Bayesian functional linear regression estimation. Extension to scalar and categorical covariates

Feriel Bouhadjera [1], Meïli Baragatti [1], Nadine Hilgert [1], Nathalie Smits [2] & Paul-Marie Grollemund [3]

[1] *MISTEA, Université Montpellier, INRAE, Institut Agro, Montpellier, France*
[2] *ABSys, Université Montpellier, INRAE, Institut Agro, Montpellier, France*
[3] *LMBP, Université Clermont Auvergne, Aubière, France*

Résumé. Nous considérons un modèle de régression linéaire où la variable à expliquer est réelle et les co-variables sont catégorielles, scalaires et fonctionnelles et agissent de manière additive dans le modèle. Notre objectif est d'estimer les paramètres de ce modèle, tout en conservant un caractère interprétable pour la partie du modèle incluant les variables fonctionnelles. Ce travail est une extension du modèle Bliss (*Bayesian functional Linear regression with Sparse Step function*), voir Grollemund et al. (2019), développé dans un cadre Bayésien et qui ne contient que des variables fonctionnelles. Nous proposons d'étendre la méthode Bliss à un modèle plus général contenant également des co-variables catégorielles et scalaires. Dans la suite, nous explicitons le modèle étendu et expliquons comment estimer les paramètres d'une manière interprétable. Une illustration est faite sur des données simulées et un jeu de données réelles sur le dépérissement de la vigne.

Mots-clés. Données fonctionnelles, méthode Bliss, régression linéaire, statistique Bayésienne.

Abstract. We consider a linear regression model with a scalar response variable and categorical, scalar and functional covariates, that act additively in the model. Our objective is to estimate the parameters of this model, while maintaining an interpretability for the part of the model including the functional covariates. This work is an extension of the Bliss model (*Bayesian functional Linear regression with Sparse Step function*), see Grollemund et al. (2019), developed in a Bayesian framework with only functional covariates. We propose to extend the Bliss method to a more general model that also contains categorical and scalar covariates. In the following, we explain the extended model and how to estimate the parameters in an interpretable way. An illustration is made on simulated data and a real data set obtained in the field of vine dieback.

Keywords. Bayesian statistics, Bliss method, functional data, linear regression.

1 Introduction

Technological advances allow us to collect large amount of data in many areas. In agronomy, a major challenge is to extract information from these massive data and establish links with final characteristics such as production yields or quality. The motivation of our study comes from a project on the vine dieback. A unique database obtained by the *Bureau National Interprofessionnel du Cognac* on the monitoring of 55 plots since 1977 offers the opportunity to analyse the determinants of multi-year vine dieback trajectories. In this project, the objectives are to be able to identify (1) the factors and interactions of biotic, abiotic and technical factors that contribute to the decline in plot yield and to the mortality of individual grapevines and (2) the time period over which these factors have an impact, both in the short term at the scale of the crop cycle and in the long term since the plot was planted. One way of modeling this problem is to build a regression model to explain the effect of climatic conditions and cultivation managements on the vine yields. The covariates involved in such a study are of different types : functional, categorical or scalar. An 'interpretable' linear regression model with functional covariates already exists in the Bayesian framework : the Bliss model, see Grollemund et al. (2019). In this model, the *prior* distribution assumes that the coefficient functions are step functions. Bliss allows to recover periods of time that influence the most the outcome, which helps answering objective (2) mentioned above. To take into account scalar and categorical covariates as well, we have to extend the Bliss model as follows.

From a statistical point of view, consider a p-dimensional scalar covariate vector $(z_1, \cdots, z_p)$, q functional covariates $(x_1(t), \cdots, x_q(t))$ on $\mathcal{C}$, which is an interval of $\mathbb{R}$, a categorical covariate α with r levels, and a scalar variable to be explained y. The statistical linear model under consideration is as follows, for a n-sample of individuals doubly indexed by (ij), where $i = 1, \cdots, r$ stands for the categorical levels and $j = 1, \cdots, n_i$ stands for the repetitions :

$$y^{(ij)} = \beta_0 + \alpha_i + \sum_{s=1}^{p} \gamma_s z_s^{(ij)} + \sum_{v=1}^{q} \int_{\mathcal{C}} \beta_v(t) x_v^{(ij)}(t) dt + \epsilon^{(ij)}, \tag{1}$$

where β_0 is the intercept, $(\gamma_1, \cdots, \gamma_p)$ and $(\alpha_1, \cdots, \alpha_r)$ are scalar parameters, and $(\beta_1(t), \cdots, \beta_q(t))$ are functional parameters. Finally, $\epsilon^{(ij)}$ are random variables. Note that this model could also be extended to 2 or more categorical covariates.

