leave-One-Out*, puis ajuste le paramètre σ^2 de sorte que les variances prédictives soient adaptées à ces erreurs.

Une remarque importante d'un point de vue pratique est que les évaluations des critères du MV et de la VC ont en fait le même coût calculatoire, grâce aux formules de *Leave-One-Out* virtuel de Dubrule (2003).

2 Une comparaison du MV et de la VC en cas de mauvaise spécification de l'ensemble de fonctions de covariance

Nous montrons ici que l'estimateur de la VC est pertinent en pratique, car il est plus robuste que le MV aux cas de mauvaises spécifications de l'ensemble de fonctions de covariance $\mathcal{R}$. Nous appelons mauvaise spécification les cas pour lesquels cet ensemble $\mathcal{R}$ est éloigné de la vraie fonction de covariance R_1. Nous procédons en deux étapes.

Dans une première étape, nous étudions le cas pour lequel $\mathcal{R} = \{\sigma^2 R_2, \sigma^2 > 0\}$, avec R_2 une fonction de corrélation fixée et différente de la vraie fonction de corrélation R_1. Ce cadre de travail nous permet d'étudier le critère d'erreur suivant pour un estimateur $\hat{\sigma}^2$ de σ^2.

$$R_{\hat{\sigma}^2,x_0} := \mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - y_0)^2|y\right] - \hat{\sigma}^2 c^2_{x_0}\right)^2\right], \tag{3}$$

avec x_0 un point de prédiction dans le domaine d'intérêt $\mathcal{X}$, y le vecteur d'observations, $\mathbb{E}\left[(\hat{y}_0 - y_0)^2|y\right]$ l'erreur quadratique moyenne de prédiction conditionnellement aux observations et $\hat{\sigma}^2 c^2_{x_0}$ l'estimation de cette erreur de prédiction. Notons que la prédiction (sous-optimale) $\hat{y}_0$, ainsi que l'estimation incorrecte de l'erreur de prédiction associée $c^2_{x_0}$ sont effectuées avec la fonction de corrélation incorrecte R_2. En résumé, le Risque dans (3) est l'erreur quadratique moyenne entre la vraie erreur quadratique moyenne $\mathbb{E}\left[(\hat{y}_0 - y_0)^2|y\right]$, de la prédiction par Krigeage $\hat{y}_0$, et la prédiction $\hat{\sigma}^2 c^2_{x_0}$ de cette erreur, donnée par l'estimateur $\hat{\sigma}^2$.

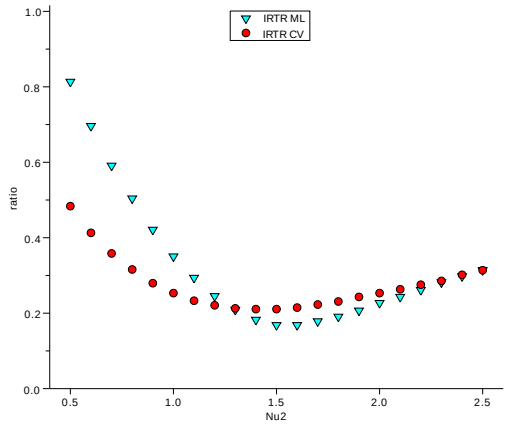

Figure 1: Plan d'expérience de 70 points de type *Latin Hypercube Sampling Maximin*, sur $[0,1]^5$. Tracé de l'erreur relative intégrée $IRTR := \int_{\mathcal{X}} \frac{\sqrt{R_{\hat{\sigma}^2,x_0}}}{\mathbb{E}[(\hat{y}_0-y_0)^2]} d_{x_0}$ (voir (3)) pour les estimateurs de MV et de la VC. R_1 est une fonction de covariance de Matérn isotropique avec la longueur de corrélation $\ell = 1.2$ et le paramètre de régularité $\nu_1 = 1.5$. R_2 est une fonction de covariance de Matérn isotropique avec la longueur de corrélation $\ell = 1.2$ et le paramètre de régularité ν_2 variant sur le tracé. Lorsque l'erreur sur la fonction de corrélation augmente, la VC devient plus efficace que le MV pour estimer le paramètre de variance.

