

Validation Croisée et Maximum de Vraisemblance en cas d'erreur de modèle sur un processus Gaussien

François Bachoc

CEA-Saclay, DEN, DM2S, SFME, LGLS, F-91191 Gif-sur-Yvette, France.

1 Cadre du doctorat

Etudiant : François Bachoc, ingénieur Supélec (2007/2010)

Début du doctorat : 04/10/10

Encadrement universitaire : Ecole doctorale de l'université Paris VII. Thèse dirigée par Joselin Garnier, professeur à Paris VII.

Encadrement CEA : Thèse encadrée par Jean-Marc Martinez, chercheur au LGLS.

2 Le problème de l'estimation des hyper-paramètres dans le Krigeage

Il est montré dans [3] que l'efficacité du modèle de substitution par Krigeage ne dépend pas asymptotiquement du choix de la fonction de covariance tant que celle-ci est équivalente à la vraie. Nous étudions le cas complémentaire où le choix de la fonction de covariance conditionne l'efficacité du Krigeage. Dans le plupart des cas, cette fonction est choisie dans un ensemble paramétrique prenant généralement la forme $\mathcal{C} = \{\sigma^2 C_\theta, \sigma^2 > 0, \theta \in \Theta\}$. Il s'agit donc d'estimer le vecteur d'hyper-paramètres (σ^2, θ) . Les méthodes que nous étudions à ce sujet sont celles du maximum de vraisemblance ([2, 4]) et de la validation croisée. La méthode de validation croisée que nous étudions consiste à minimiser en θ une erreur de prédiction par validation croisée, puis à ajuster σ^2 de manière à avoir des variances prédictives adaptées au vecteur d'erreur par validation croisée. L'implémentation est fondée sur les formules de validation croisée virtuelle de [1] et est donc de complexité comparable à celle du maximum de vraisemblance.

3 Contribution dans le cas d'une mauvaise spécification de l'ensemble de fonctions de covariance

Nous nous intéressons au cas dans lequel l'ensemble \mathcal{C} est relativement éloigné de la vraie fonction de covariance. Nous effectuons alors l'analyse en deux étapes.

Dans une première étape, nous étudions le cas dans lequel $\mathcal{C} = \{\sigma^2 C_{mod}, \sigma^2 > 0\}$ avec C_{mod} une fonction de corrélation fixée différente de la vraie fonction de corrélation C_0 . Cela nous permet d'obtenir une expression analytique du critère de qualité suivant pour un estimateur $\hat{\sigma}^2$ de σ^2 :

$$R_{\hat{\sigma}^2, x_0} = \mathbb{E} \left[\left(\mathbb{E} [(y_0 - \hat{y}_0)^2 | y] - \hat{\sigma}^2 c_{x_0}^2 \right)^2 \right]. \quad (1)$$

où x_0 est un point de prédiction, y est le vecteur d'observations, $\mathbb{E} [(y_0 - \hat{y}_0)^2 | y]$ est l'erreur de prédiction conditionnellement aux observations, et $\hat{\sigma}^2 c_{x_0}^2$ est l'estimation de cette erreur de prédiction. Nous obtenons alors une expression analytique de (1), ce qui nous permet de mettre en évidence que la maximum de vraisemblance est plus efficace lorsque C_{mod} est égal à C_0 mais

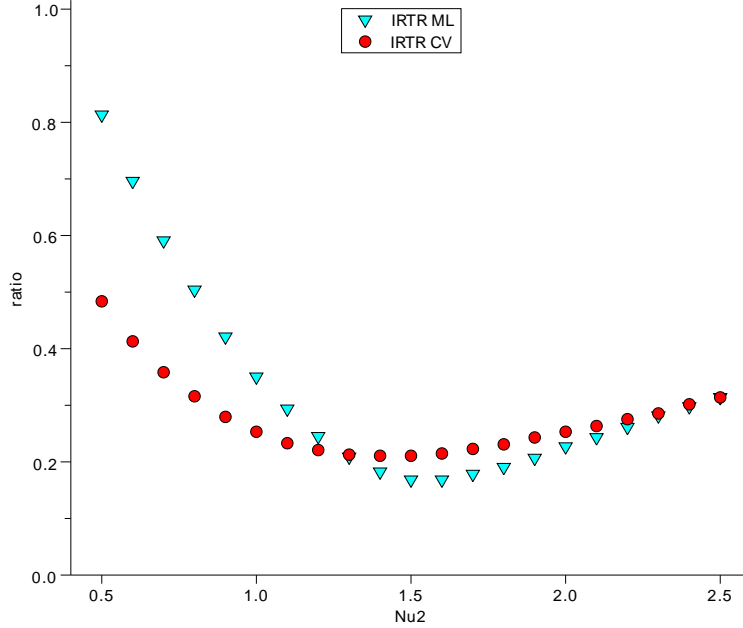


FIG. 1 – Krigeage en dimension 5 avec 70 points d’observations g n r s par LHS Maximin. La vraie fonction de cor relation est Matern avec longueur de cor relation commune $l_c = 1.2$ et param tre de r gularit  $\nu_1 = 1.5$. La fonction de cor relation utilis e pour le Krigeage est la fonction de Matern avec longueur de cor relation commune $l_c = 1.2$ et param tre de r gularit  ν_2 variant. On trace le IRTR moyen sur 50 tirages de plans d’exp riences en fonction de ν_2 . ML est plus efficace lorsque $\nu_1 = \nu_2$ et CV est plus robuste aux mauvaises sp cifications de ν_2 .

que la validation crois e devient plus efficace lorsque C_{mod} s’ loigne de C_0 . Ce comportement est illustr  par la figure 1 sur laquelle nous tracons la quantit  d’int r t

$$IRTR = \int_{x_0 \in [0,1]^d} \frac{\sqrt{R_{\hat{\sigma}^2, x_0}}}{\mathbb{E}[(\hat{y}_0 - y_0)^2]},$$

qui est une erreur relative int gr e.

La second  tape est d’ tudier le cas g n ral $\mathcal{C} = \{\sigma^2 C_\theta, \sigma^2 > 0, \theta \in \Theta\}$. Nous effectuons cela par le biais d’exp riences num riques sur des fonctions analytiques. Les r sultats obtenus confirment ceux de la premi re  tape.

R f rences

- [1] O. Dubrule. Cross validation of kriging in a unique neighborhood. *Mathematical Geology*, 15 :687–699, 1983.
- [2] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71 :135–146, 1984.
- [3] M.L Stein. *Interpolation of Spatial Data Some Theory for Kriging*. Springer, 1999.
- [4] T.J. Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *Annals of statistics*, 8 :1375–1381, 1980.