The problem we are interested in is to explain the response variable from all these covariates of different type. Statistical models that involve functional covariates with scalar and categorical covariates are not common.

Functional data models have been popularized by works such as Ramsey and Silverman (1997) and Ferraty and Vieu (2006). There are different derivatives of this model in the literature. Not to be exhaustive, we quote the generalized functional regression model (see, Müller et Stadtmüller (2005)) and the functional regression model with a functional

response (see, Yao et al. (2005)). It is known that the correlation between the covariates decreases the estimation quality. In the functional part of our model, there are obviously large correlations. To overcome this problem, robust estimators based on the minimization of a penalized least squares criterion such as Ridge or Lasso regression exist in the literature. However, despite the robustness of the proposed estimators, it is difficult to interpret the estimate.
From a practical point of view, it is important to estimate periods of time during which the climatic conditions have influenced the variable of interest. For a fixed v, if the coefficient function $\beta_v(t)$ is null over a period $\mathcal{C}$ then the integral $\int_{\mathcal{C}} \beta_v(t)x_v(t)dt$ can be written without using the values of x_v over $\mathcal{C}$ and if it is positive (resp. negative) over a period then the increase (resp. decrease) of the value of x_v over this period induces an increase (resp. decrease) of the value of y. It is important to highlight the periods where $\beta_v(t)$ is positive or negative. That's why we are interested in the estimation of the support of the coefficient function, which is an important feature of the interpretation. To that aim, the Bliss method focuses on the estimation of the support of the functional parameter, which makes the interpretation easier. The functional parameter is identified through a piecewise constant function appropriately chosen. The complex functional linear regression model is then simplified to a linear regression model. The Bliss model is estimated in a Bayesian framework which allows to take into account expert knowledge.
To fit linear regression models, several Bayesian approaches have been considered, and without being exhaustive we can list Wang et al. (2007) and Goldsmith et al. (2011). To our knowledge, the only paper specifically interested in the support of the coefficient function using a Bayesian approach is Grollemund et al. (2019). Moreover, in the last paper, the authors propose two estimators of the coefficient function with different properties. Our contribution is an extension of the Bliss method to other types of variables : scalar and categorical.

2 Model

We consider the functional parameters β_v as piecewise constant functions that can be described with a minimal number of intervals K_v :

$$\beta_v(t) = \sum_{k=1}^{K_v} \frac{b_{k,v}}{|\mathcal{I}_{k,v}|} \mathbb{1}_{\{t \in \mathcal{I}_{k,v}\}}, \qquad \text{for } v = 1, \cdots, q, \tag{2}$$

where $\mathcal{I}_{1,v}, \cdots, \mathcal{I}_{K_v,v}$ are intervals included in $\mathcal{C}$, $|\mathcal{I}_{k,v}|$ is the length of the interval and $b_{1,v}, \cdots, b_{K_v,v}$ are real parameters. The support is the union of all $\mathcal{I}_{k,v}$ if the coefficient b_k are not null. Thus, a period of time which does not influence the outcome will be outside the support. Replacing (2) in (1), the functional linear model becomes a linear model with

design depending on the intervals $\mathcal{I}_{k,v}$:

$$y^{(ij)} = \beta_0 + \alpha_i + \sum_{s=1}^{p} \gamma_s z_s^{(ij)} + \sum_{v=1}^{q} \sum_{k=1}^{K_v} b_{k,v} x_v^{(ij)}(\mathcal{I}_{k,v}) + \epsilon^{(ij)}, \tag{3}$$

