A functional spatial autoregressive model using signatures

Camille Frévent [1]

[1] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.* camille.frevent@univ-lille.fr

Résumé. Nous proposons une nouvelle approche au modèle spatial autoregressif avec covariables fonctionnelles, basé sur la notion de signatures. Celles-ci représentent une fonction comme une série de ses intégrales itérées. Elles présentent l'avantage d'être applicables à un large éventail de processus. Après avoir fourni des garanties théoriques au modèle proposé, nous avons montré dans une étude de simulation que cette nouvelle approche présente des performances compétitives par rapport au modèle traditionnel.

Mots-clés. Données fonctionnelles, FSAR, Régression spatiale, Signature, Tenseur

Abstract. We propose a new approach to the autoregressive spatial functional model, based on the notion of signature, which represents a function as an infinite series of its iterated integrals. It presents the advantage of being applicable to a wide range of processes.
After having provided theoretical guarantees to the proposed model, we have shown in a simulation study that this new approach presents competitive performances compared to the traditional model.

Keywords. Functional data, FSAR, Signature, Spatial regression, Tensor

1 Introduction

We are interested here in modelling the relationship between a real-valued random variable Y and a functional covariate $\{X(t), t \in \mathcal{T}\}$ observed in N spatial locations. A traditional approach is to assume that X belongs to $\mathcal{L}^2(\mathcal{T})$, the space of square-integrable functions on $\mathcal{T}$, and to consider the following model:

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \int_{\mathcal{T}} X_i(t)\theta^*(t)\ \mathrm{d}t + \varepsilon_i, \qquad i = 1, \dots, N,\ \ N = 1, 2, \dots$$

where the spatial dependency structure between the spatial units is described by the spatial weights matrix $V_N = (v_{ij,N})_{1 \leq i,j \leq N}$, the autoregressive parameter ρ^* is in a compact space $\mathcal{R}$ and $\theta^* \in \mathcal{L}^2(\mathcal{T})$.

Fermanian (2022) recently investigated the use of signatures in the context of a non-spatial linear regression model with functional covariates. Signatures present the advantages of being applicable to a wide range of processes that are not necessary square-integrable processes.

2 The signatures-based spatial autoregressive model

2.1 Concept of signatures

Let $\mathcal{T}$ be a compact interval and $X : \mathcal{T} \to \mathbb{R}^p$ be a $p-$dimensional continuous function, $p \geq 2$. Let $(e_i)_{i=1}^p$ be the canonical orthonormal basis of $\mathbb{R}^p$. Then the signature of X can be written as

$$Sig(X) = 1 + \sum_{d=1}^{\infty} \sum_{(i_1,\dots,i_d)} \mathcal{S}_{(i_1,\dots,i_d)}(X) e_{i_1} \otimes \cdots \otimes e_{i_d}. \text{ where } \mathcal{S}_{(i_1,\dots,i_d)}(X) = \int_{\substack{t_1 < \cdots < t_d \\ t_1,\dots,t_d \in \mathcal{T}}} \cdots \int \mathrm{d}X^{(i_1)}(t_1) \dots \mathrm{d}X^{(i_d)}(t_d).$$

2.2 Model

We consider the following signatures-based spatial autoregressive model:

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \alpha^* + \langle \theta^*, Sig(X_i) \rangle + \varepsilon_i \tag{1}$$

where the parameter θ^* is assumed to be written as

$$\theta^* = 1 + \sum_{d=1}^{\infty} \sum_{(i_1,\dots,i_d)} \beta^*_{(i_1,\dots,i_d)} e_{i_1} \otimes \cdots \otimes e_{i_d}.$$

The disturbances ε_i are assumed to be independent and identically distributed random variables such that $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^{2*}$. They are also independent of X.

