Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Méthode de l'enclos pour l'estimation non paramétrique sur petits domaines

par Jiming Jiang, Thuan Nguyen et J. Sunil Rao

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 3-12
Statistique Canada, N° 12-001-X au catalogue

Méthode de l'enclos pour l'estimation non paramétrique sur petits domaines

Jiming Jiang, Thuan Nguyen et J. Sunil Rao [1]

Résumé

Nous étudions le problème de la sélection de modèles non paramétriques pour l'estimation sur petits domaines, auquel beaucoup d'attention a été accordée récemment. Nous élaborons une méthode fondée sur le concept de la méthode de l'enclos (*fence method*) de Jiang, Rao, Gu et Nguyen (2008) pour sélectionner la fonction moyenne pour les petits domaines parmi une classe de splines d'approximation. Les études par simulations montrent que la nouvelle méthode donne des résultats impressionnants, même si le nombre de petits domaines est assez faible. Nous appliquons la méthode à un ensemble de données hospitalières sur les échecs de greffe pour choisir un modèle non paramétrique de type Fay-Herriot.

Mots clés : Modèle de Fay-Herriot ; méthode de l'enclos (*fence method*) ; sélection de modèles non paramétriques ; spline pénalisée ; estimation sur petits domaines.

1. Introduction

Ces derniers temps, l'estimation sur petits domaines (EPD) fait couler de plus en plus d'encre. Ici, le terme « petit domaine » désigne une population pour laquelle on ne peut produire des statistiques d'intérêt fiables à cause de certaines limites des données disponibles. Une région géographique (par exemple un État, un comté, une municipalité), un groupe démographique (par exemple un groupe particulier âge × sexe × race) ou un groupe démographique à l'intérieur d'une région géographique sont des exemples de petit domaine. En l'absence d'échantillons directs appropriés pour les petits domaines, des méthodes ont été élaborées afin d'« emprunter de l'information ». Les modèles statistiques, surtout les modèles à effets mixtes, ont joué un rôle important dans l'EPD. Voir Rao (2003) pour un compte rendu complet des diverses méthodes appliquées pour l'EPD.

Alors qu'il existe une abondante littérature sur l'inférence au sujet de petits domaines en utilisant des modèles à effets mixtes, y compris l'estimation des moyennes de petit domaine qui est un problème de prédiction par modèle mixte, l'estimation de l'erreur quadratique moyenne (EQM) du meilleur prédicteur linéaire sans biais empirique (EBLUP pour *empirical best linear unbiased predictor* ; voir Rao 2003) et les intervalles de prédiction (par exemple, Chatterjee, Lahiri et Li 2007), beaucoup moins d'attention a été accordée à la sélection du modèle dans le contexte de l'EPD. Pourtant, des chercheurs renommés spécialisés dans ce domaine (par exemple, Battese, Harter et Fuller 1988, Ghosh et Rao 1994) ont souligné l'importance du choix du modèle dans l'EPD. Datta et Lahiri (2001) discutent d'une méthode de sélection de modèles fondée sur le calcul du facteur de Bayes des fréquentistes pour faire le choix entre un modèle à effets fixes et un modèle à effets aléatoires. Par souci de simplicité, ils se concentrent sur le modèle à effets aléatoires équilibré unidimensionnel suivant : $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, ..., m$, $j = 1, ..., k$, où les u_i et les e_{ij} suivent une loi normale de moyenne nulle et de variance σ_u^2 et σ_e^2, respectivement. Comme le soulignent les auteurs, le choix entre un modèle à effets fixes et un modèle à effets aléatoires équivaut, dans ce cas, à tester l'hypothèse unilatérale $H_0 : \sigma_u^2 = 0$ contre $H_1 : \sigma_u^2 > 0$. Toutefois, notons que les problèmes de sélection de modèles ne peuvent pas tous être exprimés sous forme d'un test d'hypothèse. Fabrizi et Lahiri (2004) ont élaboré une méthode robuste de sélection de modèles dans le contexte des enquêtes complexes. Meza et Lahiri (2005) ont démontré les limites de la statistique C_p de Mallows en sélectionnant les covariables fixes dans un modèle de régression à erreurs emboîtées (Battese, Harter et Fuller 1988) défini par $y_{ij} = x_{ij}'\beta + u_i + e_{ij}$, $i = 1, ..., m$, $j = 1, ..., n_i$, où y_{ij} est l'observation, x_{ij} est un vecteur de covariables fixes, β est un vecteur de coefficients de régression inconnus, et les u_i et les e_{ij} sont les mêmes que dans le modèle mentionné plus haut considéré par Datta et Lahiri (2001). Des études par simulation réalisées par Meza et Lahiri (2005) ont montré que la méthode de la statistique C_p sans modification ne marche pas bien dans les conditions courantes de modèles mixtes quand la variance σ_u^2 est grande ; par ailleurs, un critère C_p modifié élaboré par ces deux derniers auteurs en ajustant les corrélations intra-grappe donne d'aussi bons résultats que le critère C_p dans les conditions de régression. Il convient de souligner que toutes ces études sont limitées aux modèles mixtes linéaires, tandis que la sélection du modèle pour l'EPD dans un contexte de modèles linéaires mixtes généralisés (MLMG) n'a jamais été abordée sérieusement.

1. Jiming Jiang, University of California, Davis. Courriel : jiang@wald.ucdavis.edu ; Thuan Nguyen, Oregon Health and Science University ; J. Sunil Rao, Case Western Reserve University.

Récemment, Jiang et coll. (2008) ont élaboré une nouvelle stratégie de sélection de modèle, qu'ils ont appelée *fence methods*, c'est-à-dire méthodes de l'enclos. Ces auteurs ont constaté que les stratégies classiques de sélection de modèles présentent certaines limites quand elles sont appliquées à des modèles mixtes. Par exemple, la méthode du BIC (Schwarz 1978) s'appuie sur la taille effective d'échantillon dont on n'est pas certain dans des situations typiques d'EPD. Pour illustrer ce point, considérons le modèle de régression à erreurs emboîtées présenté plus haut. Manifestement, la taille effective d'échantillon n'est pas le nombre total d'observations $n = \sum_{i=1}^{m} n_i$, et elle n'est pas proportionnelle non plus à m, le nombre de petits domaines, à moins que tous les n_i soient égaux et fixes. Les méthodes de l'enclos permettent d'éviter ces limites et, par conséquent, conviennent pour la résolution des problèmes de sélection des modèles mixtes, y compris les modèles linéaires mixtes et les MLMG. L'idée fondamentale qui sous-tend ces méthodes est de construire une clôture statistique pour isoler un sous-groupe de ce que l'on sait être les modèles corrects. Une fois que la clôture est construite, le modèle optimal est sélectionné parmi ceux se trouvant dans l'enclos en se servant d'un critère qui peut intégrer des quantités d'un intérêt pratiques. Nous donnons ci-après des renseignements plus détaillés sur les méthodes de l'enclos.

