Bayesian estimation of a functional linear model that is interpretable

Meïli Baragatti[1,*]

[1] *UMR MISTEA, Montpellier SupAgro-INRA, 2 place Pierre Viala, 34060 Montpellier cedex 2, France.*

[*] *meili.baragatti@supagro.inra.fr.*

Résumé. En analyse de données fonctionnelles, un modèle couramment utilisé est le modèle de régression linéaire fonctionnelle, dans lequel des réponses scalaires Y_i sont reliées à des observations d'un processus stochastique à temps continu $X_i(t)$, par la relation suivante: $Y_i = \beta_0 + \int X_i(t)\beta(t)dt + \epsilon_i$, $i = 1, \ldots, n$. L'estimation du coefficient fonctionnel d'intérêt $\beta(.)$ a été largement étudiée. Une approche intéressante appelée FLiRTI a notamment été proposée par James et al. [2009] dans un cadre fréquentiste. L'objectif de cette méthode est d'obtenir un estimateur "interprétable" de $\beta(.)$, soit ayant une forme simple. Plus spécifiquement, l'estimateur doit être nul sur certaines régions, et doit avoir une structure simple ailleurs, c'est à dire linéaire par morceaux. Cette méthodologie est particulièrement attractive pour des applications en agronomie ou biologie, étant donné que les agronomes et biologistes sont demandeurs de fonctions ayant des formes simples et avec beaucoup de régions nulles, ce qui leur est facilement interprétable. Nous souhaitons donc développer une méthode produisant de tels estimateurs "interprétables" pour $\beta(.)$, mais dans un cadre bayésien. En effet, cela nous permettra de prendre en compte la connaissance a priori d'experts, agronomes ou biologistes par exemple. Nous voulons donc étendre la modélisation FLiRTI dans un cadre bayésien, puis proposer une manière d'estimer le coefficient fonctionnel d'intérêt $\beta(.)$. Cette approche sera développée puis illustrée sur des données réelles et simulées, et enfin comparée aux approches classiques implémentées dans le package R FDA et à la méthodologie FLiRTI.

Mots-clés. régression linéaire fonctionnelle, interprétabilité, outils de sélection de variables, estimation bayésienne.

Abstract. In functional data analysis, a commonly used model is the functional linear regression model, in which scalar responses Y_i are related to sampled paths from some underlying continuous-time stochastic process $X_i(t)$, through $Y_i = \beta_0 + \int X_i(t)\beta(t)dt + \epsilon_i$, $i = 1, \ldots, n$. The estimation of the functional coefficient of interest $\beta(.)$ have been substantially studied. An interesting approach called FLiRTI has been proposed by James et al. [2009] in a frequentist framework. The objective is to obtain "interpretable" estimates for $\beta(.)$, that is which have simple shapes. In particular, the estimates should have null regions, and should have very simple structures elsewhere, typically piecewise linear structures. This methodology is particularly appealing for applications in agronomy and

biology, as agronoms and biologists are often willing to obtain simple shapes with null regions. We are interested by developing a method providing similar "interpretable" estimates for $\beta(.)$, but in a Bayesian framework. Indeed, it would enable us to easily take into account prior knowledge from experts like agronoms or biologists. We then extend the FLiRTI modelisation in a Bayesian framework, and propose a way to estimate the functional coefficient of interest $\beta(.)$. The approach developed is then illustrated on simulated and real datasets, and compared to the standard approach implemented in the FDA R-package and to the FLiRTI methodology.

Keywords. functional linear regression, interpretability, variable selection tools, Bayesian estimation.

1 Introduction

In the past ten years, functional data analysis (FDA) has been an active area of research, due to expanding methods for the acquisition and storage of data. Indeed, an observation can be viewed as a curve or a function, representing a sample path from some underlying continuous-time stochastic process $\{X(t)t \in \mathcal{T}\}$. The observed sampled paths are denoted $X_i(t)$, $i = 1, \ldots, n$. A common model studied in FDA is the functional linear regression model, in which the observed sampled paths are associated to scalar responses $Y_i, i = 1, \ldots, n$:

