Journée des doctorants 2011

Nabil Rachdi

EADS IW Suresnes - Université Toulouse III

Contexte général de la thèse.

Ma thèse porte sur l'analyse mathématique de la modélisation des incertitudes par simulation. Elle a débuté en Septembre 2008 sous la direction de Jean-Claude Fort (Université Paris V), de Thierry Klein (Université Toulouse III) et également encadrée par Fabien Mangeant (EADS IW Suresnes).
Le problème général est la prise en compte de toute l'information disponible pour modéliser les incertitudes. En particulier, nous nous plaçons dans le cadre où l'on dispose d'un nombre limité de données expérimentales et d'un (ou plusieurs) modèle sensé représenter le phénomène qui nous intéresse. Ce cadre est très représentatif des applications EADS, notamment par l'utilisation de plateformes de simulation multi-échelles pour la conception virtuelle d'avions. Cette situation est présente dans l'ingénierie en général, où plusieurs modèles numériques coexistent pour une même simulation, ainsi que des données expérimentales sur les entrées et sorties de ces modèles, pas nécessairement appariées, coûteuses à obtenir.
On se pose la question de savoir comment la simulation peut contribuer à améliorer notre connaissance (données expérimentales) sur l'incertitude d'un phénomène donné.

Résumé.

On modélise un phénomème complexe par une variable aléatoire $Y \in \mathcal{Y} \subset \mathbb{R}$ de mesure Q à (Lebesgue) densité $\rho^\star$ inconnue, et on suppose disposer d'un n-échantillon $Y_1, ..., Y_n$ (données expérimentales). L'information disponible sur Y est bien souvent insuffisante (n petit) pour mener à bien les deux principaux problèmes:

- *Prédiction* : l'étude d'une quantité d'intérêt relative à la variable Y,
- *Problème Inverse* : estimer un paramètre à partir des observations $Y_1, ..., Y_n$.

Dans les faits, le problème de la *prédiction* nécessite l'estimation d'un paramètre $\boldsymbol{\theta}$ et on peut abusivement dire que ce problème étend le second (bien qu'étant philosophiquement différent). L'approche de telles performances est faite à l'aide de modèles mathématiques $h \in \mathcal{H}$ (boîte noire etc...) pouvant être plus ou moins "fidèles" et/ou "complexes" vis-à-vis du phénomène Y:

$$\begin{array}{rcl} h : (\mathcal{X}, \mathcal{B}, P) \times \Theta & \longmapsto & \mathcal{Y} \\ (\mathbf{X}, \boldsymbol{\theta}) & \longrightarrow & h(\mathbf{X}, \boldsymbol{\theta}) \end{array}$$

où $\mathcal{X} \subset \mathbb{R}^d$, P une mesure de probabilité et $\Theta \subset \mathbb{R}^k$ est l'espace des paramètres. La mesure P ne nécessite pas d'être connue, mais on suppose avoir un m-échantillon $\mathbf{X}_1, ..., \mathbf{X}_m$, où $m = m_t + m_s$ (m_t: apprentissage, m_s: simulation), avec m grand devant n.
Nous proposons une méthode pour l'estimation d'un paramètre $\boldsymbol{\theta} \in \Theta$ pour la *prédiction* (d'une quantité d'intérêt) ou un *problème inverse*.
Dans ce papier, nous appliquons nos résultats au cas académique suivant:

- $Y = \sin(\xi) + 0.01\,\epsilon\,, \quad \xi, \epsilon \sim \mathcal{N}(0, 1)$

- $h(\mathbf{X}, \boldsymbol{\theta}) = \boldsymbol{\theta}_1 \mathbf{X} + \boldsymbol{\theta}_2 \mathbf{X}^3\,, \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^T, \quad \mathbf{X} \sim \mathcal{N}(0,1)\,.$
- $n = 50$, $m = 2.10^4$ $(m_t = 10^4\ ,\ m_s = 10^4)$

