Plans d'experiences pour la calibration de code de calcul coûteux

Adama Barry [1,2,3] François Bachoc [1] Clémentine Prieur [2]
Sarah Bouquet [3] Miguel Munoz Zuniga [3]

[1] *Institut de Mathématiques de Toulouse,*
[2] *Université de Grenoble,*
[3] *IFP Énergies Nouvelles*
adama.barry@math.univ-toulouse.fr

Résumé. Dans l'industrie, le développement de codes de calcul est souvent mis en œuvre pour étudier et analyser des phénomènes ou des systèmes physiques complexes. Certains de ces codes de calculs dépendent de deux types de variables d'entrées : les variables expérimentales ou de contrôle et les paramètres intrinsèques. Ces paramètres sont souvent des constantes physiques et/ou des paramètres de contrôle qui n'ont pas d'interprétation physique. Des valeurs précises doivent être fixées pour ces paramètres par l'ingénieur afin que le code de calcul imite le plus fidèlement possible le phénomène physique d'intérêt. La nature complexe de ces codes de calculs exige une procédure de calibration efficace, dans laquelle les paramètres inconnus sont ajustés pour améliorer l'alignement entre les sorties du code et les observations physiques. La plupart des travaux sur la calibration bayésienne se concentrent sur la construction d'un émulateur du code de calcul en sélectionnant les expériences numériques sans se préoccuper de la qualité des mesures physiques en amont. Étant donné que ces mesures physiques sont limitées par leur coût ou la difficulté de les acquérir, il serait judicieux de les sélectionner pour une calibration plus efficace.

Sur la base du cadre bayésien classique de Kennedy and O'Hagan [2001], nous proposons une stratégie hybride pour la sélection de plans d'expériences physiques et numériques pour la calibration de code de calcul coûteux. Cette stratégie permet une construction précise de l'émulateur de code de calcul, conduisant à une meilleure approximation de la densité a posteriori des paramètres de calibration et, par conséquent, à des résultats de calibration plus précis.

La première étape consiste à sélectionner les expériences physiques. Nous commencerons donc par présenter des critères permettant de mesurer la qualité d'un plan d'expériences physiques. Ces critères peuvent être regroupés en deux catégories : ceux basés sur la matrice d'information Pronzato and Walter [1985], Fedorov [1980] et ceux basés sur la distribution a posteriori exacte. Ces derniers sont mieux adaptées au problème de calibration car elles tiennent compte de la nature non linéaire du code de calcul, de l'incertitude du phénomène physique et des paramètres. Nous présenterons un algorithme d'optimisation pour la résolution du problème d'optimisation de ces critères.
La seconde étape consiste à sélectionner le plan d'expériences numériques. Nous présenterons des critères tirés de la littérature Damblin et al. [2018], Dai and Chien [2018] et ceux que nous proposons. Nos critères s'inspirent du paradigme de réduction séquentielle de l'incertitude

(SUR) Chevalier et al. [2014]. Ils sont basés sur des mesures d'incertitude pour l'objectif de calibration. Pour leur optimisation, qui est coûteuse, nous utiliserons un algorithme glouton qui exploite la procédure de Monte Carlo utilisée dans leur calcul. Une étude comparative sur un cas analytique sera présentée à la fin pour illustrer la performance des différents algorithmes.

Mots-clés. Calibration bayésienne, processus gaussien, quantification d'incertitudes, plans d'expériences physiques, plans d'expériences numériques, divergence de Kullback-Leibler.

Abstract. In industry, computer codes are often developed to study and analyse complex physical phenomena or systems. Some of these computer codes depend on two types of input variables : experimental or control variables and intrinsic parameters. These parameters are often physical constants and/or control parameters that have no physical interpretation. Precise values must be set for these parameters by the engineer so that the computer code imitates the physical phenomenon of interest as closely as possible. The complex nature of these computer codes requires an efficient calibration procedure, in which the unknown parameters are adjusted to improve the alignment between the code outputs and the physical observations. Most work on Bayesian calibration focuses on building an emulator of the computer code by selecting numerical experiments without regard to the quality of the upstream physical measurements. Given that these physical measurements are limited by their cost or the difficulty of acquiring them, it would make sense to select them for more effective calibration.