$$Y_i = \beta_0 + \int_{\mathcal{T}} X_i(t)\beta(t)dt + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\beta(.)$ is the functional coefficient of interest, and $\beta(t)$ describes the effect of the functional predictor at time t on the response. Common approaches to estimate $\beta(.)$ are the regularization using basis functions and the regularization with roughness penalties, see chapter 10 in Ramsay and Silverman [2006]. Another approach, called FLiRTI and developed by James et al. [2009], appeared interesting to us. The objective of this approach is to obtain "interpretable" estimates for $\beta(.)$, that is which have simple shapes. In particular, the estimates should have null regions, and should have very simple structures elsewhere, typically piecewise linear structures. James et al. [2009] obtained such estimates by using derivatives of $\beta(.)$ and assuming them sparse. Estimation is then done using variable selection tools. This methodology is particularly appealing for applications in agronomy and biology, as agronoms and biologists are often willing to obtain simple shapes with null regions. Indeed, that enables them to point out regions of $X_i(.)$ which are influent on Y_i, and over these influent regions they can know basically the effect of $X_i(.)$ on Y_i. Obtaining such estimates can support some of their biological assumptions, or can help them to explain some phenomenons. More complex shapes can be difficult for them to interpret, all the more if they suspect the presence of artefacts which do not reflect the reality. Hence they usually prefer basic but reliable and easily interpretable estimates. All the approaches previously mentioned are in a frequentist framework. Papers

using Bayesian approaches are not numerous, and a reason speculated by Crainiceanu and Goldsmith [2009] is that the Bayesian inferential tools are perceived as unnecessarily complex and hard to implement. But later on, Crainiceanu and Goldsmith [2010] have disproved this argument by showing that Bayesian inference can be simple and efficient.

We are then interested by developing a method providing "interpretable" estimates for $\beta(.)$, but in a Bayesian framework. Indeed, it would enable us to easily take into account prior knowledge from experts like agronoms or biologists.

2 Modelisation

The modelisation used similar to those in James et al. [2009]. In a Bayesian framework, the difference will be that we will use supplementary latent vectors to model the sparsity of the derivatives, and we will assume prior distributions for the unknown parameters.

2.1 The FLiRTI model

As in the basis approach, the functional coefficient $\beta(t)$ is projected in a p-dimensional basis $\mathbf{B}(t) = [b_1(t), b_2(t), \ldots, b_p(t)]^T$. The dimension p is chosen large enough so that $\beta(t)$ is well approximated, and the error term $e(t)$ is small:

$$\beta(t) = \mathbf{B}(t)^T\eta + e(t), \qquad p \gg n. \tag{2}$$

The simple grid basis is used in a first approach, that is $b_k(t) = 1$ if $\{\frac{k-1}{p} < t \leq \frac{k}{p}\}$ and $b_k(t) = 0$ otherwise. Combining (1) and (2) we obtain:

$$Y_i = \beta_0 + \mathbf{X}_i^T\eta + \epsilon_i^*, \quad \text{with} \quad \mathbf{X}_i^T\eta = \begin{pmatrix} \int X_i(t)b_1(t)dt \\ \vdots \\ \int X_i(t)b_p(t)dt \end{pmatrix}^T \eta. \tag{3}$$

Our objective is to obtain an estimate of $\beta(t)$ which is null on regions where $X_i(.)$ is not very influent on Y_i, and which is piecewise linear elsewhere. It is equivalent to assume that its derivatives of order 0 and 2 are sparse. Indeed, $X_i(.)$ has no influence on Y_i in regions where $\beta^{(0)}(t) = 0$, and $\beta(t)$ is piecewise linear in regions where $\beta^{(2)}(t) = 0$, hence it is easily interpretable by biologists or agronoms.
The idea is to approximate the derivatives by using a grid of $(p+1)$ evenly spaced points on $[0, 1]$, that is $\{0, \frac{1}{p}, \frac{2}{p}, \ldots, 1\}$. We assume without loss of generality that the $X_i(.)$ are scaled such that $0 \leq t \leq 1$. The derivatives are approximated by using the forward finite differences: $D\mathbf{B}(t_j) = p[\mathbf{B}(t_{j+1}) - \mathbf{B}(t_j)]$, $D^2\mathbf{B}(t_j) = p^2[\mathbf{B}(t_{j+2}) - 2\mathbf{B}(t_{j+1}) + \mathbf{B}(t_j)]$,.... With p large and $e(t)$ small, we have $\beta(t) \approx \mathbf{B}(t)^T\eta$, hence:

$$\begin{aligned} \beta^{(d)}(t) &\approx \mathbf{B}^{(d)}(t)^T\eta \\ &\approx D^d\mathbf{B}(t)^T\eta. \end{aligned}$$

