Régression linéaire fonctionnelle bayésienne explicable

Paul-Marie Grollemund [1,*,†] & Christophe Abraham [2,†] &
Meïli Baragatti [3,†] & Pierre Pudlo [4,*]

[*] *UMR I3M CNRS 5149, Université de Montpellier, Place E. Bataillon 34095, Montpellier cedex, France*
[†] *UMR Mistea, Montpellier SupAgro-INRA, 2 place Pierre Viala, 34060 Montpellier cedex 2, France*
[1] *paul-marie.grollemund@univ-montp2.fr*
[2] *christophe.abraham@supagro.inra.fr*
[3] *meili.baragatti@supagro.inra.fr*
[4] *pierre.pudlo@univ-montp2.fr*

Résumé. Nous nous plaçons dans le cadre d'un modèle de régression linéaire où la variable à expliquer est réelle et la covariable est fonctionnelle. Nous proposons un modèle bayésien basé sur la projection de ce paramètre dans une base d'histogrammes parcimonieuse et adaptative. Afin d'obtenir une estimation de la fonction coefficient explicable, nous sommes aussi amenés à introduire une nouvelle fonction de coût. Certaines grandeurs du modèle proposé étant analytiquement intractables, il est nécessaire en pratique d'utiliser des stratégies *MCMC* pour les déterminer. La structure des estimations obtenues facilite, autant qu'il soit possible, leur interprétation.

Mots-clés. Statistique bayésienne, régression linéaire fonctionelle, parcimonie

Abstract. Our Study is in the context of a linear regression model where the dependent variable is real and the covariate is functional. We propose a Bayesian model based on the adaptive decomposition of the coefficient function into a sparse function. In order to produce an interpretable estimate, we also have to introduce a new loss function which excludes non-sparse functions. Some quantities of our model are analytically intractable, so it is necessary in practice to use MCMC strategies. We produce simple estimates which are as simple as possible to interpret.

Keywords. Bayesian statistics, functional linear regression, interpretable

1 Introduction

Ces deux dernières décennies de nombreux outils ont été développés pour expliquer une variable réelle à partir d'un variable fonctionnelle. Dans ce cadre-là, un modèle statistique

est la régression linéaire fonctionnelle. Pour des observations (Y_i, X_i), $i = 1, \ldots, n$ où Y_i est un réel et X_i peut-être vu comme une fonction d'un intervalle $\mathcal{T}$ dans $\mathbb{R}$, on a:

$$Y_i = \mu + \int_{\mathcal{T}} X_i(t)\beta(t)\mathrm{d}t + \varepsilon_i \tag{1}$$

où le paramètre est β est appelé fonction coefficient. La référence pour l'étude de ce modèle est le livre de Ramsay et Silverman (2005), mais il existe un certain nombre de variantes dont nous pouvons citer, sans être exhaustifs, Cardot et al. (2003), Crambes et al. (2009), Yuan et Cai (2010), ainsi que des généralisations Müller et Stadtmüller (2005), McLean et al. (2014).

La problématique qui nous intéresse ici est d'expliquer la variable réponse à partir de la covariable en estimant avec parcimonie le paramètre β. Dans le cas d'un modèle de régression linéaire multiple avec des covariables réelles, la parcimonie se conçoit au sens où la plupart régresseurs sont estimés à 0, les excluant ainsi du modèle. Lorsque la covariable est fonctionnelle, une fonction coefficient sera dite parcimonieuse si elle admet des zones de nullité. Nous considérons de plus qu'une fonction coefficient sera explicable si elle est constante par morceaux en plus d'être parcimonieuse.

C'est pour répondre à ce genre de problématiques que James et al. (2009) ont développé une approche originale qui permet d'estimer β par une fonction simple et parcimonieuse. Cette méthode favorise la nullité de la fonction coefficient et de ses dérivées sur leurs supports. La simplicité des résultats obtenus par cette méthode permet de mettre en évidence assez clairement les zones du support sur lesquelles la covariable explique suffisamment bien la variable réponse. Cependant, cette méthode souffre d'une instabilité face à des paramètres de calibration. D'autre part, la loi de l'estimateur n'est pas explicite ce qui entraîne de devoir utiliser du *bootstrap* pour déterminer des intervalles de confiance.

Pour avoir une réponse à cette problématique qui s'affranchisse de ces problèmes, nous développons une approche bayésienne du modèle de régression linéaire fonctionnelle. Contrairement à la méthode précédente, le cadre bayésien nous permet une inférence plus complète (estimation, intervalle de confiance, erreur de prédiction,...) assez naturellement, au travers de la distribution *a posteriori*. L'approche bayésienne du modèle de régression linéaire a été largement étudiée, mais ce n'est pas le cas pour l'extension de ce modèle aux données fonctionnelles, pour lequel nous pouvons citer Crainiceanu et Goldsmith (2010), Goldsmith et al. (2011) ou encore Hui et al. (2013).