Le présent article est axé sur les modèles non paramétriques pour l'estimation sur petits domaines (EPD). Récemment, beaucoup d'attention a été accordée à ces modèles. En particulier, Opsomer, Breidt, Claeskens, Kauermann et Ranalli (2007) ont proposé un modèle non paramétrique fondé sur des splines pour l'EPD. Leur idée consiste à approximer une fonction moyenne de petit domaine non paramétrique inconnue par une spline pénalisée (P-spline). Ensuite, les auteurs utilisent un lien entre les P-splines et les modèles mixtes linéaires (Wand 2003) pour formuler le modèle d'approximation sous forme d'un modèle linéaire mixte, où les coefficients des splines sont traités comme des effets aléatoires. Pour simplifier, considérons le cas d'une covariable univariée. Nous pouvons alors exprimer une P-spline sous la forme

$$\tilde{f}(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \ldots + \gamma_q (x - \kappa_q)_+^p, \qquad (1)$$

où p est le degré de la spline, q est le nombre de nœuds, $\kappa_j, 1 \le j \le q$ sont les nœuds et $x_+ = x1_{(x>0)}$. Clairement, une P-spline est caractérisée par p, q, ainsi que l'emplacement des nœuds. Notons toutefois que, sachant p et q, l'emplacement des nœuds peut être choisi par l'algorithme de remplissage d'espace implémenté dans R [*cover.design*()]. La question de savoir comment choisir p et q persiste. La « règle empirique » générale est que p est habituellement compris entre 1 et 3, et que q est proportionnel à la taille d'échantillon, n, avec 4 ou 5 observations par nœud (Ruppert, Wand et Carroll 2003). Toutefois, étant donné la règle empirique, il pourrait encore demeurer un grand nombre de choix. Par exemple, si $n = 200$, les choix possibles pour q varient de 40 à 50, qui, combinés avec l'intervalle de 1 à 3 pour p, donnent un total de 33 choix pour la P-spline. Notre nouvelle méthode de l'enclos adaptatif offre une approche dictée par les données pour choisir p et q pour le modèle d'EPD fondé sur des splines.

La présentation de la suite de l'exposé est la suivante. À la section 2, nous décrivons les méthodes de l'enclos. À la section 3, nous élaborons une méthode adaptative de l'enclos pour résoudre le problème de sélection de modèles non paramétriques. À la section 4, nous démontrons les propriétés en échantillon fini de la nouvelle méthode au moyen d'une série d'études par simulation. À la section 5, nous donnons un exemple fondé sur des données réelles comportant l'ajustement d'un modèle de Fay-Herriot (Fay et Herriot 1979) à un ensemble de données issues d'une enquête médicale. Certains résultats techniques sont présentés en annexe.

2. Méthodes de l'enclos

Comme nous l'avons mentionné, le concept fondamental consiste à construire une clôture statistique, puis à sélectionner un modèle optimal parmi ceux se trouvant à l'intérieur de l'enclos en se basant sur un critère d'optimalité, tel que la simplicité du modèle. Soit $Q_M = Q_M(y, \theta_M)$ une mesure du manque d'ajustement, où y représente le vecteur des observations, M désigne un modèle candidat et θ_M désigne le vecteur de paramètres sous le modèle M. Ici, par manque d'ajustement, nous entendons que Q_M satisfait l'exigence de base que $\mathrm{E}(Q_M)$ est minimisée quand M est un modèle vrai et que θ_M est le vecteur de paramètres réels sous M. Alors, un modèle candidat M se trouve dans l'enclos si

$$\hat{Q}_M \le \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}, \qquad (2)$$

où $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M$, Θ_M étant l'espace des paramètres sous M, $\tilde{M}$ est un modèle qui minimise $\hat{Q}_M$ parmi $M \in \mathcal{M}$, l'ensemble des modèles candidats, et $\hat{\sigma}_{M,\tilde{M}}$ est une estimation de l'écart-type de $\hat{Q}_M - \hat{Q}_{\tilde{M}}$. La constante c_n dans le deuxième membre de (2) peut être choisie comme un nombre fixe (par exemple, $c_n = 1$) ou adaptativement (voir plus loin).

Le calcul de $\hat{Q}_M$ est habituellement simple. Par exemple, dans de nombreux cas, Q_M peut être choisi comme la log-vraisemblance négative, ou comme la somme des carrés des résidus. Par ailleurs, le calcul de $\hat{\sigma}_{M,\tilde{M}}$ peut être assez

difficile. Parfois, même si l'on peut obtenir une expression pour $\hat{\sigma}_{M,\tilde{M}}$, on ne peut garantir son exactitude en tant qu'estimation de l'écart-type dans une situation d'échantillon fini. Jiang, Nguyen et Rao (2009) ont simplifié une méthode de l'enclos adaptative proposée par Jiang et coll. (2008). Pour des raisons de simplicité, nous supposons que $\mathcal{M}$ contient un modèle complet, M_{f}, dont chaque modèle candidat est un sous-modèle. Il s'ensuit que $\tilde{M} = M_{\mathrm{f}}$. Dans la méthode adaptative simplifiée, l'inégalité de l'enclos (2) est remplacée par

$$\hat{Q}_M - \hat{Q}_{M_{\mathrm{f}}} \leq c_n, \qquad (3)$$

où la constante c_n est choisie adaptativement comme il suit. Pour chaque $M \in \mathcal{M}$, soit $p^*(M) = \mathrm{P}^*\{M_0(c) = M\}$ la probabilité empirique de sélection pour M, où $M_0(c)$ désigne le modèle sélectionné par la méthode de l'enclos basée sur (3) avec $c_n = c$ et P^* est obtenu par la méthode du bootstrap sous M_{f}. Par exemple, sous un modèle paramétrique, on peut estimer les paramètres du modèle sous M_{f}, puis utiliser un bootstrap paramétrique pour tirer des échantillons sous M_{f}. Supposons que B échantillons sont tirés ; alors, $p^*(M)$ est simplement la proportion d'échantillons (sur un total de B échantillons) dans lesquels M est sélectionné par la méthode de l'enclos basée sur (3) avec la constante c_n donnée. Soit $p^* = \max_{M \in \mathcal{M}} p^*(M)$. Notons que p^* dépend de c_n. Soit c_n^* la constante c_n qui maximise p^* et qui est celle que nous choisissons. Jiang et coll. (2008) offre l'explication suivante pour justifier la méthode de l'enclos adaptatif. Supposons qu'il existe un modèle vrai parmi les modèles candidats ; le modèle optimal est alors celui à partir duquel les données sont générées et devrait être le plus probable, sachant les données. Donc, sachant c_n, on recherche le modèle (en utilisant la méthode de l'enclos) qui est le plus appuyé par les données ou, en d'autres termes, celui dont la probabilité (a posteriori) est la plus élevée. Cette dernière est estimée par la méthode du bootstrap. Notons que, bien que les échantillons bootstrap soient générés sous M_{f}, ils sont presque identiques à ceux générés sous le modèle optimal. Il en est ainsi parce que les estimations qui correspondent aux paramètres nuls doivent, en principe, être proches de zéro, à condition que les estimateurs des paramètres sous M_{f} soient convergents. On dégage alors la constante c_n qui maximise la probabilité (a posteriori) et elle représente le choix optimal.

