Plans d'expériences et méthodologie des surfaces de réponse pour données fonctionnelles

Angelina ROCHE
Laboratoire MAP5 UMR CNRS 8145 Université Paris Descartes 45, rue des Saints Pères 75270 Paris cedex 06 angelina.roche@parisdescartes.fr

Résumé. La méthodologie des surfaces de réponse est aujourd'hui une méthode classique utilisée en ingénierie pour optimiser une réponse réelle (par exemple un rendement ou la probabilité de défaillance d'un matériau) dépendant de plusieurs covariables. Issue des travaux de Box et Wilson (1951), elle a depuis fait l'objet d'un intérêt constant, motivé par la variété des applications possibles. Nous proposons dans cette contribution une adaptation de cette méthodologie au cadre fonctionnel c'est-à-dire que nous cherchons à optimiser une variable d'intérêt dépendant d'une ou plusieurs fonctions. Nous illustrerons le fonctionnement de la méthode sur des données simulées ainsi que sur une application à la sûreté nucléaire.

Mots-clés. Surface de réponse, données fonctionnelles, plans d'expérience.

Abstract. The response surface methodology is a classical engineer method. Its aim is to optimize an output variable (the yield of a chemical reaction or the probability of failure of a material) which depends on several covariates. The method has been proposed by Box and Wilson (1951) and has since attracted a constant interest, motivated by a huge variety of applications. We propose a way to adapt response surface methodology to a functional data framework that is to say that we want to optimise a response depending on several curves. The method we propose will be illustrated on simulated datasets and on an application to nuclear safety.

Keywords. Response surface methodology, functional data analysis, design of experiment.

1 Motivations pour une adaptation à un contexte fonctionnel

La méthodologie des surfaces de réponse a pour but d'explorer les relations entre une variable réponse y et plusieurs variables dépendantes $x_1, ..., x_d$ par ajustement d'une fonction mathématique, dans l'objectif notamment d'optimiser la réponse y. Elle s'est révélée extrêmement utile dans de nombreux domaines. Plus récemment, avec le développement de codes de calcul qui peuvent être très consommateurs en temps, cette méthode a été étendue aux expérimentations numériques (Sacks *et al.*, 1989) et a été largement utilisée dans l'industrie, par exemple pour optimiser la conception de produits manufacturés,

comme des circuits électriques (Bates *et al.*, 1996) ou des pales de rotor (Lee et Hajela, 1996).

Le principe de la méthode est de trouver les conditions d'expérimentation optimales en réalisant un nombre restreint d'expériences. Le point de départ est une certaine région de l'espace $\mathbb{R}^d$ (les conditions d'expérimentation actuelles que l'on cherche à optimiser par exemple) dans laquelle une série d'expériences sont réalisées, en suivant un plan d'expériences choisi par l'utilisateur. Les observations ainsi obtenues nous permettent d'avoir une certaine idée de la forme de la surface $y = m(x_1, \ldots, x_d)$ dans cette région. Nous pouvons utiliser ainsi cette connaissance pour estimer la direction de plus forte descente (ou de plus forte croissance suivant que l'on cherche à maximiser ou minimiser m) de cette surface. Le long de cette direction, une série d'expériences est effectuée jusqu'à un point de $\mathbb{R}^d$ où la réponse est considérée optimale. D'autres expériences peuvent ensuite être réalisées autour de ce point pour affiner la position de l'optimum.

Depuis les travaux de Box et Wilson (1951), de nombreuses améliorations ont été faites principalement sur deux aspects : la modélisation de la surface dans la région considérée et le choix du plan d'expériences.

La manière classique de modéliser la forme de la surface dans la région d'intérêt est de considérer des modèles de régression polynomiaux, souvent de degré 1

$$Y = \alpha + \sum_{j=1}^{d} \beta_j x_j + \varepsilon,$$

avec $\alpha \in \mathbb{R}$, $(\beta_1, \ldots, \beta_j) \in \mathbb{R}^d$ et $\varepsilon \sim \mathcal{N}(0, 1)$, ou de degré 2

$$Y = \alpha + \sum_{j=1}^{d} \beta_j x_j + \sum_{j,k=1}^{d} \beta_{j,k} x_j x_k + \varepsilon,$$

avec $\beta_{j,k} \in \mathbb{R}$ pour tous $j, k = 1, \ldots, d$. Plus récemment, des modèles plus complexes ont été considérés, comme des modèles linéaires généralisés (voir les références citées dans Khuri, 2001) ou des modèles non-paramétriques (Facer et Müller, 2003), de façon à améliorer l'estimation de la surface $y = m(x_1, \ldots, x_d)$, en particulier lorsque la fonction m est irrégulière. Notons que plus le modèle est complexe et plus le nombre d'expériences à réaliser pour estimer ses paramètres correctement est important.