Then, one can rewrite (1) as

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \alpha^* + 1 + \sum_{d=1}^{\infty} \sum_{(i_1,\dots,i_d)} \beta^*_{(i_1,\dots,i_d)} \mathcal{S}_{(i_1,\dots,i_d)}(X_i) + \varepsilon_i. \tag{2}$$

However, this model cannot be maximized without addressing the difficulty produced by the infinite dimension of the signatures $Sig(X_i)$ (and thus the infinite number of coefficients $\beta^*_{(i_1,\dots,i_d)}$).
Thus we proposed two estimation methods that overcome this challenge.

2.2.1 Penalized signatures-based spatial regression

We consider here a slightly modified version of Model (2). We assume that the signature coefficients $\mathcal{S}_{(i_1,\dots,i_d)}(X_i)$ are involved in the model only up to a certain unknown truncation order $D^* \in \mathbb{N}$. The model thus becomes

$$Y_i = \rho^* \sum_{j=1}^{N} v_{ij,N} Y_j + \alpha^* + 1 + \sum_{d=1}^{D^*} \sum_{(i_1,\dots,i_d)} \beta^*_{(i_1,\dots,i_d)} \mathcal{S}_{(i_1,\dots,i_d)}(X_i) + \varepsilon_i. \tag{3}$$

Let the signature coefficients vector of X, $\mathcal{S}(X)$, be the sequence of all signature coefficients:

$$\mathcal{S}(X) = (1, \mathcal{S}_{(1)}(X), \dots, \mathcal{S}_{(p)}(X), \mathcal{S}_{(1,1)}(X), \mathcal{S}_{(1,2)}(X), \dots, \mathcal{S}_{(i_1,\dots,i_d)}(X), \dots)$$

and the truncated signature coefficients vector at order D of X, $\mathcal{S}^D(X)$, be defined as

$$\mathcal{S}^D(X) = (1, \mathcal{S}_{(1)}(X), \mathcal{S}_{(2)}(X), \dots, \mathcal{S}_{\underbrace{(p,\dots,p)}_{D \text{ terms}}}(X)).$$

Then, by noting $s_p(D) = \sum_{d=0}^{D} p^d = \frac{p^{D+1}-1}{p-1}$ the dimension of the truncated signature coefficients vector at order D, (3) can be rewritten as

$$S_N \mathbf{Y}_N = \alpha^* \mathbf{1}_N + \xi_{N,D^*} B^* + \boldsymbol{\varepsilon}_N, \tag{4}$$

where $S_N = (I_N - \rho^* V_N)$, $\mathbf{Y}_N$ and $\boldsymbol{\varepsilon}_N$ are two $N \times 1$ vectors of elements Y_i and ε_i, $i = 1, \dots, N$ respectively, I_N denotes the $N \times N$ identity matrix and $\mathbf{1}_N$ denotes the $N \times 1$ vector composed only of 1.

$B_D = (1, \beta_1, \beta_2, \dots, \beta_{\underbrace{(p,\dots,p)}_{D \text{ terms}}})^\top \in \mathbb{R}^{s_p(D)}$, $B^* = (1, \beta^*_1, \beta^*_2, \dots, \beta^*_{\underbrace{(p,\dots,p)}_{D^* \text{ terms}}})^\top \in \mathbb{R}^{s_p(D^*)}$ and $\xi_{N,D}$ is an $N \times s_p(D)$ matrix whose i^{th} line corresponds to $\mathcal{S}^D(X_i)$.

For a truncation order D, the associated conditional quasi log-likelihood function is

$$\begin{aligned} \ell_N(\sigma^2, \rho, \alpha, B_D) &= -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) + \ln|S_N(\rho)| \\ &\quad -\frac{1}{2\sigma^2}\left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,D}B_D\right]^\top \left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,D}B_D\right] \end{aligned} \tag{5}$$

where $S_N(\rho) = I_N - \rho V_N$.
Then, Model (4) is estimated using a penalized (ridge) regression.

