# Validation croisée virtuelle pour la calibration de modèles numériques

François Bachoc

## 1 Cadre du doctorat

**Etudiant :** François Bachoc, ingénieur Supélec (2007/2010), Option Mathématiques Appliquées de l'Ecole Centrale Paris (2009/2010) et Master recherche Mathématiques Vision et Apprentissage de l'ENS Cachan (2009/2010).
**Début du doctorat :** 04/10/10
**Encadrement universitaire :** Ecole doctorale de l'université Paris VII. Thèse dirigée par Josselin Garnier, professeur à Paris VII.
**Encadrement CEA :** Laboratoire de Génie Logiciel et Simulation (LGLS) du CEA de Saclay. Thèse encadrée par Jean-Marc Martinez, chercheur au LGLS.

## 2 Modélisation par processus gaussiens

Les codes de calcul sont largements utilisés dans l'industrie pour compléter ou remplacer les expérimentations. Ces codes de calculs sont la plupart du temps paramétrables. Il est alors souhaitable de quantifier par un formalisme aléatoire d'une part les différences entre les résultats de codes de calcul et les résultats d'expériences et d'autre part l'incertitude que l'on peut avoir sur la bonne paramétrisation du code. Dans ce contexte, le sujet de thèse s'articule autour du modèle suivant :

$$Y(x) = f(x, \beta) + Z(x) + \epsilon.$$

Dans ce modèle, $Y(x)$ est le phénomène observé pour les conditions $x$, $f$ est la fonction code de calcul paramétrée par $\beta$, $Z$ est un processus gaussien et $\epsilon$ est une erreur de mesure (ou bien due à une variabilité intrinséque du phénomène) gaussienne. Dans le cadre de la thèse nous travaillons sous l'hypothèse que le code est linéaire par rapport à $\beta$ et retrouvons donc le modèle de Krigeage (Voir par exemple [2]). Les paramètres $\beta$ peuvent être modélisés constants inconnus (cas fréquentiste) ou bien aléatoires gaussiens (cas bayésien). Enfin, nous considérons que la statistique de la partie aléatoire $Z + \epsilon$ est caractérisée par un hyper-paramètre fini-dimensionnel $\theta \in \Theta$.

## 3 Estimation des hyper-paramètres

Nous nous plaçons dans le cas ou nous disposons d'un vecteur $y$ de $n$ observations du phénomène pour $n$ conditions $x^1, ..., x^n$. L'objectif est, compte tenu de ces observations, d'une part d'estimer le paramètre $\beta$ et d'autre part de prédire le phénomène en un nouveau point $x^0$. Lorsque l'hyper-paramètre $\theta$ est connu, le traitement du problème devient analytique (equations du Krigeage, voir par exemple [2] pour le cas fréquentiste). Nous disposons notamment d'une fonction de prédiction $\langle y^0 \rangle_\theta(y)$ prédisant la valeur du phénomène en $x^0$ en fonction de $y$, avec l'hyper-paramètre $\theta$.

Pour estimer l'hyper-paramètre $\theta$ nous étudions les méthodes de maximum de vraisemblance ([5], [3]) et de validation croisée. Plus précisemment la méthode de validation croisée consiste à utiliser l'estimateur $\hat{\theta}_{CV}$ avec

$$\hat{\theta}_{CV} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} |y_i - \langle y_i \rangle_\theta (y^{-i})|^2,$$

ou $y^{-i}$ est le vecteur $y$ auquel on retire la $i$-éme composante. Le vecteur composé des $y_i - \langle y_i \rangle_\theta (y^{-i})$ est appelé vecteur d'erreurs par validation croisée.
Notons que l'on dispose d'expressions analytiques directes du vecteur d'erreurs par validation croisée ([1], [4]). Ces expressions ne font intervenir que l'hyper-paramètre $\theta$, la matrice des dérivées partielles du code $f$ en les points $x^1, ..., x^n$ et dans le cas bayésien la moyenne et la matrice de covariance a priori de $\beta$. Ainsi, il n'est pas nécessaire d'effectuer $n$ prédictions pour calculer le vecteur d'erreurs par validation croisée.

# 4 Cas d'application

L'application porte sur la calibration du code composant Flica IV développé au CEA pour les études de thermohydraulique des réacteurs nucléaires. La mise en oeuvre des méthodes portera plus particulièrement sur la calibration du modèle de frottement isotherme contenu dans Flica IV. Dans cette étude, la dimension de l'espace des conditions $x$ est 6 et le nombre de paramètres à calibrer est 2. On dispose de $n = 108$ observations.

# Références


[1] O. Dubrule. Cross validation of kriging in a unique neighborhood. Mathematical Geology, 15, 1983.

[2] T.J. Mitchell J. Sacks, W.J. Welch and H.P. Wynn. Design and analysis of computer experiments. Statistical science, 4, 1989.

[3] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71, 1984.

[4] S.S. Keerthi S. Sundararajan. Predictive approaches for choosing hyperparameters in gaussian processes. Neural computation, 13, 2001.

[5] T.J. Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. Annals of statistics, 8, 1980.