As a consequence, with:

$$A_{(d)} = \begin{pmatrix} D^d\mathbf{B}(t_1)^T \\ \vdots \\ D^d\mathbf{B}(t_{p-d})^T \end{pmatrix} \quad \text{of dimensions} \quad (p-d) \times p,$$

we can define $\gamma_{(d)} = A_{(d)}\eta$ of length $(p-d)$ which is an approximation of $\beta^{(d)}(.)$ on the grid $\{0, \frac{1}{p}, \frac{2}{p}, \ldots, \frac{p-d}{p}\}$. Indeed, ${\gamma_{(d)}}_j = D^d\mathbf{B}(t_j)^T\eta$. In a first approach we are interested by having the derivatives of order 0 and 2 sparse. Hence we will use $A_{(0)} = I_{p\times p}$ and $A_{(2)}$ (note that$A_{(2)}$ is not invertible because it is not a square matrix). The approximation of the derivatives on the grid can be expressed as:

$$\gamma_{(0)} = A_{(0)}\eta \qquad \text{and} \qquad \gamma_{(2)} = A_{(2)}\eta, \tag{4}$$

and we have a relation between $\gamma_{(0)}$ and $\gamma_{(2)}$ which is:

$$\gamma_{(2)} = A_{(2)}A_{(0)}^{-1}\gamma_{(0)} = A_{(2)}\gamma_{(0)}. \tag{5}$$

Combining (3) and (4) we can re-write the model using the approximation of the zero-derivative:

$$\begin{aligned} Y_i &= \beta_0 + \mathbf{X}_i^T\gamma_{(0)} + \epsilon_i^*, \\ Y &= \beta_0\mathbb{1} + \mathbf{X}\gamma_{(0)} + \epsilon^*, \end{aligned} \tag{6}$$

with $Y = [Y_1, \ldots, Y_n]^T$, $\epsilon^* = [\epsilon_1^*, \ldots, \epsilon_n^*]^T$.

2.2 Latent variables to model the sparsity of the derivatives

We first introduce a latent vector to model the sparsity of the zero-derivative: $\theta_{(0)}$ of length p defined by

$${\theta_{(0)}}_j = \begin{cases} 1 & \text{if } {\gamma_{(0)}}_j \neq 0, \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, p. \tag{7}$$

Given $\theta_{(0)}$, $\gamma_{\theta_{(0)}}$ is the vector of all nonzero elements of $\gamma_{(0)}$, and $\mathbf{X}_{\theta_{(0)}}$ is the matrix $\mathbf{X}$ with only the columns corresponding to the non-nul components of $\theta_{(0)}$. The model (6) can be re-written as:

$$Y = \beta_0\mathbb{1} + \mathbf{X}_{\theta_{(0)}}\gamma_{\theta_{(0)}} + \epsilon^*. \tag{8}$$

Using this modelisation, we can estimate where are the null regions of $\beta(.)$ by estimating $\theta_{(0)}$. Indeed, the vector $\theta_{(0)}$ is null when the approximated zero-derivative $\gamma_{(0)}$ is null, that is when $\beta(.)$ itself is null. Inspired by variable selection approach, the idea is to select the points of the grid where $\gamma_{(0)}$ is non null, to obtain regions where $\beta(.)$ is assumed non null.

On this regions the objective is to obtain a piecewise linear structure for $\beta(.)$, that is an approximated second-derivative $\gamma_{(2)}$ which is sparse. Inspired by variable selection approach again, the idea is to select the points of the grid where $\gamma_{(2)}$ is non null, by estimating a latent vector similar to $\theta_{(0)}$, but associated with $\gamma_{(2)}$.

