Régression bayésienne avec Bspline sous contraintes de régularité et de forme

Khader Khadraoui, Christophe Abraham

▶ **To cite this version:**

Khader Khadraoui, Christophe Abraham. Régression bayésienne avec Bspline sous contraintes de régularité et de forme. 43. Journées de statistique de la SFdS, May 2011, Tunis, Tunisie. hal-02748158

HAL Id: hal-02748158
https://hal.inrae.fr/hal-02748158

Submitted on 3 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Régression bayésienne avec les B-splines sous contraintes de régularité et de forme

Christophe Abraham & Khader Khadraoui

UMR MISTEA/INRA Montpellier, bât. 29, 02 place Viala, 34060 Montpellier cedex 06, France.

Mots clefs: Régression bayésienne, B-spline, polygone de contrôle, contraintes de forme, régularité, recuit simulé.

Abstract

A Bayesian method for regression under shape restrictions and smoothness conditions is proposed. The regression function is built from B-spline basis that controls its regularity. Then we show that its shape can be controlled simply from its cœfficients in the B-spline basis. This is achieved through the control polygon whose definition and some properties are given in this article. The regression function is estimated by the posterior mode. This mode is calculated by a simulated annealing algorithm which allows to take into account the constraints of form in the proposal distribution. A credible interval is obtained from simulations using Metropolis-Hastings algorithm with the same proposal distribution as the simulated annealing algorithm.

Résumé

On propose une méthode bayésienne de régression sous contraintes de régularité et de forme. La fonction de régression est construite à partir d'une base de B-spline qui permet de contrôler sa régularité. On montre ensuite que sa forme peut être contrôlée simplement à partir de ses cœfficients dans la base B-spline. Ce dernier résultat est obtenu par l'intermédiaire du polygone de contrôle dont la définition et certaines propriétés sont données dans cet article. La fonction de régression est estimée par le mode a posteriori. Ce dernier est calculé par un algorithme de type recuit simulé qui permet de prendre en compte la contrainte de forme dans l'étape de proposition. Un intervalle de crédibilité est obtenu à partir de simulations suivant un algorithme de type Metropolis-Hastings avec la même étape de proposition.

1 Introduction

L'estimation d'une fonction de régression f sous des contraintes de régularité et de forme est d'un intérêt considérable dans de nombreuses applications. Les exemples typiques incluent, entre autres, les courbes de doses-réponses en médecine, la construction de la fonction d'utilité, les fonctions de productions dans l'industrie, etc.

Différentes approches de régression sous contraintes sont proposées dans la littérature. Dans le cas fréquentiste, la régression sous contrainte de monotonie avec les splines a été étudiée par Mammen (1991) et Ramsay (1998). Un travail plus général étudiant des contraintes de convexité et de monotonie a été proposé par Meyer (2008). Pour une description des principales méthodes non paramétriques classiques, nous pourrons consulter Delecroix & Thomas-Hagnan (2000). Dans le cadre bayésien les travaux sont plus rares, relativement récents et concernent surtout la régression isotonique (Lavine & Mockus, 1996, Holmes & Heard, 2003, Neelson & Dunson, 2004, Chang et al., 2007 et Shively & Sager, 2009).

Dans cet article, on adopte un cadre bayésien qui permet de prendre en compte des contraintes de régularité et de forme à partir de la loi a priori. Pour obtenir une régression lisse, la fonction de régression est approchée par une spline engendrée par une base de B-spline. La régression avec les B-splines assure une estimation lisse, flexible et parcimonieuse. Le véritable problème consiste ainsi à contrôler la forme d'une spline. Nous montrons que la forme de la fonction de régression peut être contrôlée grâce au polygone de contrôle. Nous estimons f par le mode a posteriori, car contrairement à l'espérance a posteriori, il vérifie nécessairement les contraintes de forme.

Dans la Section 2, on définit les B-splines et le polygone de contrôle et on montre que la forme d'une courbe peut être contrôlée par un ensemble de points de contrôle qui ne sont pas situés sur celle-ci (de Boor, 2001). La Section 3 est consacrée à l'inférence sous des contraintes de forme et de régularité. En particulier, on montre que l'estimateur bayésien peut être calculé à partir d'un algorithme de type recuit simulé et qu'il est possible de générer suivant la loi a posteriori à partir d'un algorithme de type Metropolis-Hastings.