The first step is to select the physical experiments. We will therefore begin by presenting the criteria for measuring the quality of a design of physical experiments. These criteria can be grouped into two categories : based on the information matrix (Pronzato and Walter [1985], Fedorov [1980]) and the exact a posteriori distribution. The latter are better suited to the calibration problem because they take into account the non-linear nature of the computer code, the uncertainty of the physical phenomenon, and the uncertainty of the parameters. We will present an optimization algorithm for solving the problem of optimizing these criteria. The second stage consists of selecting the design of numerical experiments. We will present criteria taken from the literature and those that we propose. Our criteria are inspired by the sequential uncertainty reduction (SUR) paradigm Chevalier et al. [2014]. They are based on uncertainty measurements for the calibration objective. For their optimization, which is costly, we will use a greedy algorithm that exploits the Monte Carlo procedure used in their calculation. A comparative study on an analytical case will be presented at the end to illustrate the performance of the different algorithms.

Keywords. Bayesian calibration, Gaussian process, uncertainty quantification, designs of physical experiments, designs of numerical experiments, Kullback-Leibler divergence.

1 Introduction

Le code de calcul d'intérêt est représenté par une fonction paramétrique dépendant de deux types d'entrées : un vecteur de variables de contrôle désigné par $x \in \mathcal{X} \subset \mathbf{R}^d$ et un vecteur de paramètres $\theta \in \Theta \subset \mathbf{R}^p$ appelés paramètres de calibration.
Sur la base du cadre bayésien classique de Kennedy and O'Hagan [2001] (KOH), la relation entre le code de calcul et le système physique est donnée par le modèle statistique suivant

$$\mathbf{Y}_{obs}(x) = f_{code}(x, \theta_0) + \varepsilon_x, \tag{1}$$

où $\theta_0 \in \Theta$ est la vraie valeur du vecteur des paramètres et $\varepsilon_x \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ est l'erreur de mesure. Nous supposons que la variance σ_ε^2 est connue. Le cas échéant, elle peut être considérée comme un paramètre d'incertitude supplémentaire incroporé dans le cadre bayésien ci-dessous.
Notons par X le plan d'expériences physiques et Y les mesures physiques correspondantes. On modélise notre connaissance a priori sur le vecteur de paramètres de calibration par une loi a priori représentée par la densité de probabilité $\pi_0 : \theta \in \Theta \longmapsto \pi_0(\theta) \in \mathbf{R}_+$. La loi a priori est mise a jour, par la règle de Bayes, à l'aide des observations physiques pour donner la densité a posteriori

$$\begin{aligned}
\pi(\theta \mid Y_{obs}) &= \frac{\mathcal{L}(Y_{obs} \mid \theta)\pi_0(\theta)}{Z} \\
&\propto \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} \big(y_i - f_{code}(x^{(i)}, \theta)\big)^2 \Big] \pi_0(\theta) \\
&\propto \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma_\varepsilon^2} SS(\theta) \Big] \pi_0(\theta),
\end{aligned} \tag{2}$$

où $\mathcal{L}(Y_{obs} \mid \theta)$ représente la vraisemblance, $Z = \int_\Theta \mathcal{L}(Y_{obs} \mid u)\pi_0(u)du$ la constante de normalisation et $SS(\theta) = \sum_{i=1}^{n} \big(y_i - f_{code}(x^{(i)}, \theta)\big)^2$ la somme des écarts aux carrés.

Pour estimer les paramètres de calibration, nous devons échantillonner la distribution a posteriori (2) par la méthode de Monte-Carlo par chaînes de Markov (MCMC). Cela nécessite un grand nombre d'appels au code de calcul. Ce dernier étant coûteux, cette approche n'est pas réalisable. La solution consiste à approximer la densité a posteriori à l'aide d'un émulateur du code de calcul.

