On generalized additive models with dependent time series covariates

Marton Ispany, Valderio A. Reisen, Glaura Franco, Pascal Bondon, Higor Cotta, Paulo Prezotti, Faraiba Sarquis Serpa

On generalized additive models with dependent time series covariates

Márton Ispány[1], Valdério A. Reisen[24], Glaura C. Franco[3], Pascal Bondon[4], Higor H. A. Cotta[4], Paulo R. P. Filho[24], and Faradiba S. Serpa[2]

[1] University of Debrecen, Debrecen, Hungary,
ispany.marton@inf.unideb.hu,
WWW home page: https://www.inf.unideb.hu/en/ispanymarton
[2] Federal University of Espírito Santo, Vitória, Brazil
[3] Federal University of Minas Gerais, Belo Horizonte, Brazil
[4] Laboratoire des Signaux et Systèmes (L2S), CNRS-CentraleSupélec-Université Paris-Sud, Gif sur Yvette, France

Abstract. The generalized additive model (GAM) is a standard statistical methodology and is frequently used in various fields of applied data analysis where the response variable is non-normal, e.g., integer valued, and the explanatory variables are continuous, typically normally distributed. Standard assumptions of this model, among others, are that the explanatory variables are independent and identically distributed vectors which are not multicollinear. To handle the multicollinearity and serial dependence together a new hybrid model, called GAM-PCA-VAR model, was proposed in [17] which is the combination of GAM with the principal component analysis (PCA) and the vector autoregressive (VAR) model. In this paper, some properties of the GAM-PCA-VAR model are discussed theoretically and verified by simulation. A real data set is also analysed with the aim to describe the association between respiratory disease and air pollution concentrations.

Keywords: air pollution, generalized additive model, multicollinearity, principal component analysis, time series, vector autoregressive model

1 Introduction

In the recent literature of time series, there has been an outstanding growth in models proposed for data that do not satisfy the Gaussian assumption. This is mainly the case when the response variable under study is a count series or an integer valued series. Procedures developed to analyse this kind of data comprises, for example, observation driven models, see [3] and [6], integer valued autoregressive (INAR) processes, see [1] and [2], or non-Gaussian state space models, see [8] and [10].

This paper is based on the talk "An application of the GAM-PCA-VAR model to respiratory disease and air pollution data" given by the first author.

Particularly in health and environmental studies, where the response variable is typically a count time series, the generalized additive model (GAM) has been widely used to associate the dependent series, such as the number of respiratory or cardiovascular diseases to some pollutant or climate variables, see, for example, [5], [13], [14], [16], [17] and [18] among others. Therefore, in general, the researches related to the study of the association between pollution and adverse health effects usually consider only one pollutant. This simple model choice may be due to the fact that the pollutants are linearly time correlated variables, see the discussion and references in the recent paper [17].

Recently, it has become common practice to use principal component analysis (PCA) in regression models to reduce the dimensionality of an independent set of data, especially the pollutants, which in some instances can include a large number of variables. The PCA is highly indicated to this purpose, as it can handle the multicollinearity problem that can cause biased regression estimates, see, for example, [21].

Nevertheless, use of PCA in the time series context can bring some misspecifications in the fit of the GAM model, as this technique requires that the data should be independent. This problem arises due to the fact that the principal components are linear combinations of the variables. In this context, as the covariates are time series, the autocorrelation present in the observations are promptly transferred to the principal components, see [20].

One solution to this issue was recently proposed by [17], see, also, [18], who introduced a model which combines GAM, PCA and the vector autoregressive (VAR) process. The authors suggest to apply the VAR model to the covariates, in order to eliminate the serial correlation and produce white noise processes, which in turn will be used to build the principal components in the PCA. The new variables obtained in the PCA are finally used as covariates in the GAM model, originating the so called GAM-PCA-VAR model. In their work, the authors have focused on presenting the model and showing its superiority compared to the sole use of GAM or the GAM-PCA procedures, but have not deepened on the theoretical properties of the model.

Thus, to cover this gap, this work aims to state and prove some properties of the GAM-PCA-VAR model, as well as to perform some simulation study to check the results for small samples.

The paper is organized as follows. Section 2 presents the main statistical model, GAM-PCA-VAR, addressed here and its related models as GAM, PCA and VAR, in some detail. In Section 3 the theoretical results are proved for the main model. Section 4 discusses the simulation results and Section 5 is devoted to the analysis of a real data set. Section 6 concludes the work.