2 Modèle

Dans le cas où nous souhaitons expliquer le lien entre les observations $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ d'une variable réelle et un échantillon $X = (X_1, \ldots, X_n)^T$ où chaque X_i peut-être vu comme un processus en temps continu, le modèle de régression linéaire bayésien s'écrit

$$Y|X, \mu, \beta, \sigma^2 \sim \mathcal{N}_n\left(\mu\mathbf{1}_n + \int_{\mathcal{T}} X(t)\beta(t)\mathrm{d}t \ , \ \sigma^2 I_n\right). \tag{2}$$

Nous imposons au coefficient β la forme suivante :

$$\beta(t) = \sum_{k=1}^{K} \beta_k^* \mathbf{1}\left\{t \in I_k\right\} \tag{3}$$

où $\beta_k^* \in \mathbb{R}^*$ pour $k = 1, \ldots, K$ et $\mathbf{1}\left\{A\right\}$ est la fonction indicatrice. Les I_k sont donc les intervalles sur lesquels la fonction β est non nulle. Nous ne faisons pas d'hypothèses supplémentaires sur ces intervalles, notamment ils ne sont pas nécessairement disjoints. Cette contrainte est intéressante dans le cas où $\bigcup I_k \neq \mathcal{T}$ car cela implique que β admet des zones nulles.

Imposer cette forme à β revient à la projeter dans une base incomplète d'histogramme et injecter (3) dans (2) nous amène à considérer une régression linéaire sur les projetées $X_I^* = \left(X_{I_1}^*, \ldots, X_{I_K}^*\right)$ où

$$X_{I_k}^* = \int_{I_k} X(t)\mathrm{d}t.$$

Pour notre modélisation, nous considérons que ce calcul intégral est fait sans erreur et nous aboutissons à :

$$Y|X, \mu, \beta^*, \sigma^2, I \sim \mathcal{N}_n\left(\mu\mathbf{1}_n + X_I^*\beta^* \ , \ \sigma^2 I_n\right) \tag{4}$$

L'idée principale de cette modélisation est d'adapter la base de projection à une explication parcimonieuse de la réponse à partir de la covariable.

Pour travailler avec des intervalles I_k variables, il sera plus aisé de les caractériser par leurs milieux m_k et leurs étendues ℓ_k : $I_k = [m_k - \ell_k, m_k + \ell_k]$ pour $k = 1, \ldots, K$ et nous noterons alors la matrice des projetées $X_{m\ell}^*$ au lieu de X_I^* . Nous complétons alors le modèle bayésien (4) avec les lois *a priori* suivantes :

$$\begin{aligned} Y|X, \mu, \beta^*, \sigma^2, m, \ell &\sim \mathcal{N}_n\left(\mu\mathbf{1}_n + X_{m\ell}^*\beta^* \ , \ \sigma^2 I_n\right), \\ \mu|\sigma^2 \sim \ & \mathcal{N}\left(\eta_0, v_0\sigma^2\right), & \ell \sim \ & \mathcal{U}\left(]0, \ell_{\max}]^K\right), \\ \beta^*, \sigma^2 \sim \ & \mathcal{NIG}_K\left(\eta, V, a, b\right), & m \sim \ & \mathcal{U}\left(\mathcal{T}^K\right), \end{aligned}$$

où $\mathcal{NIG}_K$ désigne une loi normale-inverse-gamma de dimension K et $\eta_0, v_0, \eta, V, a, b$ et $\ell_{\max}$ sont des hyperparamètres. Il est possible de choisir ces hyperparamètres de manière à ce que les lois *a priori* soient les moins informatives possibles. Pour le cas particulier du paramètre K, nous le choisissons assez grand ce qui garantit une erreur faible dans la mesure où les intervalles peuvent se chevaucher. Nous serons amenés ultérieurement à une problématique de choix de modèle par rapport à la valeur de ce paramètre.

3 Inférence

L'objectif principal de notre problématique est de construire un estimateur ponctuel particulier du coefficient β de la régression. La nature fonctionnelle de ce paramètre ajoute une

difficulté majeure au problème d'estimation. Comme la contrainte (3) n'est pas linéaire, une première idée est de prendre directement l'espérance *a posteriori* de β marginalement en t :

$$\hat{\beta}_1(t) := \int \beta(t)\pi\left(\theta|Y,X\right)\mathrm{d}\theta$$

où $\theta = (\mu, \beta^*, \sigma^2, m, \ell)$ est l'ensemble des paramètres du modèle précédent. Mais la fonction $\hat{\beta}_1$ sort des contraintes imposées, elle n'est pas constante par morceaux car l'ensemble $\mathcal{E}$ des fonctions explicables décrit par la contrainte (3) n'est pas convexe.