2 B-splines univariées et polygone de contrôle

On propose de construire f par approximation en utilisant une base B-spline sur un intervalle $[a, b]$. On se donne une suite de points $\{t_i \in \mathbb{R} | t_i \leq t_{i+1}, i = 1, \ldots, m_k - k\}$, appelés nœuds. On précise que pour la construction d'une base B-spline, il est nécessaire d'ajouter $2k$ nœuds extérieurs aux nœuds intérieurs $(t_i)_{i=1}^{m_k-k}$. Le vecteur $\mathbf{t} \stackrel{\text{déf}}{=} (t_i)_{i=1}^{m_k+k}$ est appelé le vecteur nodal. On note B_{ik} la B-spline d'ordre k et de support $[t_i, t_{i+k}]$ dont on peut trouver une définition dans de Boor (2001). La spline d'ordre k et de vecteur nodal $\mathbf{t}$, est par définition, une combinaison linéaire des m_k B-splines B_{ik} associées au vecteur $\mathbf{t}$. On note l'ensemble de ces splines par:

$$S_{k,\mathbf{t}} \stackrel{\text{déf}}{=} \Big\{ \sum_{i=1}^{m_k} B_{ik}\beta_i \ : \ \beta_i \in \mathbb{R} \Big\}. \tag{1}$$

Il est bien connu que l'espace $S_{k,\mathbf{t}}$ est un espace de fonctions polynomiales par morceaux de degré $< k$ et de points de rupture t_i. Un élément de $S_{k,\boldsymbol{t}}$ est de classe $C^{k-1-\#t_i}$ à

chaque nœud t_i où $\#t_i$ est la multiplicité des nœuds définie par:

$$\#t_i \stackrel{\text{déf}}{=} \#\{t_j : t_i = t_j\}.$$

Il est connu également que la régression B-spline a de bonnes propriétés numériques, assure une implémentation facile et permet de contrôler la régularité (dérivabilité). Par contre, la régression B-spline reste sensible au nombre et aux positions des nœuds (Meyer, 2008).

Le polygone de contrôle associé à une spline $s_k \in S_{k,\mathbf{t}}$ est la fonction linéaire par morceaux qui interpole les sommets P_i définis par:

$$P_i \stackrel{\text{déf}}{=} (t_i^*, \beta_{i-1}), \tag{2}$$

où t_i^* est défini par

$$t_i^* \stackrel{\text{déf}}{=} (t_{i+1} + \cdots + t_{i+k-1})/(k-1). \tag{3}$$

On note par $C_{\beta,\boldsymbol{t}}$ le polygone de contrôle. Les points $P_i \in \mathbb{R}^2$ sont appelés points de contrôle. La proposition ci-dessous montre qu'il est possible de contrôler la forme d'une spline par l'intermédiaire de son polygone de contrôle.

Définition 2.1 *(i) Une fonction f continue est κ-monotone sur $[a,b]$ s'il existe exactement $\kappa - 1$ réels distincts $x_1, \ldots, x_{\kappa-1} \in]a,b[$ tels que f est croissante (resp. décroissante) sur $[a, x_1]$ et $[x_j, x_{j+1}]$ où $j = 2, 4, \ldots$ et décroissante (resp. croissante) sur $[x_{j'}, x_{j'+1}]$ où $j' = 1, 3, \ldots$.*

*(ii) La fonction f est dite **unimodale** s'il existe $x^* \in]a,b[$ tel que f est croissante sur $[a, x^*]$ et décroissante sur $[x^*, b]$.*

Proposition 2.2 *(i) (Préservation de la monotonie) Si le polygone de contrôle $C_{\beta,\mathbf{t}}$ est monotone alors la spline associée s_k est également monotone.*

(ii) (Préservation de la convexité) Si le polygone de contrôle $C_{\beta,\mathbf{t}}$ est convexe, alors la spline associée s_k est également convexe.

(iii) (Préservation de la κ-monotonie) Si le polygone de contrôle $C_{\beta,\mathbf{t}}$ est κ-monotone alors la spline associée s_k est κ'-monotone avec $\kappa' \leq \kappa$.

(iv) (Préservation de l'unimodalité) En particulier, si le polygone de contrôle $C_{\beta,\mathbf{t}}$ est unimodal alors la spline associée s_k est unimodale ou monotone.

Ainsi le paramètre principal est le polygone de contrôle, dont la spline associée épouse les formes.

3 Inférence bayésienne sous contraintes

Nous considérons le modèle usuel de régression suivant reliant le vecteur de réponse $\boldsymbol{y} = (y_1, \ldots, y_n)'$ et le vecteur de covariable $\boldsymbol{x} = (x_1, \ldots, x_n)'$:

$$\begin{cases} y_j = f(x_j) + \varepsilon_j, & j = 1, \ldots, n, \\ f(x_j) = \sum_{i=1}^{m_k} \beta_i B_{i,k}(x_j). \end{cases} \tag{4}$$

où f est une fonction de régression inconnue et les ε_j sont indépendants de loi $N(0, \sigma^2)$.