However, we can not obtain a model as (8) as we do not have a model similar to (6) containing $\gamma_{(2)}$. Indeed, the initial model (3) contains η and we can not express η using $\gamma_{(2)}$ as $A_{(2)}$ is not invertible. To circumvent this problem we define $\widetilde{\gamma}_{(2)}$ of length p and $\widetilde{A}_{(2)}$ of dimensions $p \times p$ as:

$$\widetilde{\gamma}_{(2)} = \begin{pmatrix} \\ \gamma_{(2)} \\ \\ {\gamma_{(0)}}_{p-1} \\ {\gamma_{(0)}}_{p} \end{pmatrix} \qquad \text{and} \qquad \widetilde{A}_{(2)} = \begin{pmatrix} & & A_{(2)} & & \\ & & & & \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix} \tag{9}$$

We obtain the following equation which is similar to (5):

$$\widetilde{\gamma}_{(2)} = \widetilde{A}_{(2)}\gamma_{(0)}. \tag{10}$$

Here $\widetilde{A}_{(2)}$ is invertible, hence we have $\gamma_{(0)} = \widetilde{A}_{(2)}^{-1}\widetilde{\gamma}_{(2)}$. We can then obtain a model similar to (6) but containing $\gamma_{(2)}$:

$$\begin{aligned} Y &= \beta_0 \mathbb{1} + \mathbf{X}\widetilde{A}_{(2)}^{-1}\widetilde{\gamma}_{(2)} + \epsilon^* \\ &= \beta_0 \mathbb{1} + \mathbf{Z}\widetilde{\gamma}_{(2)} + \epsilon^*, \end{aligned} \tag{11}$$

Then with $\theta_{(2)}$ of length p defined by:

$$\begin{aligned} {\theta_{(2)}}_j &= \begin{cases} 1 & \text{if } {\gamma_{(2)}}_j \neq 0, \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \dots, p-2, \\ {\theta_{(2)}}_{p-1} &= {\theta_{(2)}}_p = 0 \qquad \text{by convention,} \end{aligned} \tag{12}$$

we can obtain a model similar to (8) but containing $\theta_{(2)}$:

$$Y = \beta_0 \mathbb{1} + \mathbf{Z}_{\theta_{(2)}}\widetilde{\gamma}_{\theta_{(2)}} + \epsilon^*. \tag{13}$$

These latent vectors $\theta_{(0)}$ and $\theta_{(2)}$ can be easily interpreted. The functional coefficient $\beta(.)$ is assumed to be null on some regions of the grid, and piecewise linear elsewhere. The vector $\theta_{(0)}$ contains information about where $\beta(.)$ is null, while $\theta_{(2)}$ contains informations about the shape of $\beta(.)$: it is null on regions where $\beta(.)$ is linear, and non null just before a breakpoint (if there is a breakpoint of $\beta(.)$ at the point $\frac{k}{p}$ of the grid, $\theta_{(2)}$ is non null at the point $\frac{k-1}{p}$). By convention $\theta_{(2)}$ is null on the last two points of the grid, because we can not calculate $\gamma_{(2)}$ on these points (see (4) or (5)).

2.3 Prior distributions

To work in a Bayesian framework, we need to put prior distributions on the unknow parameters . These priors will be detailed in the presentation.

3 Implementation and inference

The vectors of interest are $\theta_{(0)}$ and $\theta_{(2)}$ as they enable us to model the sparsity of $\beta(.)$ and of its second derivative. The use of these latent vectors is inspired from Stochastic Search Variable selection (SSVS) procedures, which have been introduced by George and McCulloch [1993] and Chipman et al. [2001]. The inference scheme we proposed is then inspired from these procedures.

3.1 Estimation of $\theta_{(0)}$

We start from the complete posterior distribution $\pi(\beta_0, \gamma_{\theta_{(0)}}, \theta_{(0)}, \sigma^2 \mid Y)$ obtained from (8). The number of possible $\theta_{(0)}$-vectors is 2^p, which is quite large for grids commonly used. The idea is to use a Gibbs sampling algorithm to explore this posterior distribution and search for high probability $\theta_{(0)}$ values, without having to visit all the possible values. But in a first step only $\theta_{(0)}$ is of interest, the other parameters $\beta_0, \gamma_{\theta_{(0)}}$ and σ^2 can be considered as nuisance parameters. We can then use the collapsing technique studied by Liu [1994] and Liu et al. [1994], by integrating out these parameters from the complete posterior distribution. This technique improves the algorithm and facilitates the convergence of the Markov chain. Integrating out $\beta_0, \gamma_{\theta_{(0)}}$ and σ^2 we obtain $\pi(\theta_{(0)} \mid Y)$. A simple Metropolis-Hastings algorithm (Hastings [1970] and Metropolis et al. [1953) can then enable us to explore the posterior distribution of the vector of interest $\theta_{(0)}$. Using the post-burn-in simulations, we can then estimate it.