De façon analogue au choix du modèle de régression, le choix du plan d'expériences doit répondre à deux critères antagonistes : d'une part le nombre d'expériences à réaliser doit être le plus petit possible de façon à minimiser les coûts, d'autre part l'estimation de la surface doit être la plus précise possible. Les plans d'expériences classiques sont les plans factoriels, les plans composites centrés (CCD) et de Box-Benhken. De nombreux autres plans d'expériences existent, nous renvoyons à Georgiou *et al.* (2014) pour les avancées les plus récentes sur ce sujet et à Khuri et Mukhopadhyay (2010) pour un état de l'art et la description des plans les plus classiques.

Cependant, aucune méthode n'existe pour optimiser une variable de sortie lorsqu'une des covariables est une fonction ou une courbe. Or les besoins sont réels tant du point de vue de l'optimisation de sorties de codes de calcul que de résultats d'expériences physiques.

Par exemple, la probabilité de défaillance de la cuve d'un réacteur nucléaire lors d'un accident de perte de réfrigérant primaire est liée à l'évolution de la température, de la pression et de l'effusivité thermique (qui mesure la capacité du matériau à échanger de la chaleur avec son environnement par contact) dans le coeur du réacteur. Ces paramètres sont contrôlés par l'injection d'eau froide dans le circuit primaire. Une meilleure compréhension de ce lien permettrait d'améliorer la procédure à suivre pour minimiser les dégâts causés par ce type d'accident.

Hormis les applications industrielles, d'autres applications sont envisageables, comme des applications médicales. Il a été mis en évidence par exemple par Glaser *et al.* (2001, 2013) que la probabilité de survenue d'un oedème cérébral chez les enfants hospitalisés pour acidocétose diabétique (qui est une complication du diabète de type I) dépend de la manière dont le traitement est administré et non seulement de la quantité administrée.

2 Plans d'expérience et données fonctionnelles

Nous proposons une méthode couplant réduction de la dimension avec des plans d'expériences multivariés classiques. L'idée (simple) pour générer un plan d'expérience autours d'un point $x_0 \in \mathbb{H}$ (avec $\mathbb{H}$ un espace de Hilbert séparable) est la suivante : nous choisissons une base d'approximation $(\varphi_j)_{j\geq 1}$ de $\mathbb{H}$, une dimension d'approximation d et un plan d'expérience dans $\mathbb{R}^d$ $\{\mathbf{x}_i, i = 1, \ldots, n\} = \{(x_{i,1}, \ldots, x_{i,d}), i = 1, \ldots, n\}$ autour du point $0 \in \mathbb{R}^d$. Le plan d'expérience $\{x_i, i = 1, \ldots, n\}$ est défini ensuite de la façon suivante :

$$x_i := x_0 + \sum_{j=1}^{d} x_{i,j}\varphi_j.$$

Cette méthode a pour avantage d'être flexible : tous les types de plans d'expérience multivariés peuvent être considérés ainsi que toutes les bases d'approximation de $\mathbb{H}$. Suivant le contexte, il est possible de choisir une base d'approximation fixe telle que la base de Fourier, une base de splines ou d'ondelettes. Lorsqu'un échantillon d'apprentissage est disponible $\{(X_i, Y_i), i = 1, \ldots, n\}$, nous pouvons également utiliser l'information de cet échantillon pour définir une base adaptée aux données. Deux exemples de telles bases sont les suivantes :

- La base de l'ACP très appréciée pour ses propriétés d'optimalité puisqu'elle vérifie

$$\frac{1}{n}\sum_{i=1}^{n} \|X_i - \widehat{\Pi}_d X_i\|^2 = \min_{\Pi_d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \|X_i - \Pi_d X_i\|^2 \right\},$$

où $\widehat{\Pi}_d$ est la projection orthogonale sur span$\{\varphi_1, \ldots, \varphi_d\}$ et le minimum du membre de droite est pris sur tous les projecteurs orthogonaux sur des sous-espaces de $\mathbb{H}$ de dimension d.

- La base PLS (Wold, 1975, Preda et Saporta, 2005) qui, contrairement à la base de l'ACP qui n'est calculée qu'à partir de X, permet de prendre en compte l'interaction entre X et Y. Nous renvoyons à Delaigle et Hall (2012) pour la définition de l'algorithme de calcul de la base PLS dans une contexte fonctionnel et des résultats théoriques.

3 Application à la méthodologie des surfaces de réponse

L'algorithme proposé est directement inspiré de la version classique, multivariée, de la méthodologie des surfaces de réponse.