with

$$x_v^{(ij)}(\mathcal{I}_{k,v}) = \frac{1}{|\mathcal{I}_{k,v}|} \int_{\mathcal{I}_{k,v}} x_v^{(ij)}(t)dt, \quad \text{for} \quad \mathcal{I}_{k,v} = [m_{1,v} \pm \ell_{1,v}, \cdots, m_{K_v,v} \pm \ell_{K_v,v}],$$

where for a fixed v, $m_{k,v}$ is the center and $\ell_{k,v}$ is the half length of the interval $\mathcal{I}_{k,v}$. Hence, let define the parameters set to model (1) of dimension $3K + r + p + 2$ by :

$$\theta \quad = \quad \left(\beta_0, (\alpha_1, \cdots, \alpha_r), (\gamma_1, \cdots, \gamma_p), (\theta_1, \cdots, \theta_q), \sigma^2\right),$$

where

$$\theta_v \quad = \quad ((b_{1,v}, \cdots, b_{K_v,v}), (m_{1,v}, \cdots, m_{K_v,v}), (\ell_{1,v}, \cdots, \ell_{K_v,v})).$$

For simplicity of notations, we focus on model (1) with only one scalar and one functional covariates. We assume that we observe $\{y^{(ij)},\ i = 1, \cdots, r \text{ and } j = 1, \cdots, n_i\}$ replicates of outcome, $(\alpha_1, \cdots, \alpha_r)$ levels of the categorical variable α and $\{z^{(ij)},\ i = 1, \cdots, r \text{ and } j = 1, \cdots, n_i\}$ are scalar observations and $\{x^{(ij)}(\mathcal{I}_k),\ i = 1, \cdots, r,\ j = 1, \cdots, n_i \text{ and } k = 1, \cdots, K\}$ are scalar observations. Notice that $n = \sum_{i=1}^{r} n_i$. The model (3) becomes :

$$y^{(ij)} | \beta_0, \alpha_i, \gamma, z^{(ij)}, (b_1, \cdots, b_K), x^{(ij)}(t) \rightsquigarrow \mathcal{N}\left(\beta_0 + \alpha_i + \gamma z^{(ij)} + \sum_{k=1}^{K} b_k x^{(ij)}(\mathcal{I}_k), \sigma^2\right). \tag{4}$$

We complete the Bayesian model (4) with the following *prior* distributions :

$$\beta_0 | \sigma^2 \rightsquigarrow \mathcal{N}\left(0, u_0\sigma^2\right), \qquad \alpha_i | \sigma^2 \rightsquigarrow \mathcal{N}\left(0, v_0\sigma^2\right), \qquad \gamma | \sigma^2 \rightsquigarrow \mathcal{N}\left(0, w_0\sigma^2\right),$$

and $\pi(\sigma^2) \propto 1/\sigma^2$. Furthermore, for $k = 1, \cdots, K$

$$\begin{aligned} b_k | \sigma^2, m_k, \ell_k &\rightsquigarrow \mathcal{N}_K\left(0, n\sigma^2(G + \nu\lambda_{max}(G) I_K)^{-1}\right), \\ m_k &\overset{i.i.d.}{\rightsquigarrow} \mathcal{U}(\mathcal{C}), \\ \ell_k &\overset{i.i.d.}{\rightsquigarrow} \exp\left(a \times |\mathcal{C}|\right), \end{aligned} \tag{5}$$

where G is the Gram matrix given by $G = x.(\mathcal{I}.)^t x.(\mathcal{I}.)$ with

$$x.(\mathcal{I}.) = \left\{x^{(ij)}(\mathcal{I}_k),\ \ i = 1, \cdots, r,\ j = 1, \cdots, n_i, \text{ and } k = 1, \cdots, K\right\}.$$