Il existe deux cas extrêmes correspondant à $c_n = 0$ et à $c_n = \infty$ (c'est-à-dire très grand). Notons que, si $c_n = 0$, $p^* = 1$, parce que quand $c_n = 0$, la méthode aboutit toujours au choix de M_{f}. De même, s'il existe un modèle le plus simple unique (par exemple un modèle avec une dimension minimale), disons, M_*, alors $p^* = 1$ pour la constante c_n très grande. En effet, quand c_n est suffisamment grande, tous les modèles se trouvent dans l'enclos, donc la méthode aboutit toujours au choix de M_* si la simplicité est utilisée comme critère d'optimalité pour sélectionner le modèle à l'intérieur de l'enclos. Ces deux cas extrêmes sont traités attentivement dans Jiang et coll. (2008) et dans Jiang et coll. (2009). Toutefois, comme il est souligné dans Jiang et coll. (2008), les méthodes pour traiter les cas extrêmes, à savoir les tests de dépistage et l'ajustement par rapport à la ligne de base/la vérification des seuils sont rarement nécessaires en pratique. Par exemple, dans la plupart des applications, il existe un (grand) nombre de variables candidates, mais l'on pense qu'un (petit) sous-ensemble seulement de ces variables est important. Cela signifie que le modèle optimal n'est ni M_* ni M_{f}. Donc, il n'est pas nécessaire de s'inquiéter des cas extrêmes et les méthodes pour le traitement de ces cas peuvent être omises. Dans la plupart des applications, le graphique de p^* en fonction de c_n a une forme en W dont le pic du milieu correspond à c_n^*.

Le graphique de droite de la figure 2 en est un exemple. Il s'agit du graphique de p^* en fonction de c_n pour l'exemple discuté à la section 5. Ce graphique montre la forme en « W » typique, telle que nous l'avons décrite, et le pic du milieu correspond à l'endroit où se trouve la valeur optimale de c_n, c'est-à-dire c_n^*.

Jiang et coll. (2009) ont établi la cohérence de l'enclos adaptatif simplifié et étudié ses propriétés en échantillon fini.

3. Sélection de modèles d'EPD non paramétriques

Afin de simplifier l'illustration, nous considérons le modèle d'EPD suivant :

$$y_i = f(X_i) + B_i u_i + e_i, \quad i = 1, \ldots, m, \qquad (4)$$

où y_i est un vecteur de dimension $n_i \times 1$ représentant les observations provenant du i^{e} petit domaine, $f(X_i) = [f(x_{ij})]_{1 \leq j \leq n_i}$ avec $f(x)$ une fonction (lisse) inconnue, B_i est une matrice connue de dimensions $n_i \times b$, u_i est un vecteur de dimension $b \times 1$ d'effets aléatoires propres au petit domaine, et e_i est un vecteur de dimension $n_i \times 1$ d'erreurs d'échantillonnage. Nous supposons que les u_i, e_i, $i = 1, \ldots, m$ sont indépendants avec $u_i \sim N(0, G_i)$, $G_i = G_i(\theta)$, et $e_i \sim N(0, R_i)$, $R_i = R_i(\theta)$, θ étant un vecteur inconnu de composantes de variance. Notons que, à part $f(X_i)$, le modèle est le même que le modèle mixte linéaire « longitudinal » classique (par exemple Laird et Ware 1982, Datta et Lahiri 2000).

Le modèle de spline d'approximation est donné en remplaçant $f(x)$ par $\tilde{f}(x)$ dans (1), où les coefficients β et γ sont estimés par les moindres carrés pénalisés, c'est-à-dire par

$$\text{minimisation de } |y - X\beta - Z\gamma|^2 + \lambda |\gamma|^2, \qquad (5)$$

où $y = (y_i)_{1 \le i \le m}$, la $(i, j)^{\text{e}}$ ligne de X est $(1, x_{ij}, ..., x_{ij}^p)$, la $(i, j)^{\text{e}}$ ligne de Z est $[(x_{ij} - \kappa_1)_+^p, ..., (x_{ij} - \kappa_q)_+^p]$, $i = 1, ..., m$, $j = 1, ..., n_i$ et λ est un paramètre de pénalité, ou de lissage. Pour déterminer λ, Wand (2003) a utilisé le lien intéressant suivant avec un modèle mixte linéaire. Pour illustrer l'idée, considérons un cas simple dans lequel $B_i = 0$ (c'est-à-dire qu'il n'existe pas d'effets aléatoires de petit domaine) et les composantes de e_i sont indépendantes et suivent une loi $N(0, \tau^2)$. Si les γ sont traités comme des effets aléatoires qui sont indépendants et suivent une loi $N(0, \sigma^2)$, la solution de (5) est la même que le meilleur estimateur linéaire sans biais (BLUE pour *best linear unbiased estimator*) pour β, et que le meilleur prédicteur linéaire sans biais (BLUP pour *best linear unbiased predictor*) pour γ, si λ est identique au ratio τ^2/σ^2. Donc, la valeur de λ peut être estimée par les estimateurs du maximum de vraisemblance (MV) ou du maximum de vraisemblance restreint (MVR) de σ^2 et τ^2 (par exemple, Jiang 2007). Cependant, certaines études donnent à penser que cette approche produit un biais de sous-lissage (Kauermann 2005). Considérons, par exemple, le cas particulier dans lequel $f(x)$ est, en fait, la spline quadratique avec deux nœuds donnés par (10). (Notons que cette fonction est lisse en ce sens qu'elle possède une dérivée continue.) Il est évident que, dans ce cas, la meilleure spline d'approximation devrait être la fonction $f(x)$ proprement dite avec seulement deux nœuds, c'est-à-dire $q = 2$ (naturellement, nous pourrions utiliser une spline comportant de nombreux nœuds pour « approximer » la spline quadratique à deux nœuds, mais cela paraît très inefficace dans le cas qui nous occupe). Toutefois, si nous utilisons le lien avec un modèle linéaire mixte mentionné plus haut, l'estimateur MV (ou MVR) de σ^2 est convergent uniquement si $q \to \infty$ (c'est-à-dire si le nombre d'apparitions des effets aléatoires de la spline tend vers l'infini). L'incohérence apparente a deux conséquences inquiétantes : i) la signification de λ pourrait être conceptuellement difficile à interpréter et ii) le comportement de l'estimateur de λ pourrait être imprévisible.

La méthode de l'enclos offre une approche naturelle pour choisir le degré de la spline, p, le nombre de nœuds, q, et le paramètre de lissage, λ, simultanément. Notons toutefois une différence importante par rapport aux situations considérées dans Jiang et coll. (2008) et dans Jiang et coll. (2009) en ce sens que le vrai modèle sous-jacent ne fait pas partie de la classe des modèles candidats, c'est-à-dire les splines d'approximation (1). De surcroît, le rôle de λ dans le modèle devrait être clarifié : λ contrôle le degré de lissage du modèle sous-jacent. Une mesure naturelle du manque d'ajustement est $Q_M = |y - X\beta - Z\gamma|^2$. Cependant, $\hat{Q}_M$ ne s'obtient pas par minimisation de Q_M sur β et γ sans contrainte. Nous avons plutôt $\hat{Q}_M = |y - X\hat{\beta} - Z\hat{\gamma}|^2$, où $\hat{\beta}$ et $\hat{\gamma}$ sont la solution de (5) et dépendent donc de λ. Le paramètre λ optimal doit être sélectionné par la méthode de l'enclos, en même temps que p et q, comme il est décrit plus bas.