1. Génération d'un plan d'expériences fonctionnel $(x_i^{(0)}, i = 1, \ldots, n_0)$ dans une certaine région de l'espace et réalisation des expériences correspondantes (les résultats sont notés $(Y_i^{(0)}, i = 1, \ldots, n_0)$).

2. Estimation des paramètres d'un modèle d'ordre 1 : $Y = \alpha + \langle \beta, x \rangle + \varepsilon$ (avec $\alpha \in \mathbb{R}$ et $\beta \in \mathbb{H}$) à l'aide de l'échantillon $((x_i^{(0)}, Y_i^{(0)}), i = 1, \ldots, n)$ obtenu à l'étape 1, et calcul d'une direction de descente.

3. Optimisation le long de la direction de descente.

4. (*optionnel*) Réalisation de nouvelles expériences autour du point optimal et estimation des coefficients d'un modèle d'ordre deux $Y = \alpha + \langle \beta, x \rangle + \langle Hx, x \rangle + \varepsilon$ (avec $\alpha \in \mathbb{R}$, $\beta \in \mathbb{H}$ et $H : \mathbb{H} \to \mathbb{H}$ un opérateur auto-adjoint) pour affiner la localisation du minimum.

Nous testons l'algorithme sur deux exemples de données simulées. Les résultats sont encourageants et mettent en évidence l'importance d'un choix approprié de la base $(\varphi_j)_{j \geq 1}$. Des bases générées à partir d'un échantillon d'apprentissage sont considérées. La base PLS semble la mieux adaptée à notre contexte puisqu'elle permet de prendre en compte la corrélation entre X et Y et par conséquent donne de très bons résultats. La question du choix de la dimension d est également étudiée d'un point de vue pratique : nous constatons que l'algorithme fonctionne d'autant mieux que la dimension d est grande. Cela est dû notamment à une meilleure exploration de l'espace $\mathbb{H}$. La contrainte sur la dimension est donc plutôt liée au nombre d'expériences qu'il est possible de réaliser.

Nous appliquons ensuite la méthode de génération d'un plan d'expériences à des données transmises par le CEA Cadarache, issues de résultats d'un code de calcul. Ces données sont composées de courbes de température, de pression et d'effusivité thermique observées lors d'une simulation numérique d'accident de perte de réfrigérant primaire.

Bibliographie

[1] Bates, R. A., Buck, R. J., Riccomagno, E., et Wynn, H. P. (1996). *J. Roy. Statist. Soc. Ser. B,* **58** (1), 77–94, 95–111. With discussion and a reply by the authors.
[2] Box, G. E. P. et Wilson, K. B. (1951). *J. Roy. Statist. Soc. Ser. B.* **13**, 1–38; discussion: 38–45.
[3] Delaigle, A. et Hall, P. (2012). *Ann. Statist.* **40** (1), 322–352.
[4] Facer, M. R. et Müller, H.-G. (2003). *J. Multivariate Anal.* **87** (1), 191–217.
[5] Georgiou, S. D., Stylianou, S., et Aggarwal, M. (2014). *Comput. Statist. Data Anal.* **71**, 1124–1133.
[6] Glaser, N. S., Barnett, P., McCaslin, I., Nelson, D., Trainor, J., Louie, J., Kaufman, F., Quayle, K., Roback, M., Malley, R., *et al.* (2001). *New Engl. J. Med.* **344** (4), 264–269.
[7] Glaser, N. S., Wootton-Gorges, S. L., Buonocore, M. H., Tancredi, D. J., Marcin, J. P., Caltagirone, R., Lee, Y., Murphy, C., et Kuppermann, N. (2013). *Pediatrics,* **131** (1), e73–e80.
[8] Khuri, A. I. (2001). In: *Proceedings of the Third World Congress of Nonlinear Analysts, Part 3 (Catania, 2000)* volume 47 pp. 2023–2034,.
[9] Khuri, A. I. et Mukhopadhyay, S. (2010). *Wiley Interdiscip. Rev. Comput. Stat.* **2** (2), 128–149.
[10] Lee, J. et Hajela, P. (1996). *J. Aircraft,* **33** (5), 962–969.
[11] Preda, C. et Saporta, G. (2005). *Comput. Statist. Data Anal.* **48** (1), 149–158.
[12] Sacks, J., Welch, W. J., Mitchell, T. J., et Wynn, H. P. (1989). *Statist. Sci.* **4** (4), 409–435. With comments and a rejoinder by the authors.
[13] Wold, H. (1975). In: *Perspectives in probability and statistics (papers in honour of M. S. Bartlett on the occasion of his 65th birthday)* pp. 117–142. Applied Probability Trust Univ. Sheffield, Sheffield.