L'étape commune à la prédiction et au problème inverse est l'estimation d'un paramètre $\theta \in \Theta$, nous proposons l'algorithme d'apprentissage suivant:

$$\widehat{\boldsymbol{\theta}}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Psi \left(\sum_{j=1}^{m_t} \widetilde{\rho}_{\mathcal{F}}(h(\mathbf{X}_j, \boldsymbol{\theta}))\,,\, Y \right)\,,$$

où $\mathcal{F}$ est un espace fonctionnel, $\Psi : \mathcal{F} \to L_1(Q)$ est un $\mathcal{F}$-contraste que l'on définira et $\widetilde{\rho}_{\mathcal{F}} : \mathcal{Y} \to \mathcal{F}$ une fonction poids.
Nous montrons des résultats non asymptotiques sur les performances de cet algorithme (inégalités oracles) dans un cadre général. Dans le cas académique cité plus haut, on s'intéresse à la prédiction de la densité de la variable Y. On représente sur la figure (1) les trois densités suivantes: celle de Y, $\rho^\star$ (en rouge) par MC intensif, celle reconstruite par noyaux à partir des données expérimentales $Y_1, ..., Y_n$ (en vert), et celle reconstruite à noyaux à partir des m_s données simulées sous $\widehat{\boldsymbol{\theta}}_\Psi$: $h(\mathbf{X}_j, \widehat{\boldsymbol{\theta}}_\Psi)$, $j = m_t + 1, ..., m$ (en bleu).

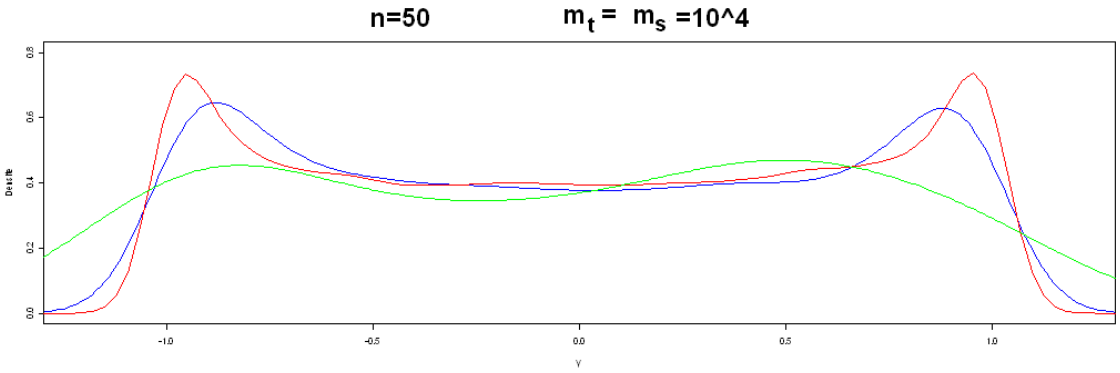

Figure 1: Comparaison des densités de probabilité

Nous étudions les conditions sur la modélisation (modèle h, n, m etc ...) permettant d'améliorer la prédiction d'une quantité d'intérêt basée uniquement sur les données (expérimentales) $Y_1, ..., Y_n$. Nous analysons également les *procédures croisées* en prédiction, i.e quand l'estimation du paramètre $\boldsymbol{\theta}$ (par ex. regression) n'a "pas de lien" avec la quantité d'intérêt recherchée (par ex. probabilité de dépassement).

Bibliographie.

N. Rachdi, JC Fort, T. Klein (submitted 2010), *Oracle inequalities for new M-estimation and model selection problems*
N. Rachdi *et al* (2010), *Modeling uncertainties in Complex Systems* (Proceedings of the Sixth International Conference on Sensitivity Analysis of Model Output , Milan)
N. Rachdi (in preparation), *Learning with computer codes under uncertainties*
A.W van der Vaart (1998). *Asymptotics Statistics.*
A.W van der Vaart, Jon A. Wellner (1996). *Weak Convergence and Empirical Processes.*