2 Émulation par processus gaussien et approximation de la densité a posteriori

On pose un a priori sur le code de calcul comme étant la réalisation d'un processus gaussien

$$\mathrm{Y}_{code} \sim \mathbf{GP}(m_\beta, k_\psi), \tag{3}$$

où $m : u \in \mathcal{X} \times \Theta \longmapsto m(u) \in \mathbf{R}$ est la fonction moyenne a priori et $k_\psi : (u, v) \in (\mathcal{X} \times \Theta)^2 \longmapsto k_\psi(u, v) \in \mathbf{R}$ est la fonction de covariance a priori avec β et ψ le vecteur des paramètres de régressions et le vecteur des hyperparamètres.

Supposons que l'on dispose de $D_M = \Big((x_1,\theta_1),\ldots,(x_M,\theta_M)\Big)^T$ le plan d'expériences numériques et $f_{code}(D_M) = \Big(f_{code}(x_1,\theta_1),\ldots,f_{code}(x_M,\theta_M)\Big)^T$ les observations numériques correspondantes. Le processus gaussien conditionné aux observations numériques reste également gaussien :

$$Y^M_{code} := \Big[Y_{code} \mid Y_{code}(D_M) = f_{code}(D_M)\Big] \sim \mathbf{GP}\Big(\mu_M, k_M\Big), \tag{4}$$

où μ_M et k_M representent la fonction moyenne a posteriori et la fonction de covariance a posteriori dont les expressions suivent :

$$\mu_M(\upsilon) = m_\beta(\upsilon) - k(\upsilon, D_M)\big[k(D_M)\big]^{-1}\big[f_{code}(D_M) - m_\beta(D_M)\big], \tag{5}$$

$$k_M(\upsilon,\upsilon') = k(\upsilon,\upsilon') - k(\upsilon, D_M)\big[k(D_M)\big]^{-1}k(D_M,\upsilon'), \tag{6}$$

et

$$m_\beta(D_M) = \Big(m_\beta(x_i,\theta_i)\Big)_{i=1,\ldots,M} \quad , \quad k(\upsilon, D_M) = \Big(k\big(\upsilon,(x_i,\theta_i)\big)\Big)_{i=1,\ldots,M},$$
$$k(D_M,\upsilon') = \Big(k\big((x_i,\theta_i),\upsilon'\big)\Big)_{i=1,\ldots,M} \quad , \quad k(D_M) = \Big(k\big((x_i,\theta_i),(x_j,\theta_j)\big)\Big)_{i,j=1,\ldots,M}.$$

Les paramètres du modèle de processus gaussien (β,ψ) sont omis dans les notations par souci de simplicité. Ils peuvent être estimés à l'aide des observations numériques par la technique de modularisation (voir Damblin et al. [2018]). Une hypothèse supplémentaire d'indépendance entre la loi a priori du vecteur des paramètres et celle du code de calcul conduit à l'approximation de la densité a posteriori donnée comme suit :

$$\begin{aligned}\pi(\theta \mid Y_{obs}, f_{code}(D_M)) &= \frac{\mathcal{L}^c(Y_{obs} \mid f_{code}(D_M),\theta)\pi(\theta)}{\int_\Theta \mathcal{L}^c(Y_{obs} \mid f_{code}(D_M),u)\pi(\theta')du} \\ &\propto \mathcal{L}^c(Y_{obs} \mid f_{code}(D_M),\theta)\pi(\theta),\end{aligned} \tag{7}$$

où

$$\mathcal{L}^c\big(Y_{obs} \mid f_{code}(D_M,\theta)\big) = \frac{1}{(2\pi)^{n/2} \mid W_M(\theta) \mid^{1/2}} \times \exp\Big[-\frac{1}{2}\big(Y_{obs} - \mu_M(X_{obs},\theta)\big)^T\big[W_M(\theta)\big]^{-1}\big(Y_{obs} - \mu_M(X_{obs},\theta)\big)\Big]$$

est la vraisemblance conditionnée aux observations numériques, avec $W_M(\theta) = \sigma^2_\varepsilon I_n + k_M\big((X_{obs},\theta),(X_{obs},\theta)\big)$.