Une autre différence tient au fait qu'il pourrait ne pas y avoir de modèle complet parmi les modèles candidats. Par conséquent, nous remplaçons l'inégalité d'enclos (3) par la suivante :

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \le c_n, \tag{6}$$

où $\tilde{M}$ est le modèle candidat possédant le $\hat{Q}_M$ minimal. Nous utilisons le critère d'optimalité à l'intérieur de l'enclos suivant, qui combine la simplicité et le degré de lissage du modèle. Pour les modèles compris dans l'enclos, choisir celui dont la valeur de q est la plus petite ; s'il existe plus d'un modèle répondant à ce critère, choisir celui dont la valeur de p est la plus petite. Cela donne le meilleur choix de p et q. Une fois que p et q sont choisis, nous choisissons le modèle *à l'intérieur de l'enclos* dont la valeur de λ est la plus grande. De nouveau, notons que λ fait partie du modèle M qui est sélectionné (ou « estimé ») par la méthode de l'enclos. La constante de réglage c_n est choisie adaptativement en utilisant la méthode adaptative simplifiée de Jiang et coll. (2009), où le bootstrap paramétrique est employé pour calculer p^* (voir la section 2).

La preuve du théorème qui suit est donnée en annexe. Par souci de simplicité, supposons que la matrice $W = (X\ Z)$ est de plein rang. Soit $P_{W^\perp} = I_n - P_W$, où $n = \sum_{i=1}^m n_i$ et $P_W = W(W'W)^{-1}W'$.

Théorème. En ce qui concerne les calculs, la méthode de l'enclos susmentionnée est équivalente à la procédure suivante : i) d'abord utiliser l'enclos (adaptatif) pour choisir p et q en utilisant (6) avec $\lambda = 0$ et $\hat{Q}_M = y' P_{W^\perp} y$ (voir le lemme plus bas), ainsi que le critère mentionné plus haut pour choisir p, q dans l'enclos ; ii) en désignant par M_0^* le modèle correspondant aux paramètres p et q choisis, trouver le λ maximal tel que

$$\hat{Q}_{M_0^*, \lambda} - \hat{Q}_{\tilde{M}} \le c_n^*, \tag{7}$$

où, pour tout modèle M avec les X et Z correspondants, nous avons

$$\begin{aligned}
\hat{Q}_{M,\lambda} &= |y - X\hat{\beta}_\lambda - Z\hat{\gamma}_\lambda|^2, \\
\hat{\beta}_\lambda &= (X'V_\lambda^{-1}X)^{-1}X'V_\lambda^{-1}y, \\
\hat{\gamma}_\lambda &= \lambda^{-1}(I_q + \lambda^{-1}Z'Z)^{-1}Z'(y - X\hat{\beta}_\lambda), \\
X'V_\lambda^{-1}X &= X'X - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'X, \\
X'V_\lambda^{-1}y &= X'y - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'y,
\end{aligned}$$

et la constante c_n^* est choisie par la méthode de l'enclos adaptatif décrite à la section 2 (V_λ est définie ci-après, mais

n'est pas directement nécessaire ici pour le calcul, à cause des deux dernières équations).

Notons qu'à l'étape (i) du théorème, il n'est pas nécessaire de s'occuper de λ. L'élément qui motive (7) est que cette inégalité est satisfaite quand $\lambda = 0$, de sorte que l'on voudrait savoir jusqu'à quel point λ peut aller. En fait, le λ maximal est une solution de l'équation $\hat{Q}_{M_0^*,\lambda} - \hat{Q}_{\tilde{M}} = c_n^*$. Le but des deux dernières équations est d'éviter l'inversion directe de $V_\lambda = I_n + \lambda^{-1} ZZ'$, dont la dimension est égale à n, c'est-à-dire la taille totale de l'échantillon. Notons que V_λ ne possède pas une structure diagonale par bloc à cause de ZZ', de sorte que, si n est grand, l'inversion directe de V_λ peut donner lieu à des calculs difficiles.

La preuve du théorème requiert le lemme qui suit, dont la preuve est donnée en annexe.

Lemme. Pour tout M et y, $\hat{Q}_{M,\lambda}$ est une fonction croissante de λ avec $\inf_{\lambda>0} \hat{Q}_{M,\lambda} = \hat{Q}_M$.

4. Simulations

Nous considérons une extension du modèle de Fay-Herriot (Fay et Herriot 1979) dans des conditions non paramétriques. Le modèle peut s'écrire sous la forme

$$y_i = f(x_i) + v_i + e_i,\ i = 1,\ ...,\ m, \qquad (8)$$

où les v_i, e_i, $i = 1, ..., m$ sont indépendants et tels que $v_i \sim N(0,\ A)$, $e_i \sim N(0,\ D_i)$, où A est inconnue, mais la variance d'échantillonnage D_i est supposée connue. La principale différence par rapport au modèle de Fay-Herriot classique est $f(x_i)$, où $f(x)$ est une fonction lisse inconnue.

Pour simplifier, nous supposons que $D_i = D, 1 \le i \le m$. Alors, le modèle peut être exprimé sous la forme

$$y_i = f(x_i) + \varepsilon_i,\ i = 1,\ ...,\ m, \qquad (9)$$

où $\varepsilon_i \sim N(0,\ \sigma^2)$ avec $\sigma^2 = A + D$, qui est inconnue. Donc, le modèle est le même que le modèle de régression non paramétrique.

Nous considérons trois cas distincts qui couvrent divers aspects et situations. Dans le premier cas (cas 1), la fonction sous-jacente réelle est une fonction linéaire, $f(x) = 1 - x$, $0 \le x \le 1$, d'où le modèle se réduit au modèle de Fay-Herriot classique. L'objectif est de découvrir si la méthode de l'enclos permet de valider ce modèle quand il est valide. Dans le second cas (cas 2), la fonction sous-jacente réelle est une spline quadratique à deux nœuds donnée par

$$f(x) = 1 - x + x^2 - 2(x-1)_+^2 + 2(x-2)_+^2,\ 0 \le x \le 3 \qquad (10)$$

(la forme est un demi-cercle entre 0 et 1 ouvert vers le haut, un demi-cercle entre 1 et 2 ouvert vers le bas et un demi-cercle entre 2 et 3 ouvert vers le haut). Notons que cette fonction est lisse en ce sens qu'elle possède une dérivée continue. Ici, nous avons l'intention de déterminer si l'enclos permet d'identifier la fonction sous-jacente réelle dans la situation « parfaite », c'est-à-dire quand la fonction $f(x)$ proprement dite est une spline. Le dernier cas (cas 3), est peut-être celui représentant la situation la plus proche de la pratique, dans laquelle aucune spline ne peut fournir une approximation parfaite de $f(x)$. Autrement dit, la fonction sous-jacente réelle ne figure pas parmi les candidates. Dans ce cas, $f(x)$ est choisie telle que $0,5\sin(2\pi x)$, $0 \le x \le 1$, qui est l'une des fonctions considérées par Kauermann (2005).