2.2.2 Spatial autoregressive model based on signatures projections

In this section we consider Model (2). Using the notation $B^*_\infty = (1, \beta^*_1, \dots, \beta^*_{(i_1,\dots,i_d)}, \dots)^\top$, and $\xi_{N,\infty}$ the matrix whose i^{th} line corresponds to $\mathcal{S}(X_i)$, one can rewrite (2) as

$$S_N \mathbf{Y}_N = \alpha^* \mathbf{1}_N + \xi_{N,\infty} B^*_\infty + \boldsymbol{\varepsilon}_N\,, \qquad N = 1, 2, \dots$$

Then, the associated conditional quasi log-likelihood function of the vector $\mathbf{Y}_N$ given $\{Sig(X_i),\ i = 1, \ldots, N\}$ is given by:

$$\ell_N(\sigma^2, \rho, \alpha, B_\infty) = -\frac{N}{2}\ln(\sigma^2) - \frac{N}{2}\ln(2\pi) + \ln|S_N(\rho)| - \frac{1}{2\sigma^2}\left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,\infty}B_\infty\right]^\top \left[S_N(\rho)\mathbf{Y}_N - \alpha\mathbf{1}_N - \xi_{N,\infty}B_\infty\right]. \quad (6)$$

Estimation
We assume that there exist new coefficients $\boldsymbol{\zeta}_i = (\zeta_{i,1}, \zeta_{i,2}, \ldots)^\top$ and $\Phi^* = (\phi_1^*, \phi_2^*, \ldots)^\top$ such that

$$\langle \theta^*, Sig(X_i)\rangle = \mathcal{S}(X_i)B_\infty^* = 1 + \langle \boldsymbol{\zeta}_i, \Phi^*\rangle,$$

and a positive sequence of integers C_N increasing asymptotically as the sample size $N \to \infty$ such that

$$\langle \boldsymbol{\zeta}_i, \Phi^*\rangle = \sum_{c=1}^{C_N} \zeta_{i,c}\phi_c^* + \sum_{c=C_N+1}^{\infty} \zeta_{i,c}\phi_c^*$$

where the second term vanishes asymptotically when $N \to \infty$.

Then $\langle \theta^*, Sig(X_i)\rangle$ can be approximated by $1+\sum_{c=1}^{C_N} \zeta_{i,c}\phi_c^*$ and $\xi_{N,\infty}B_\infty$ can be approximated by $\mathbf{1}_N + Z_{C_N}\Phi_{C_N}^*$ where Z_{C_N} is the $N \times C_N$ matrix whose i^{th} line is given by

$$\boldsymbol{\zeta}_i^{C_N\top} = (\zeta_{i,1}, \ldots, \zeta_{i,C_N})$$

and $\Phi_{C_N}^* = (\phi_1^*, \ldots, \phi_{C_N}^*)^\top$.

Then the new parameters can be estimated using a traditional estimation approach for a (non functional) SAR model.

3 Finite sample properties

A simulation study was then conducted to compare the performances of the proposed signatures-based spatial autoregressive model considering the penalized spatial regression and the signatures projections strategies. We also compared them with the functional linear model proposed by Ahmed et al (2022).

3.1 Design of the simulation study

We considered a grid with 60×60 locations, where we randomly allocate N=200 spatial units. Then the data was generated according to the following three models where

$$X_i(t) = (X_{i,1}(t), \ldots, X_{i,p}(t))^\top, X_{i,k}(t) = \alpha_{i,k}t + f_{i,k}(t),$$

$$\theta^*(t) = (\theta_1^*(t), \ldots, \theta_p^*(t))^\top, \theta_k^*(t) = \Psi_k t + g_{i,k}(t),$$
$$\alpha_{i,k} \sim \mathcal{U}([-3,3]) \text{ and } \Psi_k \sim \mathcal{U}([-3,3]).$$

$f_{i,k}$ and $g_{i,k}$ are Gaussian processes with exponential covariance matrix with length-scale 1, and $\varepsilon_i \sim \mathcal{N}(0,1)$ for $i \in [\![1, N]\!]$. X_i is observed at 101 equally spaced times of $[0,1]$.