Une idée qui assure d'avoir une estimation satisfaisant les contraintes est d'estimer le mode *a posteriori* $\hat{\theta}^{\mathrm{mod}}$ puis de reconstruire $\hat{\beta}_2(t)$ partir de $\hat{\theta}^{\mathrm{mod}}$ et de la contrainte (3). En pratique, cet estimateur a de très mauvaises propriétés car la distribution *a posteriori* est multimodale à cause de l'interchangeabilité des paramètres (β_i^*, m_i, ℓ_i). La meilleure solution est d'introduire la fonction de coût naturelle

$$L(d,\beta) = +\infty \mathbf{1}\left\{d \in \mathcal{E}^c\right\} + \|d-\beta\|^2 \mathbf{1}\left\{d \in \mathcal{E}\right\}. \tag{5}$$

L'estimateur de Bayes associé à ce coût est la fonction qui minimise le coût moyen *a posteriori*

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \int L(\beta,\beta')\pi\left(\theta|Y,X\right)\mathrm{d}\beta'. \tag{6}$$

Comme le coût est infini hors du support de la distribution *a priori*, $\hat{\beta}$ conduit naturellement à une fonction constante par morceaux obéissant aux contraintes.

4 Résultats numériques

Pour calculer une estimation à partir des données, il est tout d'abord évidemment nécessaire de déterminer la loi de la distribution *a posteriori* jointe. Étant donnée les lois *a priori* non conjuguées, il n'est pas possible de la caractériser analytiquement. Dans ce cas, il est possible d'utiliser une méthode *MCMC* pour échantillonner cette loi (voir Tierney (1994)). Ici, nous utilisons un *échantillonneur de Gibbs* car les distributions conditionnelles complètes s'écrivent assez simplement. Pour des considérations computationnelles, nous travaillons

avec les paramètres m et ℓ coordonnées par coordonnées :

$$\begin{aligned}
\mu|Y,X,\beta^*,\sigma^2,m,\ell &\sim \mathcal{N}\left(\frac{\eta_0 v_0^{-1}+\mathbf{1}_n^T\left(Y-X^*_{m\ell}\beta^*\right)}{n+v_0^{-1}},\frac{\sigma^2}{n+v_0^{-1}}\right)\\
\beta^*|Y,X,\mu,\sigma^2,m,\ell &\sim \mathcal{N}\left({X^*_{m\ell}}^T(Y-\mu\mathbf{1}_n)+V^{-1}\eta,{X^*_{m\ell}}^T X^*_{m\ell}+V^{-1}\right)\\
\sigma^2|Y,X,\mu,\beta^*,m,\ell &\sim \mathcal{IG}\left(a+\frac{n+K+1}{2},b_{\sigma^2}\right)\\
\pi\left(m_k|Y,X,\mu,\beta^*,\sigma^2,m_{-k},\ell\right) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|Y-\mu\mathbf{1}_n-X^*_{m\ell}\beta^*\right\|^2\right\}\mathbf{1}\left\{m_k\in\mathcal{T}\right\}\\
\pi\left(\ell_k|Y,X,\mu,\beta^*,\sigma^2,\ell_{-k},m\right) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left\|Y-\mu\mathbf{1}_n-X^*_{m\ell}\beta^*\right\|^2\right\}\mathbf{1}\left\{\ell_k\in\left]0,\ell_{max}\right]\right\}
\end{aligned}$$

avec $b_{\sigma^2}=b+\frac{1}{2}\left\|Y-\mu\mathbf{1}_n-X^*_{m\ell}\beta^*\right\|^2+\frac{1}{2v_0}\left(\mu-\eta_0\right)^2+\frac{1}{2}\left\|\beta^*-\eta\right\|^2_{V^{-1}}$. Les distributions conditionnelles complètes des paramètres m_k et ℓ_k ne sont pas des lois usuelles. Cependant, en considérant que m_k et ℓ_k évoluent dans une grille finie de leurs supports, il est possible de caractériser ces lois en calculant numériquement leurs fonctions de probabilité. Cette grille de discrétisation peut être trivialement $\mathcal{T}_G=\{t_1,\ldots,t_p\}$ l'ensemble des instants d'observation des courbes X_i.

L'espérance *a posteriori* du coût (6), à β fixé, est alors simplement calculable à partir des résultats de l'*échantillonneur de Gibbs*. Pour déterminer la fonction β qui minimise cette quantité, nous utilisons un recuit simulé (Bélisle (1992)). Par définition du coût (5), les fonctions qui ne sont pas dans l'ensemble $\mathcal{E}$ sont automatiquement éliminées, nous restreignons donc les propositions du recuit simulé à l'ensemble $\mathcal{E}$.

Sur des jeux de données simulés, nous obtenons des estimations visuellement explicables et reconstituant assez fidèlement le signal (voir Fig. 1).