3.1 La distribution a posteriori

Soit S l'ensemble des vecteurs $\beta = (\beta_i)_{i=1}^{m_k}$ tels que f respecte la contrainte de forme. Par exemple, d'après la proposition 2.2, si on impose à f d'être croissante, on a :

$$S = \{f \,|\, \beta_i \in \mathbb{R}, \beta_1 \leq \beta_2 \leq \cdots \leq \beta_{m_k}\},$$

pour une régression unimodale concave, on a :

$$S = \{f \,|\, \beta_i \in \mathbb{R}, \beta_1 < \beta_2, \beta_{m_k-1} > \beta_{m_k}, \beta_i - 2\beta_{i-1} + \beta_{i-2} \leq 0\},$$

et pour une régression unimodale, on a :

$$S = \{f \,|\, \beta_i \in \mathbb{R}, \beta_1 = \beta_2 < \beta_3 \leq \cdots \leq \beta_l \geq \beta_{l+1} \geq \cdots \geq \beta_{m_k-1} > \beta_{m_k}, \text{pour } l = 4, \ldots, m_k - 1\}.$$

Il suffit alors de choisir une loi a priori pour les paramètres (β, σ^2) du modèle (4): on pose $\beta|\sigma^2 \sim N^S_{m_k}(m, \sigma^2 V)$, où $N^S_{m_k}(\cdot, \cdot)$ est une loi normale tronquée à S, et $\sigma^2 \sim IG(\xi, \varsigma)$. On obtient ainsi une loi a priori π^S, conditionnée par la contrainte $f \in S$, dont la densité est donnée, à une constante près, par:

$$\pi^S(\beta, \sigma^2) \propto (\sigma^2)^{-(\xi + (m_k/2) + 1)} \exp\Big\{ - \frac{(\beta - m)' V^{-1} (\beta - m) + 2\varsigma}{2\sigma^2} \Big\} \mathbf{1}_{\{\beta \in S\}}.$$

La forme de f est garantie par l'indicatrice $\mathbf{1}_{\{\beta \in S\}}$ et la régularité est contrôlée par k l'ordre de la spline.
Il est facile de voir que la densité de la loi a posteriori est proportionnelle à $\mathbf{1}_{\{\beta \in S\}}$ multipliée par la densité de loi normale inverse gamma que l'on aurait obtenu sans considérer la restriction à S.

3.2 Implémentation de l'estimateur

Nous proposons d'estimer f par le mode a posteriori. Pour calculer le mode, il suffit de trouver, $\beta^\star$, l'argument minimum de:

$$Q(\beta) = (\beta - m^*)' (V^*)^{-1} (\beta - m^*),$$

où

$$m^* = (V^{-1} + B'B)^{-1}(V^{-1}m + B'\boldsymbol{y});$$
$$V^* = (V^{-1} + B'B)^{-1},$$

sous la contrainte $\beta \in S$. Nous calculons $\beta^\star$ par un algorithme de type recuit simulé. L'algorithme du recuit est composé de deux étapes: une étape de proposition qui permet l'exploration de l'espace d'états et une étape d'acceptation/rejet. La contrainte de forme sera garantie grâce à l'étape de proposition. À partir de la valeur β^t à l'iteration t de l'algorithme, on propose de remplacer β^t par $\widetilde{\beta} = (\beta_1^t, \ldots, \beta_{l-1}^t, \widetilde{\beta}_l, \beta_{l+1}^t, \ldots, \beta_{m_k}^t)'$ tel que:

(*i*) $l \sim \mathcal{U}_{\{1,\ldots,m_k\}}$, tirage de la composante de β^t à modifier;

(*ii*) $\widetilde{\beta}_l \sim \mathcal{U}_{\{S(\beta^t,l)\cap[\beta_l^t \pm \varepsilon_0]\}}$ où $S(\beta^t, l) = \{\widetilde{\beta}_l : (\beta_1^t, \ldots, \beta_{l-1}^t, \widetilde{\beta}_l, \beta_{l+1}^t, \ldots, \beta_{m_k}^t)' \in S\}$ et ε_0 est une constante positive qui contrôle la variabilité de la proposition.

La valeur initiale β^1 peut être choisie selon un examen visuel des données $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Pour avoir une bande de confiance autour de l'estimateur, nous proposons des simulations suivant la loi a posteriori par un algorithme de type Metropolis-Hastings construit à partir de la même proposition que pour l'algorithme de recuit..