Nous étudions les situations où la taille d'échantillon est petite ou moyenne, à savoir $m = 10$, 15 ou 20 pour le cas 1, $m = 30$, 40 ou 50 pour le cas 2, et $m =$ 10, 30 ou 50 pour le cas 3. La covariable x_i est tirée de la loi uniforme [0, 1] dans le cas 1, et de la loi uniforme [0, 3] dans le cas 2, puis maintenue fixe dans toutes les simulations. À l'exemple de Kauermann (2005), nous posons que x_i représente les points équidistants dans le cas 3. Nous choisissons pour l'écart-type de l'erreur σ dans (9) la valeur de 0,2 dans les cas 1 et 2. Cette valeur est telle que l'écart-type du signal est, dans chaque cas, à peu près le même que l'écart-type de l'erreur. Pour le cas 3, nous examinons trois valeurs différentes pour σ, à savoir 0,2, 0,5 et 1,0. Ces valeurs sont également du même ordre que l'écart-type du signal dans ce cas.

Les splines d'approximation candidates pour les cas 1 et 2 sont les suivantes : $p = 0, 1, 2, 3$, $q = 0$ et $p = 1, 2, 3$, $q = 2, 5$ (de sorte qu'il existe, en tout, 10 candidates). Pour le cas 3, comme Kauermann (2005), nous considérons uniquement des splines linéaires (c'est-à-dire $p = 1$) ; en outre, nous prenons en considération le nombre de nœuds dans l'intervalle donné par la « règle empirique » (c'est-à-dire environ 4 ou 5 observations par nœud ; voir la section 1), ainsi que le modèle passant par l'origine ($p = q = 0$) et le modèle linéaire ($p = 1$, $q = 0$). Donc, pour $m = 10$, $q = 0$, 2, 3, pour $m = 30$, $q = 0, 6, 7, 8$ et pour $m = 50$, $q = 0$, 10, 11, 12, 13.

Le tableau 1 donne les résultats fondés sur 100 simulations sous les cas 1 et 2. Comme dans Jiang et coll. (2009), nous considérons à la fois le pic le plus élevé, ce qui consiste à choisir c_n avec p^* le plus élevé, ainsi que la limite inférieure (LI) de confiance à 95 %, ce qui consiste à choisir une valeur plus faible de c_n correspondant à un pic de p^* afin d'être prudent, si le p^* correspondant est supérieur à la limite inférieure de confiance à 95 % de p^* pour toute valeur de c_n plus grande qui correspond à un pic de p^*. Nous voyons que la performance de l'enclos adaptatif est satisfaisante même si la taille de l'échantillon est petite. En outre, il semble que la méthode de la limite de confiance inférieure donne de meilleurs résultats dans le cas de petits échantillons, mais ne produit pratiquement pas de

différence dans les grands échantillons. Ces constatations corroborent celles de Jiang et coll. (2009).

Tableau 1
Sélection de modèles non paramétriques – cas 1 et cas 2 : probabilités empiriques, en pourcentage, que le modèle optimal soit sélectionné fondées sur 100 simulations

	Cas 1			Cas 2		
Taille de l'échantillon	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
Pic le plus élevé	62	91	97	71	83	97
LI de confiance	73	90	97	73	80	96

Le tableau 2 donne les résultats pour le cas 3. Notons que, contrairement aux cas 1 et 2, il n'existe pas ici de modèle optimal (un modèle optimal doit être un modèle vrai conformément à notre définition). Donc, au lieu de donner les probabilités empiriques de sélection du modèle optimal, nous donnons la distribution empirique des modèles sélectionnés dans chaque cas. Nous voyons que, à mesure que σ augmente, la distribution des modèles sélectionnés est de plus en plus étalée. Nous observons un profil inverse à mesure que m augmente. La méthode de la limite inférieure (LI) de confiance semble donner de meilleurs résultats pour le choix d'un modèle de spline. Parmi les modèles de spline, la méthode de l'enclos semble donner de façon écrasante la préférence à un petit plutôt qu'à un grand nombre de nœuds.

Soulignons que la méthode de l'enclos nous permet de choisir non seulement p et q, mais aussi λ (voir la section 3). Dans chaque simulation, nous calculons $\hat{\beta} = \hat{\beta}_\lambda$ et $\hat{\gamma} = \hat{\gamma}_\lambda$, donné sous l'expression (7), fondé sur le λ choisi par la méthode de l'enclos adaptatif. Nous calculons les valeurs ajustées au moyen de (1) en remplaçant β et γ par $\hat{\beta}$ et $\hat{\gamma}$, respectivement. Puis, nous calculons la moyenne des valeurs prédites sur les 100 simulations. La figure 1 donne les valeurs prédites moyennes pour les trois cas $(m = 10, 30, 50)$ avec $\sigma = 0{,}2$ sous le cas 3. Les valeurs réelles de la fonction sous-jacente, $f(x_i) = 0{,}5 \sin(2\pi x_i)$, $i = 1, \ldots, m$ sont également tracées aux fins de comparaison.

5. Un exemple fondé sur des données réelles

Nous nous servons d'un ensemble de données tirées de Morris et Christiansen (1995) provenant de 23 hôpitaux (sur un total de 219) dans lesquels au moins 50 greffes de rein avaient été effectuées au cours d'une période de 27 mois (tableau 3). Les y_i sont les taux d'échecs de greffe pour les opérations de greffe de rein, autrement dit $y_i =$ nombre d'échecs de greffe $/ n_i$, où n_i est le nombre de greffes de rein à l'hôpital i durant la période de référence. La variance du taux d'échecs de greffe, D_i, est approximée par $(0{,}2)(0{,}8)/n_i$, où 0,2 est le taux d'échecs observé pour l'ensemble des hôpitaux. Donc, nous supposons que D_i est connue. En outre, nous disposons pour chaque hôpital d'un indice de gravité x_i, qui est la fraction moyenne de femmes, de noirs, d'enfants ou de personnes gravement malades ayant reçu un rein à l'hôpital i. Nous traitons l'indice de gravité comme une covariable.