3 Implementation

The full posterior distribution can be written explicitly from the Bayesian model given in (5) by

$$\begin{aligned}
\beta|y,\sigma^2,m,\ell &\rightsquigarrow \mathcal{N}_K\left((\underline{x}^t\underline{x}+\underline{V})^{-1}\underline{x}^t y, \sigma^2(\underline{x}^t\underline{x}+\underline{V})^{-1}\right),\\
\sigma^2|y,\beta,m,\ell &\rightsquigarrow \Gamma^{-1}\left(\frac{n+K+r}{2}+1, \frac{1}{2}\left\{\text{RSS}+\beta^t\underline{V}\beta\right\}\right),\\
f(m_k|y,\beta,\sigma^2,m_{-k},\ell) &\propto \exp\left\{-\text{RSS}/2\sigma^2\right\}\times f(b|m_k,\ell_k,\sigma^2)\times f(m_k), \qquad k=1,\cdots,K\\
f(\ell_k|y,\beta,\sigma^2,\ell_{-k},m) &\propto \exp\left\{-\text{RSS}/2\sigma^2\right\}\times f(b|m_k,\ell_k,\sigma^2)\times f(\ell_k), \qquad k=1,\cdots,K
\end{aligned}$$

where $\underline{x} = (1_n \mid A \mid z \mid x.(\mathcal{I}_\cdot))$, $\beta^t = (\beta_0,(\alpha_1,\cdots,\alpha_r),\gamma,(b_1,\cdots,b_K))^t$ and $\text{RSS} = ||y-\underline{x}\beta||^2$. Furthermore, let define

$$A = \begin{pmatrix}
1_{n_1} & 0_{n_1} & \dots & \dots & \dots & \dots & 0_{n_1}\\
0_{n_2} & 1_{n_2} & 0_{n_2} & \dots & \dots & \dots & 0_{n_2}\\
\vdots & & & \dots & & & \vdots\\
0_{n_i} & \dots & 0_{n_i} & 1_{n_i} & 0_{n_i} & \dots & 0_{n_i}\\
\vdots & & & & \vdots & & \\
0_{n_r} & \dots & \dots & \dots & \dots & 0_{n_r} & 1_{n_r}
\end{pmatrix} \qquad \underset{-}{V} = \begin{pmatrix}
v_0^{-1} & 0 & 0 & 0\\
0 & u_0^{-1} & 0 & 0\\
0 & 0 & w_0^{-1} & 0\\
0 & 0 & 0 & n^{-1}(G+\nu\lambda_{max}(G)\mathcal{I}_K)
\end{pmatrix}.$$

The full *posterior* conditional distributions of parameters m_k and ℓ_k are not know distributions. However, considering that m_k and ℓ_k evolve in a finite grid of their supports, it is possible to characterize these distributions by numerically computing their probability functions. Furthermore, the resulting Bayesian model depends on hyperparameters which are u_0, v_0, w_0, ν, a and K. It is possible to choose these hyperparameters so that the *prior* distributions are the less informative possible. For the parameter K, we choose it rather large which guarantees a low error insofar as the intervals can overlap. In Grollemund et al. (2019), the authors propose some default values for the hyperparameters. As usual with hierarchical models, sampling from the posterior distribution can be done with a Gibbs algorithm (see : Robert and Casella (2013), Chapter 7).

4 Applications

Numerical results based both on simulated data and real data from the vineyard will be presented.

Acknowledgments

The present work was supported by the LabEx NUMEV (ANR-10-LABX-0020) within the I-Site MUSE, and by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

Bibliography

Grollemund, P.M., Abraham, C., Baragatti, M., Pudlo, P. (2019). Bayesian Functional Linear Regression with Sparse Step Functions, *Bayesian Analysis*, 14(1) :111–135.
Ramsay, J. and Silverman, B. (1997). Functional Data Analysis. *Springer-Verlag New-York*.
Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis : theory and practice. *Springer Science & Business Media*.
Müller, H.-G. et Stadtm̈uller, U. (2005). Generalized functional linear models. *Annals of Statistics*, pages 774–805.
Yao, F., Müller, H.-G., Wang, J.-L., et al. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6) :2873–2903.
Wang, X., Ray, S., and Mallick, B. K. (2007). Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, 102(479) :962–973.
Goldsmith, J., Wand, M. P., and Crainiceanu, C. (2011). Functional regression via variational Bayes. *Electronic journal of statistics*, 5 :572–602.
Robert, C.P. and Casella, G. (2013). Monte Carlo statistical methods. Springer-Verlag New York. 22, 72.