3.3 Application

La méthode bayésienne de régression sous contrainte est appliquée sur les données d'une étude récente (Jeanson et al., 2009) qui porte sur les cinétiques d'acidification (diminution de pH). Des problèmes d'interférences électriques ont fait artificiellement remonter la mesure du pH au cours des fermentations. Dans cette application, nous proposons la reconstruction d'une courbe d'acidification en imposant la décroissance du pH au cours du temps. La Figure 1 illustre la reconstruction de la courbe où le mode a posteriori vérifie la décroissance de la mesure du pH.

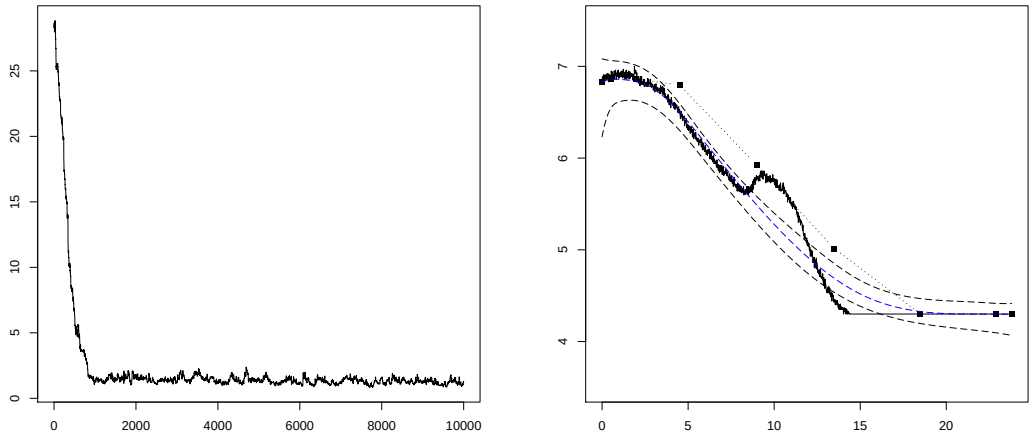
Bibliographie

[1] Cai, B. et Dunson, D.B. (2007), Bayesian multivariate isotonic regression splines: Applications to carcinogenicity studies, Journal of the American Statistical Association, **102**, 1158-1171.

[2] Chang, I-S., Chien, L-C., Hsiung, C. A. Wen, C-C. et Wu, Y-J. (2007), Shape restricted regression with random bernstein polynomials, Complex Datasets and Inverse Problems: Tomography, Networks and Beyond, **54** :187-202.

[3] de Boor, C. (2001), B(asic)-spline basics, Computer Aided Geometric Design (CAGD), Handbook, **1** :1-34.

[4] Delecroix, M. et Thomas-Agnan, C. (2000), Spline and Kernel Regression under Shape restrictions, Jhon Wiley Sons, New-York.
[5] Denison, D. G., Holmes, C. C., Mallick, B. K. et Smith, A. F. M. (2002), Bayesian Methods for Nonlinear Classification and Regression, John Wiley Sons, New-York.
[6] Holmes, C. C. et Heard, N. A. (2003), Generalized monotonic regression using random change points, Statistics in Medecine, **22** :623-638.
[7] Jeanson, S., Hilgert, N., Coquillard, M., Seukpanya, C., Faiveley, M., Neuveu, P., Abraham, C., Georgescu, V., Fourcassi, P. et Beuvier, E. (2009), Milk acidification by lactococcus lactis is improved by decreasing the level of dissolved oxygen rather than decreasing redox potential in the potential in the milk prior to inoculation, International Journal of Food Microbiology, **131** :75-81.
[8] Lavine, M. et Mockus, A. (1995), A non parametric bayes method for isotonic regression, Journal of Statistical Planning and Inference, **46** :235-248.
[9] Mammen, E. (1991), Estimating a smooth monotone regression function, Annals of Statistics, **19** :724-740.
[10] Meyer, M. C. (2008), Inference using shape-restricted regression splines, The Annals of Applied Statistics, **2** :1013-1033.
[11] Neelon, B. et Dunson, D. B. (2004), Bayesian isotonic regression and trend analysis, Biometrics, **60** :398-406.
[12] Shively, T. S. et Sager, T. W. (2009), A bayesian approach to non-parametric monotone function estimation, Journal of the Royal Statistical Society, **71** :159-175.



(a) *Minimisation de la fonction $Q(\beta)$ par le recuit simulé.*

(b) *Bande de confiance à 5% pour le mode a posteriori.*

Figure 1: (a) Les simulations par le recuit simulé. (b) Le mode a posteriori (- - en bleu), la bande de confiance à 5% (- -) et le polygone de contrôle (. . . . .).