3.2 Estimation of $\theta_{(2)}$

Once we have an estimate $\widehat{\theta_{(0)}}$ of $\theta_{(0)}$, we estimate $\theta_{(2)}$. We have to keep in mind that $\gamma_{(0)}$ and $\gamma_{(2)}$ are related by (5) and (10). As a consequence, $\theta_{(0)}$ and $\theta_{(2)}$ are also related. For instance, if $\theta_{(0)}$ is null on a region of the grid ($\beta(.)$ is null), $\theta_{(2)}$ should also be null on a corresponding region (the second derivative of $\beta(.)$ should be null). In practice, once we have an estimate of $\theta_{(0)}$, some constraints on the estimate of $\theta_{(2)}$ are necessary to ensure that the relations (5) and (10) can be verified.
Once the constraints have been defined, the estimation of $\theta_{(2)}$ is done in a similar way to $\theta_{(0)}$. The complete posterior distribution $\pi(\beta_0, \widetilde{\gamma}_{\theta_{(2)}}, \theta_{(2)}, \sigma^2 \mid Y)$ obtained from (13) is integrated out in $\beta_0, \widetilde{\gamma}_{\theta_{(2)}}$ and σ^2, which yields $\pi(\theta_{(2)} \mid Y)$. Here again a Metropolis-Hastings algorithm enables us to explore the posterior distribution of the vector of interest $\theta_{(2)}$, and the the post-burn-in simulations enable us to estimate it.

3.3 Estimation of $\beta(.)$ on the grid

Once we have estimates for $\theta_{(0)}$ and $\theta_{(2)}$, we can easily obtain an estimate for $\beta(.)$. Indeed, we have estimated non-nul regions and positions of breakpoints for $\beta(.)$. As $\beta(.)$ is assumed to be piecewise linear, we just need to estimate the non-nul values of $\beta(.)$ on the breakpoints. The values of $\beta(.)$ between two breakpoints can then be easily deduced using the linearity. From the model (6) we can obtain the complete posterior distribution $\pi(\beta_0, \gamma_{(0)}, \sigma^2 \mid Y)$. Having integrating out β_0 and σ^2, we obtain the posterior of $\gamma_{(0)}$, $\pi(\gamma_{(0)} \mid Y)$. To explore this posterior distribution of interest we still use a Metropolis-Hastings algorithm. To estimate $\gamma_{(0)}$ and hence $\beta(.)$ on the grid from the post-burn-in simulations $\{\gamma_{(0)}^{(t)}, t = b+1, \ldots, b+m\}$, we classicaly use the empirical mean which is a consistent estimator of the posterior expectation $\mathbb{E}(\gamma_{(0)} \mid Y)$.

4 Results and Discussion

During the presentation we will we show results of our approach on simulations and on a real example, and we will compare it to the standard approach implemented in the FDA R-package and to the FLiRTI method. Eventually, we will discuss this approach.

Bibliography

[1] H. Chipman, E.I. George, and R.E. McCulloch (2001). *The practical implementation of Bayesian model selection.* In Model selection - IMS Lecture Notes. P. LAHIRI. Institute of Mathematical Statistics.

[2] C. Crainiceanu, A-M. Staicu, and C-Z. Di (2009). *Generalized multilevel functional regression.* Journal of the American Statistical Association, 104(488):1550-1561.

[3] C. Crainiceanu and J. Goldsmith (2010). *Bayesian functional data analysis using winbugs.* Journal of statistical software, 32(11).

[4] E.I. George and R.E. McCulloch (1993). *Variable selection via Gibbs sampling.* Journal of the American Statistical Association, 88(423):881-889.

[5] G.M. James, J. Wang, and J. Zhu (2009). *Functional linear regression that is interpretable.* The Annals of Statistics, 37:2083-2108.

[6] J.S. Liu (1994). *The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem.* Journal of the American Statistical Association, 89(427):958-966.

[7] J.S. Liu, W.H. Wong, and A. Kong (1994). *Covariance structure and convergence rate of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes.* Biometrika, 81:27-40.

[8] J.O. Ramsay and B.W. Silverman (2006). *Functional Data Analysis.* Springer-Verlag, New-York.