Tableau 2
Sélection de modèles non paramétriques – cas 3 : distributions empiriques, en pourcentage, des modèles sélectionnés

	Taille de l'échantillon	$m = 10$		$m = 30$		$m = 50$	
	N^bre de nœuds	0, 2, 3		0, 6, 7, 8		0, 10, 11, 12, 13	
		(p, q)	%	(p, q)	%	(p, q)	%
$\sigma = 0{,}2$	Pic le plus élevé	(0, 0)	1	(1, 0)	9	(1, 10)	100
		(1, 0)	31	(1, 6)	91		
		(1, 2)	68				
	LI de confiance	(1, 0)	24	(1, 0)	9	(1, 10)	100
		(1, 2)	76	(1, 6)	91		
$\sigma = 0{,}5$	Pic le plus élevé	(0, 0)	14	(1, 0)	21	(1, 0)	13
		(1, 0)	27	(1, 6)	77	(1, 10)	84
		(1, 2)	56	(1, 7)	2	(1, 11)	2
		(1, 3)	3			(1, 12)	1
	LI de confiance	(0, 0)	8	(1, 0)	8	(1, 0)	2
		(1, 0)	23	(1, 6)	89	(1, 10)	94
		(1, 2)	65	(1, 7)	3	(1, 11)	2
		(1, 3)	4			(1, 12)	2
$\sigma = 1$	Pic le plus élevé	(0, 0)	27	(0, 0)	15	(0, 0)	10
		(1, 0)	20	(1, 0)	18	(1, 0)	26
		(1, 2)	49	(1, 6)	63	(1, 10)	60
		(1, 3)	4	(1, 7)	4	(1, 11)	2
						(1, 12)	2
	LI de confiance	(0, 0)	20	(0, 0)	1	(0, 0)	2
		(1, 0)	13	(1, 0)	13	(1, 0)	13
		(1, 2)	59	(1, 6)	82	(1, 10)	80
		(1, 3)	8	(1, 7)	4	(1, 11)	2
						(1, 12)	3

Tableau 3
Données hospitalières tirées de Morris et Christiansen (1995)

Région	y_i	x_i	$\sqrt{D_i}$
1	0,302	0,112	0,055
2	0,140	0,206	0,053
3	0,203	0,104	0,052
4	0,333	0,168	0,052
5	0,347	0,337	0,047
6	0,216	0,169	0,046
7	0,156	0,211	0,046
8	0,143	0,195	0,046
9	0,220	0,221	0,044
10	0,205	0,077	0,044
11	0,209	0,195	0,042
12	0,266	0,185	0,041
13	0,240	0,202	0,041
14	0,262	0,108	0,036
15	0,144	0,204	0,036
16	0,116	0,072	0,035
17	0,201	0,142	0,033
18	0,212	0,136	0,032
19	0,189	0,172	0,031
20	0,212	0,202	0,029
21	0,166	0,087	0,029
22	0,173	0,177	0,027
23	0,165	0,072	0,025

Ganesh (2009) a proposé pour les taux d'échecs de greffe le modèle de Fay-Herriot suivant : $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$, où les v_i sont les effets aléatoires spécifiques à l'hôpital et les e_i sont les erreurs d'échantillonnage. Nous supposons que les v_i, e_i sont indépendants avec $v_i \sim N(0, A)$ et $e_i \sim N(0, D_i)$. Ici, la variance A est inconnue. En se fondant sur le modèle, Ganesh a obtenu des intervalles crédibles pour les contrastes choisis. Cependant, l'inspection des données brutes suggère certaines tendances non linéaires, de sorte que la question se pose de savoir si l'on peut donner à la partie effets fixes du modèle une forme fonctionnelle plus souple.

Pour répondre à cette question, nous considérons le modèle de Fay-Herriot comme un membre spécial d'une classe de modèles de spline d'approximation dont nous avons discuté à la section 3. Plus précisément, nous supposons que

$$y_i = f(x_i) + v_i + e_i, \ i = 1, ..., m, \qquad (11)$$

où $f(x)$ est une fonction lisse inconnue et tous les autres termes sont les mêmes que dans le modèle de Fay-Herriot. Nous considérons ensuite la classe suivante de modèles de spline d'approximation :

$$\hat{f}(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + ... + \gamma_q (x - \kappa_q)_+^p \qquad (12)$$

avec $p = 0, 1, 2, 3$ et $q = 0, 1, \ldots, 6$ (le cas $p = 0$ est uniquement pour $q = 0$). Ici la borne supérieure 6 est choisie selon la « règle empirique » (parce que $m = 23$, donc $m/4 = 5{,}75$). Notons que le modèle de Fay-Herriot correspond au cas où $p = 1$ et $q = 0$. La question est alors de trouver le modèle optimal, en ce qui concerne p et q, pour cette classe.

Nous appliquons la méthode de l'enclos adaptatif décrite à la section 3 à ce cas. Ici, pour obtenir les échantillons bootstrap nécessaires pour déterminer c_n^*, nous commençons par calculer l'estimateur MV sous le modèle $\tilde{M}$, qui minimise $\hat{Q}_M = y' P_{W^\perp} y$ parmi les modèles candidats [c'est-à-dire (12) ; voir le théorème à la section 3], puis nous tirons des échantillons bootstrap paramétriques sous le modèle $\tilde{M}$ en traitant les estimateurs MV comme étant les paramètres réels. Cela est raisonnable parce que $\tilde{M}$ est le meilleur modèle d'approximation pour ce qui est de l'ajustement, même si sous le modèle (11), il pourrait ne pas exister de modèle vrai parmi les modèles candidats. Nous choisissons la taille d'échantillon bootstrap égale à 100.

La méthode de l'enclos donne lieu à la sélection du modèle $p = 3$ et $q = 0$, c'est-à-dire une fonction cubique sans nœuds, comme modèle optimal. Pour être certains que la taille d'échantillon bootstrap $B = 100$ est adéquate, nous avons répété l'analyse 100 fois, en utilisant chaque fois des échantillons bootstrap différents (rappelons que dans la méthode de l'enclos adaptatif, il faut tirer des échantillons bootstrap afin de déterminer c_n^*, si bien que la question est de savoir si différents échantillons bootstrap donnent lieu à la sélection de modèles différents). Tous les résultats ont abouti à la sélection du même modèle, à savoir une fonction cubique sans nœuds (même si les quantités intermédiaires dérivées du bootstrap, telles que p^* et c_n^*, variaient d'une itération bootstrap à l'autre). Nous avons également effectué l'analyse des données en utilisant $B = 1\,000$ et le modèle sélectionné est demeuré le même. Donc, il semble que la taille d'échantillon bootstrap de $B = 100$ est adéquate. Le graphique de gauche de la figure 2 donne le tracé de p^* en fonction de c_n dans la sélection du modèle par la méthode de l'enclos adaptatif.