Dans Bachoc (+2013a), une expression analytique de (3) est obtenue, ce qui permet de montrer que le MV est plus efficace lorsque R_2 est proche de R_1, mais que la VC est plus robuste lorsque R_2 devient sensiblement différent de R_1. Nous illustrons cela dans la figure 1, pour laquelle nous traçons la quantité *Integrated Risk on Target Ratio*, $IRTR := \int_{\mathcal{X}} \frac{\sqrt{R_{\hat{\sigma}^2,x_0}}}{\mathbb{E}[(\hat{y}_0-y_0)^2]} d_{x_0}$, qui est une erreur relative intégrée dans le domaine $\mathcal{X}$.

Dans une seconde étape, nous étudions le cas général $\mathcal{R} = \{\sigma^2 R_\theta, \sigma^2 > 0, \theta \in \Theta\}$. L'étude est effectuée par des expériences numériques sur les fonctions analytiques d' Ishigami et de Morris. Les mauvaises spécifications de $\mathcal{R}$ que nous étudions sont une structure isotrope et l'utilisation de fonctions de covariance irrégulières (les fonctions analytiques sont régulières). Les résultats obtenus confirment ceux de la première étape.

La conclusion générale est que la VC est plus robuste que le MV aux cas de mauvaises spécifications de l'ensemble paramétrique de fonctions de covariance.

3 Influence de l'irrégularité du plan d'expérience dans le cas bien spécifié

Nous étudions ici le cas bien spécifié dans lequel la vraie fonction de covariance appartient à l'ensemble paramétrique de fonctions de covariance $\mathcal{R}$. De plus, nous n'utilisons plus spécifiquement la décomposition σ^2, θ. Nous avons donc $\mathcal{R} = \{R_\theta, \theta \in \Theta\}$, avec Θ un domaine compact de $\mathbb{R}^p$ et R_θ une fonction de covariance stationnaire. Nous avons également la vraie fonction de covariance $R_1 = R_{\theta_0}$, avec θ_0 dans l'intérieur de Θ.

Nous choisissons d'étudier l'influence de l'irrégularité du plan d'expérience sur l'estimation des hyper-paramètres de covariance dans un cadre asymptotique. Dans la littérature, deux cadres asymptotiques existent: asymptotique par remplissage et par expansion (Stein (1999) p.62). Pour l'asymptotique par remplissage, la suite de points d'observations est dense dans un domaine borné. Pour l'asymptotique par expansion, une distance minimale existe entre deux points d'observations différents, de sorte que le domaine d'observation est non borné. Le cadre asymptotique par expansion est meilleur pour étudier l'influence de l'irrégularité du plan d'expérience sur l'estimation des hyper-paramètres de covariance. En effet, l'éventuelle distribution asymptotique des estimateurs d'hyper-paramètres de covariance, dans le cadre par remplissage, est indépendante du plan d'expérience (voir par exemple Ying, 1991), ce qui n'est pas le cas dans le cadre par expansion.

Nous notons $(v_i)_{i\in\mathbb{N}^*}$ une suite de points d'observations sur $\mathbb{N}^d$ telle que pour tout $N \in \mathbb{N}$, $\left\{v_i, 1 \leq i \leq N^d\right\} = \{1, ..., N\}^d$. La suite de points d'observations du processus Gaussien et la suite des $v_i + \epsilon X_i$, $1 \leq i \leq n$, avec $-\frac{1}{2} < \epsilon < \frac{1}{2}$ et $X_i \sim_{iid} \mathcal{L}_X$. $\mathcal{L}_X$ est une loi de probabilité de support $S_X \subset [-1, 1]^d$, à densité de probabilité strictement positive sur S_X. Deux remarques peuvent être faites sur ce cadre asymptotique.

- La condition $-\frac{1}{2} < \epsilon < \frac{1}{2}$ assure une distance minimale entre deux points d'observation différents. Nous sommes donc bien dans un cadre asymptotique par expansion. La consistance et la normalité asymptotique de l'estimateur du MV a été prouvée dans Sweeting (1980) et dans Mardia et Marshall (1984). A notre connaissance, il n'existe pas de résultats asymptotiques pour l'estimateur de la VC dans la littérature.

