Le facteur de Bayes appliqué à la validation des codes de calcul

Guillaume Damblin [1,2] & Merlin Keller [1] & Pierre Barbillon [2] & Alberto Pasanisi [3] & Eric Parent [2]

[1] *guillaume.damblin@edf.fr*
EDF R&D, 6 quai Watier 78 401 Chatou Cedex
[2] *eric.parent.agroparistech@gmail.com*
AgroParisTech, 16 rue Claude Bernard 75 005 Paris

Résumé. Nous présentons dans cet article une nouvelle approche pour la validation d'un code de calcul simulant un système physique d'intérêt. La validation est appréhendée comme un problème de test statistique qui confronte l'hypothèse nulle selon laquelle le code prédit parfaitement le système physique d'intérêt, avec l'hypothèse alternative selon laquelle une erreur systématique subsiste entre le système physique et les prédictions du code. Lorsque le code dépend d'un paramètre inconnu, l'hypothèse nulle correspond à l'existence d'une valeur du paramètre permettant un ajustement parfait du code au système physique, tandis que l'hypothèse alternative correspond à la situation pour laquelle chaque valeur du paramètre définit une fonction d'erreur non nulle entre le code et le système physique. En supposant dans un premier temps que le code de calcul est linéaire par rapport au paramètre, le facteur de Bayes est calculé à partir des mesures physiques disponibles afin de discriminer laquelle des deux hypothèses statistiques est la plus probable. Une attention particulière sera portée au choix des lois *a priori* pour lesquelles nous proposons plusieurs techniques de construction.

Mots-clés. validation, expériences simulées, calage bayésien, facteur de Bayes, lois *a priori*.

Abstract. This paper presents a new way of dealing with the validation of computer codes. In fact, this task can be addressed by testing if the code behaves as a perfect representation of the physical system or if there remains a systematic discrepancy between the code and the reality. When the code is parametrized by an unknown parameter, the first hypothesis means that a value of the parameter makes a perfect agreement between the code and the physical system. Otherwise, each value of the parameter gives a non-zero error function between the code and the reality. By means of physical measurements, Bayes factor is computed for Bayesian model selection by assuming the code is linear with respect to its parameter. We mainly focus on how to build the *prior* distributions using several techniques.

Keywords. validation, computer experiments, Bayesian calibration, Bayes factor, *prior* distributions.

1 Introduction

La validation des codes de calcul complexes rencontre à l'heure actuelle un intérêt grandissant dans les études industrielles du fait de l'utilisation intensive de la simulation numérique pour prédire des systèmes physiques complexes et pour lesquels l'expérience physique est infaisable ou économiquement très coûteuse à mettre en œuvre. La validation peut être définie comme l'action d'évaluer la capacité d'un code de calcul à prédire une quantité physique d'intérêt. Souvent, elle nécessite une étape préalable de calage du code, formalisée ci-dessous.

Soit $r(x) \in \mathbb{R}$ une quantité physique d'intérêt fonction d'un vecteur x de variables mesurables et soit $f_\theta(x)$ un code de calcul paramétré par un vecteur θ. Plusieurs méthodologies statistiques de validation ont été proposées dans la littérature. Elles reposent toutes sur une quantification exhaustive des incertitudes embarquées à l'intérieur du code, propagées ensuite à la réponse $f_\theta(x)$. Souvent, la principale source d'incertitude provient de la méconnaissance du paramètre θ.

Supposons que nous disposions de n mesures y_i du système physique r en plusieurs configurations x_i. Alors, pour $1 \leq i \leq n$, il est naturel d'introduire l'équation

$$y_i = r(x_i) + \epsilon_i \tag{1}$$

dans lequel $\epsilon_i \sim \mathcal{N}(0, \lambda^2)$ est un bruit i.i.d. traduisant l'erreur d'observation. Suivant la méthodologie décrite dans Kennedy et O'Hagan (2001), il n'existe aucun paramétrage permettant au code de prédire exactement $r(x)$. On pose donc,

$$r(x) = f_\theta(x) + b(x) \tag{2}$$

dans lequel b est une fonction d'erreur et θ le paramétrage optimal inconnu. En combinant les équations (1) et (2), on obtient l'équation

$$y_i = f_\theta(x_i) + b(x_i) + \epsilon_i \tag{3}$$

avec une hypothèse *a priori* de processus Gaussien pour b :

$$b(x) \sim \mathcal{PG}(0, \sigma^2 \Sigma_\Psi^x)$$

.

La méthodologie de validation décrite dans Bayarri et al. (2007) combine l'estimation de θ avec la prédiction de r à partir du modèle (3). La caractéristique majeure de cette approche est la prise en compte de l'erreur b de facto, sans justification statistique. Par suite, l'incertitude calculée sur $r(x)$ résulte de l'incertitude sur la réponse du code à laquelle vient s'ajouter la fonction d'erreur b estimée conjointement à θ dans (3), ce qui

peut laisser perplexe quant à la confiance à attribuer au code en termes de représentation du système physique.