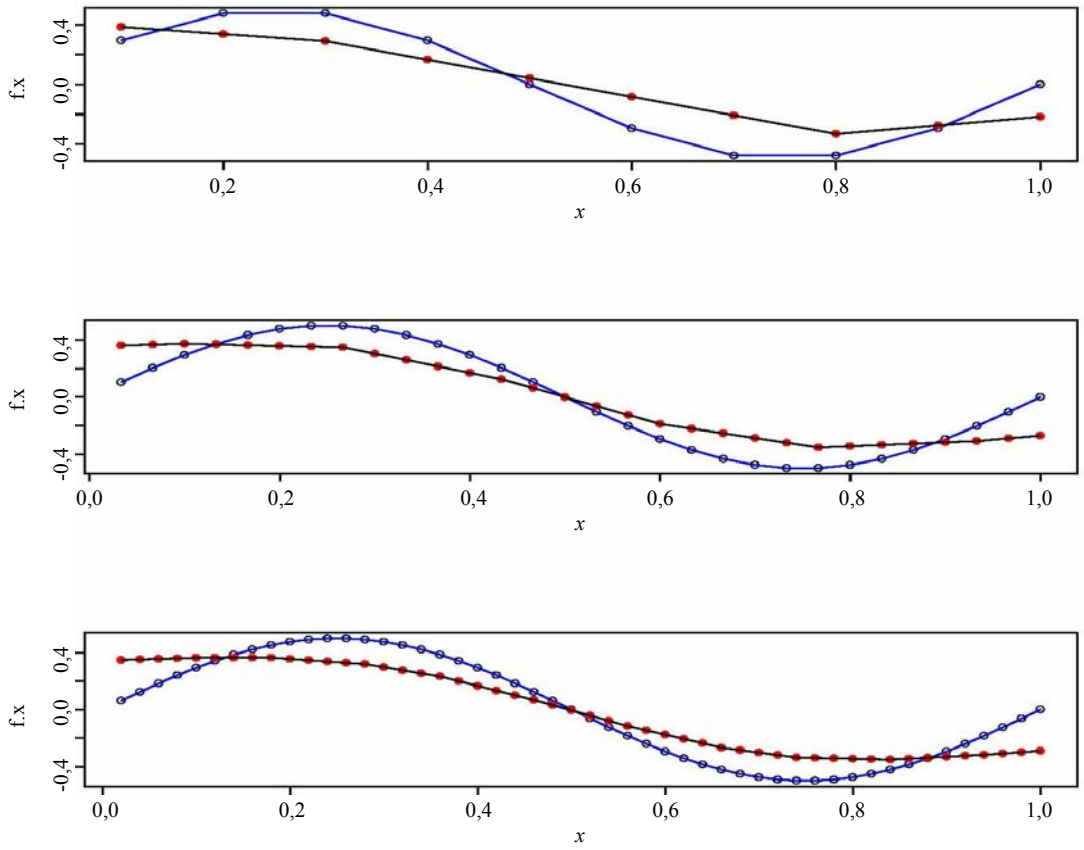

Figure 1 **Simulation du cas 3. Graphique du haut : valeurs prédites moyennes pour $m = 10$. Graphique du milieu : valeurs prédites moyennes pour $m = 30$. Graphique du bas : valeurs prédites moyennes pour $m = 50$. Dans tous les cas, les points représentent les valeurs prédites, tandis que les cercles correspondent à la fonction sous-jacente réelle**

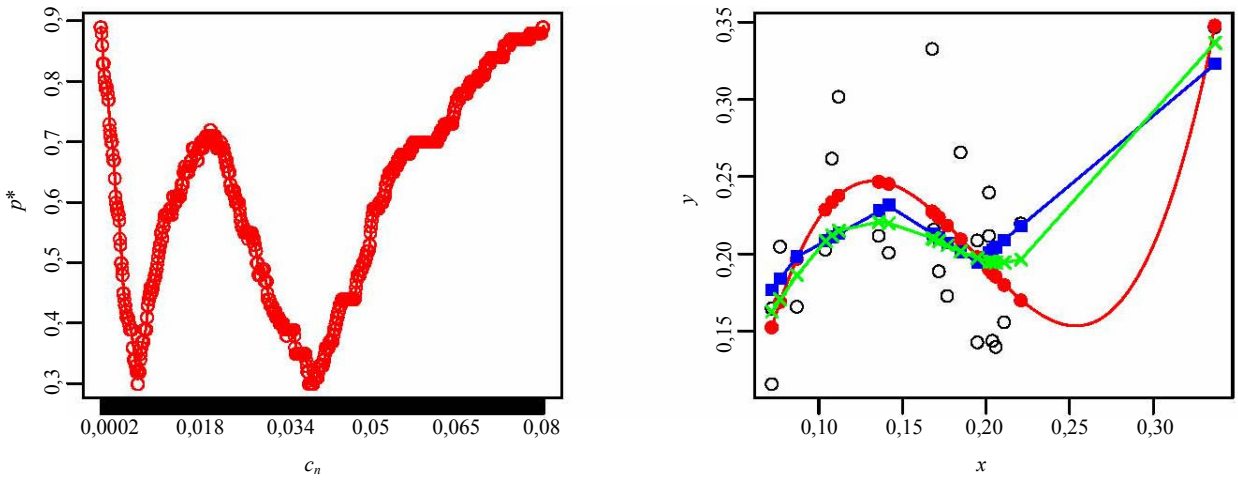

Figure 2 **À gauche : graphique de $p*$ en fonction de c_n pour la recherche sur l'espace complet des modèles. À droite : données brutes et valeurs et courbes prédites ; les points et leur courbe correspondent à la fonction cubique résultant de la recherche dans l'espace complet des modèles ; les carrés et leur courbe correspondent à la spline linéaire avec quatre nœuds résultant de la recherche dans l'espace restreint des modèles ; les X et leur courbe représentent les valeurs prédites par le MAG**

Quelques comparaisons sont toujours utiles. Nous effectuons notre première comparaison avec la méthode de l'enclos proprement dite, mais en utilisant un espace plus restreint de modèles candidats. Plus précisément, nous considérons (12) en imposant la contrainte de splines linéaires uniquement, c'est-à-dire $p = 1,$ et un nombre de nœuds compris dans l'intervalle de la « règle empirique », c'est-à-dire $q = 4, 5, 6,$ plus le modèle passant par l'origine $(p = q = 0)$ et le modèle linéaire $(p = 1,\ q = 0).$ Dans ces conditions, la méthode de l'enclos donne lieu à la sélection d'une spline linéaire à quatre nœuds (c'est-à-dire $p = 1,$ $q = 4)$ comme modèle optimal. La valeur de λ correspondant à ce modèle est approximativement égale à 0,001. Le tracé de p^* en fonction de c_n pour cette sélection de modèle est fort semblable au graphique de gauche de la figure 2 et, par conséquent, est omis. En outre, le graphique de droite de la figure 2 donne les valeurs et les courbes prédites sous les deux modèles sélectionnés par la méthode de l'enclos parmi les différents espaces de modèles, ainsi que les points de données originaux.

Une autre comparaison peut être effectuée en traitant (11) comme un modèle additif généralisé (MAG) avec erreurs hétéroscédastiques. Nous pouvons obtenir un ajustement pondéré avec degré de lissage optimisé en utilisant un critère de validation croisée généralisée (VCG). Ici, les poids utilisés sont $w_i = 1/(A + D_i),$ où l'estimation du maximum de vraisemblance pour A est utilisée comme estimation par remplacement. Rappelons que les D_i sont connus. Cette fonction prédite est également superposée dans le graphique de droite de la figure 2. Soulignons à quel point elle ressemble à celle prédite par la méthode de l'enclos avec espace restreint.