- Les plans d'expériences sont aléatoires, et le paramètre ϵ est un paramètre de régularité. Le cas $\epsilon = 0$ correspond à une grille régulière de points d'observation, tandis que l'irrégularité augmente lorsque $|\epsilon|$ se rapproche de $\frac{1}{2}$.

Pour le MV, nous notons, pour un nombre d'observations n, $(\mathbf{I}_{ML})_n$ la matrice d'information de Fisher modifiée. C'est une matrice aléatoire $p \times p$ définie par

$$(\mathbf{I}_{ML})_{n,i,j} = \frac{1}{n} Tr(\Gamma_{\theta_0,n}^{-1} \frac{\partial \Gamma_{\theta_0,n}}{\partial \theta_i} \Gamma_{\theta_0,n}^{-1} \frac{\partial \Gamma_{\theta_0,n}}{\partial \theta_j}),$$

avec $\Gamma_{\theta_0,n}$ la matrice aléatoire $n \times n$ de covariance pour l'hyper-paramètre de covariance θ_0. Dans Bachoc (+2013b), il est montré que $(\mathbf{I}_{ML})_n$ converge presque sûrement vers

une matrice déterministe $\mathbf{I}_{ML}$, qui ne dépend que de ϵ. Sous des conditions générales d'identifiabilité (assurant en outre la positivité de $\mathbf{I}_{ML}$), il est aussi montré que

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \rightarrow_{\mathcal{L}} \mathcal{N}(0, 2\mathbf{I}_{ML}^{-1}). \tag{4}$$

Pour la VC, un résultat similaire est montré dans Bachoc (+2013b). Il est montré, sous les conditions d'identifiabilité et lorsque l'ensemble paramétrique de fonctions de covariance est constitué de fonctions de corrélation, que

$$\sqrt{n}(\hat{\theta}_{CV} - \theta_0) \rightarrow_{\mathcal{L}} \mathcal{N}(0, \mathbf{I}_{CV,2}^{-1}\mathbf{I}_{CV,1}\mathbf{I}_{CV,2}^{-1}). \tag{5}$$

Les matrices $\mathbf{I}_{CV,1}$ et $\mathbf{I}_{CV,2}$ sont des limites presque sûres de matrices aléatoires, dont les expressions sont similaires à celle de l'information de Fisher modifiée.

Nous étudions ensuite, dans le cas de l'estimation d'un seul paramètre de covariance, les variances asymptotiques données par (4) et (5), en fonction du paramètre de régularité ϵ. Nous effectuons une analyse exhaustive du modèle de Matérn. Nous mettons en évidence que l'irrégularité du plan d'expériences est généralement un avantage pour l'estimation, en particulier lorsque ϵ est proche de $\frac{1}{2}$. Néanmoins, il existe des cas pour lesquels une perturbation de la grille régulière (faible pour le MV, faible ou forte pour la VC) peut dégrader l'estimation de la fonction de covariance.

Bibliographie

[1] Bachoc, F. (+2013a), Cross Validation and Maximum Likelihood Estimations of Hyper-parameters of Gaussian Processes with Model Misspecification, En révision mineure pour Computational Statistics and Data Analysis.

[2] Bachoc, F. (+2013b), Asymptotic analysis of the role of the spatial sampling for hyper-parameter estimation of Gaussian processes, Soumis.

[3] Dubrule, 0. (1983), Cross Validation of Kriging in a Unique Neighborhood, Mathematical Geology, 15, 687-699.

[4] Mardia, K.V. et Marshall, R.J. (1984), Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression, 71, 135-146.

[5] Rasmussen, C.E. et Williams, C.K.I (2006), Gaussian Processes for Machine Learning, The MIT Press, Cambridge.

[6] Santner, T.J., Williams, B.J. et Notz, W.I, (2003), The Design and Analysis of Computer Experiments, Springer, New York.

[7] Stein, M.L. (1999), Interpolation of Spatial Data: Some Theory for Kriging, Springer, New York.

[8] Sweeting, T.J. (1980), Uniform Asymptotic Normality of the Maximum Likelihood Estimator, 8, 1375-1381.

[9] Ying, Z. (1991), Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process, 36, 280-296.