Dans la section suivante, nous proposons une définition alternative où la validation s'écrit comme un problème de sélection d'un modèle statistique où l'on teste $b \equiv 0$ contre $b \not\equiv 0$. Si $b \equiv 0$, alors le code est validé car il est capable de prédire le comportement du système "sans erreur".

2 La sélection de modèle pour la validation

Dans la section précédente, la validation est décrite comme le calcul de l'incertitude de prédiction d'un système physique sur la base d'un modèle statistique reliant les mesures physiques disponibles avec la réponse du code de calcul. La méthode de validation que nous proposons ici consiste à détecter à partir des mesures disponibles $y = \{y_1, \cdots, y_n\}$, s'il existe ou non une erreur systématique b entre le code et la réalité, ceci en amont de l'objectif de prédiction. Deux modèles sont donc en compétition pour prédire r en utilisant f_θ. Le premier suppose qu'il existe une valeur θ_0 du paramètre telle que f_{θ_0} prédit parfaitement r. D'où,

$$M_0 \ : \ y_i = f_{\theta_0}(x_i) + \epsilon_i. \tag{4}$$

Le second suppose qu'il existe une fonction d'erreur telle que

$$M_1 \ : \ y_i = f_{\theta_1}(x_i) + b(x_i) + \epsilon_i, \tag{5}$$

avec $b(x) \sim \mathcal{PG}(0, \sigma^2 \Sigma^x_\Psi)$. Considérons les deux modèles M_0 et M_1 caractérisés par leur vraisemblance respective $f_0 = f_0(y|p_0)$ et $f_1 = f_1(y|p_1)$ où $p_0 = (\theta_0, \lambda^2)$ et $p_1 = (\theta_1, \sigma^2, \Psi, \lambda^2)$. Pour chacun des modèles M_0 et M_1,

$$P(M_i|y) \propto P(M_i) \int_{p_i} f_i(y|p_i) \Pi_i(p_i) \mathrm{d}p_i \tag{6}$$

avec $\Pi_i(p_i)$ désignant la loi *a priori* des paramètres et $P(M_1) = 1 - P(M_0)$. Le facteur de Bayes (Bernardo et Smith, 1994) s'écrit comme le rapport des vraisemblances marginales de chacun des modèles conditionnellement à y :

$$B(y) = \frac{P(M_1|y)}{P(M_0|y)} \times \frac{P(M_0)}{P(M_1)} \tag{7}$$

$$= \frac{\int_{p_1} f_1(y|p_1) \Pi_1(p_1) \mathrm{d}p_1}{\int_{p_0} f_0(y|p_0) \Pi_0(p_0) \mathrm{d}p_0}. \tag{8}$$

Dans ce travail, on suppose que le code est linéaire en θ. En considérant $h(x)$ un vecteur de fonctions de régression, le modèle numérique prend donc la forme suivante :

$$f_\theta(x) = h(x)^T \theta \tag{9}$$

Dans la littérature, Bachoc (2014) calibre un code de mécanique des fluides dans un domaine de fonctionnement du code pour lequel l'hypothèse (9) est vérifiée. Sous cette hypothèse et en choisissant la loi *a priori* conjuguée inverse gamma-normale pour le couple (λ^2, θ_0), la probabilité *a posteriori* $P(M_0|y)$ du modèle M_0 est analytique. La probabilité $P(M_1|y)$ doit elle être calculée numériquement. Lorsque $B(y) > 1$, le facteur de Bayes indique que les mesures y sont plus probablement issues du modèle M_1 que du modèle M_0 et inversement si $B(y) < 1$. Néanmoins, une des difficultés provient de la dépendance importante du facteur de Bayes aux lois *a priori* $\Pi_0(p_0)$ et $\Pi_1(p_1)$. Dans la section suivante, nous proposons plusieurs critères pour construire ces lois *a priori* de manière à ne pas biaiser la procédure de sélection en privilégiant un des deux modèles par un choix inapproprié de celles-ci. Notons enfin que si l'hypothèse (9) n'est pas vérifiée, il est toujours possible d'appliquer la procédure de test que nous proposons, sous réserve de temps de calculs plus élevés car $P(M_0|y)$ n'est plus analytique.

3 Application

Nous adaptons le formalisme de la section précédente aux caractéristiques du cas industriel que nous allons traiter. En plus de l'hypothèse de linéarité du modèle, la variance du bruit d'observation λ^2 est supposée connue et le noyau de covariance de b est choisi Gaussien et isotrope. Quitte à normaliser les mesures y en les divisant par λ :

$$Cov(b(x), b(x')) = \mathbb{E}[b(x)b(x')] = \mathbf{1}_{x=x'} + \sigma^2 \exp\Big(- \frac{||x - x'||^2}{\Psi} \Big) \tag{10}$$