Model 1. $Y_i = \rho^* \sum_{j=1}^N v_{ij,N} Y_j + \int_0^1 X_i(t)^\top \theta^*(t)\, \mathrm{d}t + \varepsilon_i$.
Model 2. $Y_i = \rho^* \sum_{j=1}^N v_{ij,N} Y_j + ||\alpha_i|| + \varepsilon_i$, $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,p})^\top$.
Model 3. $Y_i = \rho^* \sum_{j=1}^N v_{ij,N} Y_j + \langle \mathcal{S}^{D^*}(X_i), \mathcal{S}^{D^*}(\theta^*) \rangle + \varepsilon_i$ where $D^* = 2$.

The spatial weight matrix V_N was constructed using the k nearest neighbors method, and we considered the cases $k = 4$ and $k = 8$, $p = 2, 6, 10$ and $\rho^* = 0, 0.2, 0.4, 0.6, 0.8$.

For each model, several approaches were compared:

1. The approach proposed by Ahmed et al (2022) using a cubic B-splines basis with 12 equally spaced knots to approximate the X_i from the observed data and a functional PCA. As proposed by Ahmed et al (2022), we used a threshold on the number of coefficients such that the cumulative inertia was below 95%. However we also investigated this approach without using a pre-defined threshold on the number of coefficients.

2. Our proposed approach based on the penalized spatial regression.

3. Our proposed approach based on signatures projections. We considered a maximum truncation order for the signatures to reach a maximal number of coefficients of 10^3. Then a PCA was performed on the truncated signature coefficients vectors. Four strategies were considered: standardizing or not the signature coefficients before computing the PCA and using a threshold on the maximal number of coefficients such that the cumulative inertia was below 95%, or not using a threshold.

For each model and each value of k, p and ρ^*, 200 data sets were generated. Each data set was then split into a training, a validation and a test set such that the optimal number of coefficients (for the methods using PCA) or the optimal truncation order (for the penalized regression) was selected on the validation set based on the mean square error (MSE) criterion and the performances were finally evaluated on the test set using the MSE.

3.2 Results of the simulation study

All the approaches give similar estimations of the spatial autocorrelation coefficient ρ^* and the latter is reasonably well estimated.

Regarding the MSE, in all cases, our proposed approach based on the PCA on the signature coefficients presents better performances when the signature coefficients are standardized. Not using a threshold on the number of coefficients does not appear to change the performance except for Models 1 and 3 with $p = 10$, where it allows to slightly decrease the MSE. The approach of Ahmed et al (2022) presents similar performances with or without a threshold on the number of coefficients.

With Model 1, which is naturally favourable to the approach of the aforementioned authors, the latter presents the best performances. It should be noted, however, that our approach based on a penalized regression presents close MSEs.

In the case of Model 3 (which is naturally favorable to our methods), our approaches based on a penalized regression or a PCA (with standardization) on the signatures coefficients present much lower MSEs than the approach of Ahmed et al (2022).

Finally, with Model 2, the approach of the above-mentioned authors and our proposition using a PCA (with standardization) on the signatures coefficients give similar performance. Our approach using penalized regression presents slightly lower MSEs.

4 Discussion

Here we proposed an alternative to the traditional spatial autoregressive model with functional covariates. This new approach is based on the notion of signatures and presents the advantages of being applicable to a wide range of processes that are not necessary square-integrable processes, and to better capture the differences between the curves. We then proposed two methods for estimating the model, respectively based on a penalized regression and on signatures projections.
The simulation study shows that our approach is competitive with those in the literature.

Bibliographie

Ahmed, M. S., Broze, L., Dabo-Niang, S., & Gharbi, Z. (2022). Quasi-maximum Likelihood Estimators for Functional Linear Spatial Autoregressive Models. *Geostatistical Functional Data Analysis*, 286-328.

Fermanian, A. (2022). Functional linear regression with truncated signatures. *Journal of Multivariate Analysis*, 192