De l'équation (7), nous pouvons noter que la qualité de l'approximation dépend de la qualité de l'émulateur qui à son tour dépend des observations numériques donc du plan d'expériences numériques. De plus la qualité de la densité a posteriori que l'on cherche à approximer (2) dépend de la qualité des observations physiques. D'où la nécessité de sélectionner avec soin le plan d'expériences physiques et le plan d'expériences numériques pour une calibration efficace.

3 Algorithme hybride pour la calibration de code de calcul

L'algorithme hybride que nous proposons pour la calibration se déclinent en cinq étapes principales :

1. Construire un émulateur de processus gaussien initial à l'aide d'un plan d'expériences numériques D_{M_0} et les observations numériques correspondantes $f(D_{M_0})$.
2. Utiliser l'émulateur pour sélectionner le plan d'expériences physiques

$$X_{obs} \in \arg\max \mathbf{C}(X), \tag{8}$$

   où C est un critère de sélection qui mesure la quantité d'information contenue dans un plan d'expériences physiques.
3. Effectuer les mesures sur le terrain et recueillir les observations physiques Y_{obs}.
4. Pour $k = M_0, \ldots, M$:
   — Sélectionner en deux étapes :

$$\theta_{k+1} \in \arg\max_{\Theta} \alpha_k(\theta) \tag{9}$$

$$x_{k+1} \in \arg\max_{X_{obs}} \beta_k(x, \theta_{k+1}), \tag{10}$$

   où α_k et β_k sont des fonctions d'acquisitions.
   — Enrichir le plan d'expériences numériques $D_{k+1} = D_k \cup \{(x_{k+1}, \theta_{k+1})\}$ et les observations numériques $f_{code}(D_{k+1}) = \left(f_{code}(D_{k+1})^T, f_{code}(x_{k+1}, \theta_{k+1})\right)^T$.
   — Mettre à jour l'émulateur de processus gaussien.
5. Approximation de la distribution a posteriori, échantillonnage MCMC et estimation des intervalles de crédibilité des paramètres de calibration.

L'exposé se concentrera sur les critères de la littérature basés sur la matrice d'information (Pronzato and Walter [1985], Fedorov [1980]) et ceux que nous proposons qui utilisent la distribution a posteriori pour mesurer la qualité d'un plan d'expériences physiques.
Pour la sélection séquentielle de plans d'expériences numériques, nous présenterons les stratégies de Damblin et al. [2018], Dai and Chien [2018] et Gardner et al. [2019] et celles que nous proposons inspirées du paradigme de réduction séquentielle d'incertitudes (Chevalier et al. [2014]). Des algorithmes d'optimisation adaptés aux problèmes (8), (9) et (10) seront également présentés. Et enfin des illustrations et des résultats des performances des algorithmes sur des cas test analytiques seront présentés.

Références

Clément Chevalier, David Ginsbourger, Victor Picheny, Julien Bect, Emmanuel Vazquez, and Yann Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4) :455–465, 2014. ISSN 00401706, 15372723. URL http://www.jstor.org/stable/24587032.

Xiaowu Dai and Peter Chien. Another look at statistical calibration : A non-asymptotic theory and prediction-oriented optimality, 2018. URL https://arxiv.org/abs/1802.00021.

Guillaume Damblin, Pierre Barbillon, Merlin Keller, Alberto Pasanisi, and Éric Parent. Adaptive numerical designs for the calibration of computer codes. *SIAM/ASA Journal on Uncertainty Quantification*, 6(1) :151–179, Jan 2018. ISSN 2166-2525. doi : 10.1137/15m1033162. URL http://dx.doi.org/10.1137/15M1033162.

Valery V. Fedorov. Convex design theory 1. *Statistics*, 11 :21–43, 1980.

Paul Gardner, Charles Lord, and Robert Barthorpe. Sequential bayesian history matching for model calibration. 05 2019. doi : 10.1115/VVS2019-5149.

Marc Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, 63 :425–464, 02 2001. doi : 10.1111/1467-9868.00294.

Luc Pronzato and Eric Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1) :103–120, 1985. ISSN 0025-5564. doi : https ://doi.org/10.1016/0025-5564(85)90068-9. URL https://www.sciencedirect.com/science/article/pii/0025556485900689.