Maintenant $p_0 = \theta_0$ et $p_1 = (\theta_1, \sigma^2, \Psi)$. Les paramètres θ_0 et θ_1 jouant le même rôle dans les deux modèles, nous choisissons *a priori*

$$\theta_0 = \theta_1 \sim \mathcal{N}(\beta, V).$$

Comme $P(M_0|y)$ est explicite, calculer $B(y)$ revient à calculer $P(M_1|y)$. Lorsque n est grand, l'approximation de la loi *a posteriori* $\Pi(\theta, \sigma^2, \Psi)$ par une loi normale multivariée permet d'approcher $P(M_1|y)$ par son approximation de Laplace. Lorsque le nombre de mesures y est restreint, on préférera utiliser des techniques d'échantillonnage d'importance, voire de lissage par noyau de $\Pi(\theta, \sigma^2, \Psi)$ si la dimension de p_1 n'est pas trop élevée. Supposons *a priori* que $\sigma^2 \sim \mathcal{IG}(a, b)$ et $\Psi \sim \mathcal{B}(c, d)$. Le facteur de Bayes (7) prend alors la forme suivante :

$$B(y|a, b, c, d) = \int_{\Psi, \sigma^2} BF(y|\Psi, \sigma^2) \Pi_1(\Psi|c, d) \Pi_1(\sigma^2|a, b) \mathrm{d}\Psi \mathrm{d}\sigma^2 \tag{11}$$

où $BF(y, \Psi, \sigma^2)$ est appelé facteur de Bayes local, qui est un terme bien adapté pour étudier la variabilité du facteur de Bayes aux lois *a priori* comme l'illustre la figure 1.

Considérons les lois prédictives *a priori* :

$$\frac{1}{2}m_0(\tilde{y}) + \frac{1}{2}m_1(\tilde{y}|, a, b, c, d), \tag{12}$$

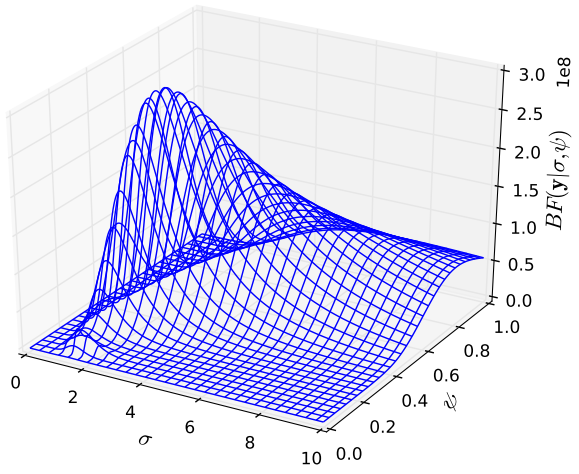
avec

$$m_0(\tilde{y}) = \int f_0(\tilde{y}|\theta_0)\Pi(\theta_0)\mathrm{d}\theta_0, \tag{13}$$

et

$$m_1(\tilde{y}|a, b, c, d) = \int f_1(\tilde{y}|p_1)\Pi(p_1|a, b, c, d)\mathrm{d}p_1. \tag{14}$$

FIGURE 1 – *Facteur de Bayes local*



Le calage des lois *a priori* peut être effectué afin de ne privilégier aucun des deux modèles avant le recueil des mesures y. Supposons que $P(M_0) = P(M_1) = \frac{1}{2}$. On se propose de déterminer (a, b, c, d) avec l' objectif de contrôler le risque de première espèce en maximisant la puissance du test. On veut donc s'assurer que

$$\mathbb{P}[BF(\tilde{y}|a, b, c, d) > 1|M_0] \leq 0.05 \tag{15}$$

où la probabilité est calculée sur la loi prédictive *a priori* du modèle M_0 de densité de probabilité donnée par (13), avec

$$\mathbb{P}[BF(\tilde{y}|a, b, c, d) > 1|M_1] \tag{16}$$

la plus grande possible, où la probabilité est calculée sur la loi prédictive *a priori* du modèle M_1 de densité de probabilité donnée par (14).

Une deuxième approche possible, que nous testons également, consiste à considérer des lois *a priori* non informatives sur chacun des paramètres des deux modèles suivant la méthodologie proposée par Berger et Pericchi (1996).

Cette stratégie de sélection de modèles sera appliquée à EDF pour valider un code de suivi des performances énergétiques d'une installation de production d'électricité.

4 Quelques références bibliographiques

D'un point de vue fréquentiste, le test du rapport de vraisemblance dans le contexte des modèles numériques a été proposé par Loeppky et al. (2006) pour choisir entre M_0 et M_1. En ce qui concerne l'état de l'art des méthodes de calcul du facteur de Bayes, Kass et Raftery (1995) constitue l'article de référence.

Références

Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of gaussian processes. *Journal of Multivariate Analysis*, 125 :1–35.

Bayarri, M. J., Berger, J. O., Sacks, P. R., Cafeo, J. A., Cavendish, J., Lin, C.-H., et Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49 :138–154.

Berger, J. et Pericchi, L. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433) :109–122.

Bernardo, J. M. et Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, London, 1st edition.

Kass, R. et Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430) :773–795.

Kennedy, M. et O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B, Methodological*, 63 :425–464.

Loeppky, D., Bingham, D., et Welch, W. (2006). Computer model calibration or tuning in practice. *Technical Report*.