Pour étendre la classe de modèles pris en considération par lissage fondé sur la VCG, nous avons utilisé la procédure BRUTO (Hastie et Tibshirani 1990) qui consiste à augmenter la classe de modèles pour examiner un ajustement nul et un ajustement linéaire pour la fonction spline et qui intègre la sélection de modèles résultante (c'est-à-dire modèles nul, linéaire ou lisse) dans un algorithme de rétroajustement pondéré en utilisant le critère VCG pour l'efficacité des calculs. Fait intéressant, ici, la procédure BRUTO trouve simplement un ajustement linéaire global pour la forme fonctionnelle des effets fixes. Il s'agit certes d'une comparaison intéressante, mais les propriétés théoriques de BRUTO pour des modèles tels que (11) n'ont pas vraiment été étudiées en profondeur.

Enfin, comme nous l'avons mentionné à la section 3, en utilisant le lien entre le modèle mixte P-spline et linéaire, nous pouvons formuler (12) sous la forme d'un modèle linéaire mixte, où les coefficients de la fonction spline sont traités comme des effets aléatoires. Le problème se résume alors à un problème de sélection de modèles mixtes (paramétriques), de sorte que la méthode de Jiang et coll. (2009) peut être appliquée. En fait, il s'agissait de notre approche initiale pour l'ensemble de données utilisé et le modèle que nous avons trouvé était le même que celui sélectionné par la procédure BRUTO. Cependant, nous avons certaines réserves quant à cette approche, comme nous l'avons expliqué à la section 3.

6. Conclusion

Bien que le présent article porte principalement sur la sélection de modèles d'EPD non paramétriques, notre méthode pourrait être applicable à des problèmes de sélection de modèles à effets mixtes fondés sur des splines dans d'autres domaines, comme l'analyse de données longitudinales (par exemple Wang 2005).

Dans le cas où un modèle vrai existe parmi les modèles candidats, tels les cas 1 et 2 à la section 4, il est possible d'établir la cohérence de la méthode de l'enclos proposée pour sélectionner le modèle de la même façon qu'à la section 3 de Jiang et coll. (2009) (quoique le résultat de ce dernier article ne s'applique pas directement). Toutefois, en pratique, la situation dans laquelle la modélisation non paramétrique est la plus utile est celle où un modèle vrai n'existe pas, ou qu'il ne figure pas parmi les candidats, comme dans le cas 3 de la section 4. Dans ces conditions, aucun résultat de cohérence ne peut évidemment être prouvé. Il reste à préciser quel serait un comportement asymptotique désirable pour étudier ce dernier cas.

Remerciements

Les travaux de Jiming Jiang sont financés en partie par les bourses de la NSF DMS-203676 et DMS-0402824. Ceux de J. Sunil Rao sont financés partiellement par les bourses de la NSF DMS-0203724 et DMS-405072, et par la bourse des NIH K25-CA89868.

Annexe

1. *Preuve du lemme.* Écrivons $g(\lambda) = \hat{Q}_{M,\lambda}.$ Nous pouvons montrer (détails omis) que $g'(\lambda) = 2\lambda y' B_\lambda A_\lambda B_\lambda' y,$ où $A_\lambda = B'(W'W + \lambda BB')^{-1} B,$ $B_\lambda = W(W'W + \lambda BB')^{-1} B$ avec $B' = (0\ I_q)$ et $W = (X\ Z).$ Donc, $g'(\lambda) \geq 0$ pour $\lambda > 0.$ En outre, $Q_{M,\lambda} \to \hat{Q}_M$ quand $\lambda \to 0.$

2. *Preuve du théorème.* Considérons l'inégalité d'enclos

$$\hat{Q}_{M,\lambda} - \hat{Q}_{\tilde{M},\tilde{\lambda}} \leq c_n, \qquad \text{(A.1)}$$

où $(\bar{M}, \bar{\lambda})$ minimise $\hat{Q}_{M,\lambda}.$ Considérons aussi l'inégalité d'enclos obtenue en utilisant $\hat{Q}_M = y' P_{W^\perp} y,$ qui est

$$\hat{Q}_M - \hat{Q}_{\bar{M}} \leq c_n. \qquad \text{(A.2)}$$

En vertu du lemme, il faut que $\bar{\lambda} = 0$, et $\bar{M} = \tilde{M}$, donc $\hat{Q}_{\bar{M},\bar{\lambda}} = \hat{Q}_{\tilde{M}}$. Il s'ensuit, de nouveau en vertu du lemme, que pour la même constante c_n, (A.2) est vérifiée si et uniquement si (A.1) est vérifiée pour une valeur donnée de λ. Par conséquent, les modèles compris à l'intérieur de l'enclos, en ce qui concerne p et q, sont les mêmes sous les deux méthodes. Il est alors facile de voir, d'après le critère de sélection, que le même modèle $M_0 = M_0(c_n)$, en ce qui concerne p et q, sera sélectionné sous les deux méthodes pour la constante donnée c_n. Il s'ensuit alors que la constante c_n^* sélectionnée en utilisant la méthode adaptative sera la même sous les deux méthodes. Donc, en utilisant de nouveau l'argument susmentionné, le modèle optimal M_0^*, en ce qui concerne p et q, sera le même sous les deux méthodes.

Les formules sous l'expression (7) peuvent être calculées en utilisant les expressions du BLUE et du BLUP (par exemple Jiang 2007, paragraphe 2.3.1) et l'égalité suivante (par exemple, Sen et Srivastava 1990, page 275) : si U est de dimensions $n \times q$ et V est de dimensions $q \times n$, alors $(P + UV)^{-1} = P^{-1} - P^{-1}U(I_q + VP^{-1}U)^{-1}VP^{-1}$ à condition que les inverses existent.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.

Chatterjee, S., Lahiri, P. et Li, H. (2007). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, to appear.

Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, G.S., et Lahiri, P. (2001). Discussions on a paper by Efron & Gous. (Éd., P. Lahiri) *Model Selection*, IMS Lecture Notes/Monograph 38.

Fabrizi, E., et Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Technical Report, Dept. of Math., Univ. of Maryland.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *Journal of Multivariate Analysis*, in press.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 55-93.

Hastie, T., et Tibshirani, R.J. (1990). *Generalized Additive Models*. New York : Chapman and Hall.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York : Springer.

Jiang, J., Rao, J.S., Gu, Z. et Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 36, 1669-1692.

Jiang, J., Nguyen, T. et Rao, J.S. (2009). A simplified adaptive fence procedure. *Statistics and Probability Letters*, 79, 625-629.

Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127, 53-69.

Laird, N.M., et Ware, J.M. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.

Meza, J., et Lahiri, P. (2005). Une note sur la statistique C_p sous un modèle de régression à erreur emboîtée. *Techniques d'enquête*, 31, 115-120.

Morris, C.N., et Christiansen, C.L. (1995). Hierarchical models for ranking and for identifying extremes with applications. *Bayes Statistics* 5, Oxford Univ. Press.

Opsomer, J.D., Breidt, F.J., Claeskens, G., Kauermann, G. et Ranalli, M.G. (2007). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, à paraître.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Ruppert, R., Wand, M. et Carroll, R. (2003). *Semiparametric Regression*. Cambridge Univ. Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Sen, A., et Srivastava, M. (1990). *Regression Analysis*. New York : Springer.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223-249.

Wang, J.-L. (2005). Nonparametric regression analysis of longitudinal data. *Encyclopedia of Biostatistics*, 2ième Éd.