5 Conclusion et discussions

La méthode que nous proposons ici répond à la problématique initiale. L'estimation que nous obtenons admet des zones de nullité lorsque la fonction cible est proche de 0 et s'ajuste assez fidèlement ailleurs quelque soit la régularité du signal. Les résultats obtenus réalisent en pratique un bon compromis entre parcimonie, simplicité et ajustement. En cela, cette méthode permet de faciliter autant que possible l'interprétation de l'estimation de la fonction coefficient.

Lors de l'exposé, nous présenterons d'autres résultats de notre estimateur et une description du *posterior* sur des jeux de données simulés et réels. Nous nous comparerons à d'autres méthodes classiques en analyse de données fonctionnelles (*FDA*) ou qui fournissent des résultats explicable (*FLiRTI*, *Bayesian Fused Lasso*).

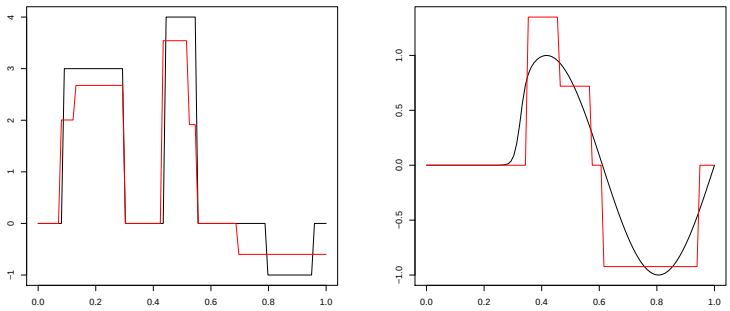

Figure 1: Ces deux graphiques illustrent les résultats que nous obtenons sur deux jeux de données simulées. Dans les deux cas, il y a 50 *individus* et les courbes X_i sont évaluées en 100 points de discrétisation. Les données Y_i sont calculées à partir du modèle (2), avec un bruit gaussien, une fonction coefficient donnée et des X_i simulées suivant un processus gaussien. Nous présentons pour chacun des jeux de données, l'estimation que nous obtenons (en rouge) de la fonction coefficient que nous nous donnons (en noir). Dans le premier cas (graphique de gauche), nous prenons β appartenant à l'ensemble $\mathcal{E}$ et pour le second (graphique de droite) β n'appartenant pas à $\mathcal{E}$.

À partir de ce travail, plusieurs directions de recherches semblent intéressantes. Premièrement, le choix bayésien et la structure de notre modélisation donne un cadre adapté pour inclure de la connaissance *a priori* dans le modèle. Il est par exemple raisonnable pour certains praticiens d'avoir une idée à l'avance des zones potentiellement importantes. Finalement, une idée motivée par des problématiques pratiques consiste à inclure dans le modèle une variable qualitative et son interaction avec la variable fonctionnelle.

Bibliographie

[1] Bélisle, C. (1992) Convergence theorems for a class of simulated annealing algorithms on Rd, *J. Appl. Prob.*

[2] Cardot, H., Ferraty, F. et Sarda, P. (2003) Spline estimators for the functional linear model, *Statistica Sinica*

[3] Crainiceanu, C. et Goldsmith, J. (2010) Bayesian Functional Data Analysis Using WinBUGS, *J. Stat. Softw.*

[4] Crambes, C., Kneip, A. et Sarda, P. (2009) Smoothing splines estimators for functional linear regression, *Ann. Statist.*

[5] Goldsmith, J., Wand, M.,P. et Crainiceanu, C. (2011) Functional regression via variational Bayes, *Electron J. Stat.*

[6] Hui, S.,K., Meyvis, T. et Assael, H. (2013) Analyzing Moment-to-Moment Data Using

a Bayesian Functional Linear Model: Application to TV Show Pilot Testing, *Marketing Science*
[7] James, G.M., Wang, J. et Zhu, J. (2009), Functional linear regression that's interpretable, *Ann. Statist.*.
[8] Kyung, M., Gill, J., Ghosh, M. et Casella, G. (2005) Penalized Regression, Standard Errors, and Bayesian Lassos, *International Society for Bayesian Analysis.*
[9] McLean, M. W., Hooker, G., Staicu, A., Scheipl, F. et Ruppert, D. (2014) Functional Generalized Additive Models, *J. of Comput. and Graph. Stat.*
[10] Müller, H. G. et Stadtmüller , U. (2005) Generalized functional linear models, *Ann. Statist.*
[11] Ramsay, J.O. et Silverman, B.W. (2005), *Functional Data Analysis*, 2nd ed. Springer, New York.
[12] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. et Knight, K. (2005) Sparsity and smoothness via the fused lasso, *J. R. Statist. Soc.*
[13] Tierney, L. (1994) Markov chains for exploring posterior distributions, *Ann. Statist.*