GDR MASCOT NUM version du 10/06/08
Sujet de thèse pour septembre 2009

Plan d'expériences et estimation en régression *sparse*

J. Bigot (IMT Toulouse), F. Gamboa (IMT Toulouse), B. Sudret (EDF Clamart)

Cadre de travail

On considère un modèle de régression scalaire où la réponse est du type

$$Y = \langle \alpha^*, \phi(x) \rangle + \varepsilon, \quad (x \in U), \tag{1}$$

où

- U est l'espace des variables exogènes (variables d'entrée). Typiquement U est un sous-ensemble de $\mathbb{R}^l$ ($l \geq 1$) avec l grand. Par exemple $U := \{-1, 1\}^l$ dans le cas des modèles factoriels à deux niveaux ou $U := [0, 1]^l$ dans le cas de la régression polynomiale multidimensionnelle sur un compact.
- $\phi : \mathbb{R}^l \longmapsto \mathbb{R}^k$ ($k \geq 1$) est une fonction donnée. Par exemple, dans le cas d'un modèle factoriel à deux niveaux les composantes de ϕ sont des monômes obtenus par multiplication d'un sous-ensemble des variables d'entrée. Dans le cas de la régression multidimensionnelle sur un compact, les composantes de ϕ sont des polynômes.
- ε est l'erreur de mesure qui peut par exemple être modélisée par une variable aléatoire gaussienne.

Le modèle est observé sur n points de l'espace des variables exogènes : $x_1, \ldots, x_n$. On se place dans le cas où le nombre d'observations n est beaucoup plus petit que le nombre de fonctions de régression l. Sans hypothèse supplémentaire, il est évidemment vain d'espérer pouvoir développer des méthodes de statistique inférentielle sur le paramètre inconnu α^*. On se place ici dans le cadre dit *sparse* ou le nombre de composantes non nulles du paramètre α^* est moindre qu'une fraction du nombre des obsevations. C'est-à-dire que l'on suppose que l'ensemble des paramètres est :

$$\Theta := \left\{ \alpha = (\alpha_1, \cdots, \alpha_k)^T \in \mathbb{R}^k : \ \sharp\{j \in \{1, \cdots, k\} : \ \alpha_j \neq 0\} \leq \kappa n \right\}, \tag{2}$$

où $0 < \kappa < 1$, est une constante donnée. Le calcul de l'estimateur des moindres carrés contraint à cet ensemble est malheureusement un problème numérique de complexité exponentielle et il n'est pas possible de mettre en oeuvre directement de l'inférence statistique sur cet ensemble de paramètres. Il est néanmoins possible de contourner cette difficulté en considérant l'estimateur des moindres carrés non contraint de norme l^1 minimum. En effet, Candes *et al* ([4]) et Donoho ([8]) ont montré que la résolution exacte ou stable d'un système linéaire sous-déterminant sur Θ pouvait *souvent* être obtenu en minimisant la norme l^1. *Souvent* signifie ici que si la matrice du système linéaire est choisie au hasard, alors il est possible de résoudre exactement ou de façon stable ce système pourvu que l'on s'intéresse à une solution appartenant à Θ. En fait, on est dans le cadre d'un phénomène connu dans les problèmes inverses mal posés appelé *Super-résolution* (voir [7], [9], [11], [10]).

Programme de recherche

Plan d'expériences adaptés

Un premier axe de recherche s'articulera autour du problème de la détermination et la construction effective de plan d'éxpériences adaptés au problème d'estimation dans un modèle de régresssion paramétrique *sparse*.

Problème inverse mal posé

On replacera le modèle linéaire *sparse* précédent dans le cadre des problèmes inverses mal posés rencontrés en traiterment du signal et développés dans [11] et [10]. Dans ce contextes, on étudiera les qualités de la reconstruction par la méthode l^1.

Algorithmes pour la recherche d'une solution sparse

Les problèmes de minimisation convexe avec une contrainte de type ℓ_1 peuvent se résoudre à partir d'algorithmes itératifs qu'on retrouve sous l'appelation algarithmes greedy ou méthodes d'aggrégation dans la littérature. Ces algorithmes ont été à l'origine développés dans la communauté des problèmes inverses [6] et connaissent actuellement un large développement en statistique mathématique [1], [5]. Dans le contexte des plans d'expériences, ce type d'algorithme pour la recherche d'une solution sparse reste à développer et à étudier théoriquement.

Application aux métamodèles par chaos polynomial

Les développements par chaos polynomial permettent de représenter la réponse aléatoire $Y = \mathcal{M}(X)$ d'un modèle $\mathcal{M} : \mathbb{R}^M \longrightarrow \mathbb{R}^N$ dont les entrées sont modélisées par un vecteur aléatoire X de densité p_X donnée [12] :

$$Y = \mathcal{M}(X) = \sum_{\alpha \in \mathbb{N}^M} y_\alpha \, \Psi_\alpha(X) \tag{3}$$

Le calcul des coefficients $\{y_\alpha, \alpha \in \mathbb{N}^M\}$ peut se poser comme un problème de régression [2], qui s'avère la plupart du temps *sparse* [3].

On propose d'investiguer la possibilité d'appliquer les algorithmes de régression l^1 (Dantzig selector, LARS, etc.) à ces problèmes, pour lesquels le nombre de régresseurs est potentiellement infini. Le choix de plans d'expériences adaptés à ce problème rejoint le premier point du programme de recherche.

EDF R&D s'intéresse aux métamodèles pour la propagation d'incertitudes dans la modélisation de systèmes industriels complexes, notamment mécanique. La mise en œuvre d'approches *sparse* en régression pourrait être appliqué aux calculs par éléments finis stochastiques de structures industrielles (enceintes de confinement de centrales nucléaires, organes de robinetterie, etc.).

Références

[1] Wolfgang Dahmen Andrew R. Barron, Albert Cohen and Ronald A. DeVore. Approximation and learning by greedy algorithms. *Annals of statistics*, 36 :64–94, 2008.

[2] M. Berveiller, B. Sudret, and M. Lemaire. Stochastic finite elements : a non intrusive approach by regression. *Eur. J. Comput. Mech.*, 15(1-3) :81–92, 2006.

[3] G. Blatman and B. Sudret. Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique*, 336 :518–523, 2008.

[4] Emmanuel Candes and Terence Tao. The Dantzig selector : statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6) :2313–2351, 2007.

[5] A. Dalalyan, A. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *to appear in Machine Learning*, 2008.

[6] M. Daubechies, I. Defrise and De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57 :1413–1457, 2004.

[7] David L. Donoho. Superresolution via sparsity constraints. *SIAM J. Math. Anal.*, 23(5) :1309–1331, 1992.

[8] David L. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6) :797–829, 2006.

[9] David L. Donoho, Iain M. Johnstone, Jeffrey C. Hoch, and Alan S. Stern. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, 54(1) :41–81, 1992. With discussion and a reply by the authors.

[10] F. Gamboa and É. Gassiat. The maximum entropy method on the mean : applications to linear programming and superresolution. *Math. Programming*, 66(1, Ser. A) :103–122, 1994.

[11] F. Gamboa and E. Gassiat. Sets of superresolution and the maximum entropy method on the mean. *SIAM J. Math. Anal.*, 27(4) :1129–1152, 1996.

[12] C. Soize and R. Ghanem. Physical systems with random uncertainties : chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput.*, 26(2) :395–410, 2004.