2 The GAM-PCA-VAR model

The generalized additive model (GAM), see [11] and [19], with a Poisson marginal distribution is typically used to relate a non-negative integer valued response variable Y with a set of covariates or explanatory variables $X_1, \ldots, X_p$. In GAM

the expected value $\mu = \mathsf{E}(Y)$ of the response variable depends on the covariates via the formula

$$g(\mu) = \beta_0 + \sum_{i=1}^{p} f_i(X_i),$$

where g denotes the link function, β_0 is the intercept parameter and f_i's are functions with a specified parametric form, e.g., they are linear functions $f_i(x) = \beta_i x$, $\beta_i \in \mathbb{R}$, $i = 1, \ldots, p$, or non-parametric, e.g., they are simple smoothing functions like splines or moving averages. The unknown parameters β_0 and f_i, $i = 1, \ldots, p$ can be estimated by various algorithms, e.g., backfitting or restricted maximum likelihood (REML) method. However, if the data observed for variables Y and X_i, $i = 1, \ldots, p$, form a time series the observations cannot be considered as a result of independent experiments and the covariates present strong interdependence, e.g., multicollinearity or concurvity, the standard fitting methods result in remarkable bias, see, e.g., [7] and [17].

Let $\{Y_t\} \equiv \{Y_t\}_{t\in\mathbb{Z}}$ be a count time series, i.e., it is composed of non-negative integer valued random variables. We suppose that the explanatory variables form a zero-mean stationary vector time series $\{\boldsymbol{X}_t\} \equiv \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ of dimension p, i.e., $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{pt})^\top$ where $\top$ denotes the transpose, with the covariance matrix $\Sigma_{\boldsymbol{X}} = \mathsf{E}(\boldsymbol{X}_t \boldsymbol{X}_t^\top)$. Let $\mathcal{F}_t$ denote the σ-algebra which contains the available information up to time t for all $t \in \mathbb{Z}$ from the point of view of the response variable, e.g., $\boldsymbol{X}_t$ is $\mathcal{F}_{t-1}$-measurable. The GAM-PCA-VAR model is introduced in [17] as a probabilistic latent variable model. In this paper, we define this model in a more general form as

$$Y_t \,|\, \mathcal{F}_{t-1} \sim \mathrm{Poi}(\mu_t), \tag{1}$$

$$\boldsymbol{X}_t = \Phi \boldsymbol{X}_{t-1} + A\boldsymbol{Z}_t \tag{2}$$

with link

$$g(\mu_t) = \beta_0 + \sum_{i=1}^{p} \sum_{j=0}^{\infty} f_{ij}(Z_{i(t-j)}), \tag{3}$$

where $\mathrm{Poi}(\cdot)$ denotes the Poisson distribution, the latent variables $\{\boldsymbol{Z}_t\}$, $\boldsymbol{Z}_t = (Z_{1t}, \ldots, Z_{pt})^\top$, form a zero-mean Gaussian vector white noise process of dimension p with diagonal variance matrix $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, A is an orthogonal matrix of dimension $p \times p$, Φ is a matrix of dimension $p \times p$, g is a known link function, β_0 denotes the intercept, and f_{ij}'s are unknown functions. For a zero-mean Gaussian vector white noise process $\{\boldsymbol{Z}_t\}$ with covariance matrix Σ we shall use the notation $\{\boldsymbol{Z}_t\} \sim \mathrm{GWN}(\Sigma)$. See also [4, Definition 11.1.2]. Clearly, for all i, the univariate time series $\{Z_{it}\} \sim \mathrm{GWN}(\lambda_i)$, and $\{Z_{it}\}$ is mutually independent from $\{Z_{jt}\}$ for all $j \neq i$. We assume that all the eigenvalues of Φ are less than 1 in modulus which implies that equation (2) has a unique stationary causal solution. In the case of a Poisson distributed response variable the two widely used link functions are the identity link, $g(z) = z$, and the canonical logarithmic link, $g(z) = \log z$. The set $(\beta_0, \{f_{ij}\}, A, \Lambda, \Phi)$ forms the parameters of the GAM-PCA-VAR model to be estimated. We remark that

in the case of canonical logarithmic link function no additional assumption is needed for the parameters, while in the case of identity link function all the parameters in equation (3), i.e., β_0 and f_{ij}'s, have to be non-negative. It should be also emphasized that the underlying intensity process $\{\mu_t\}$ of $\{Y_t\}$ is also a time series with a complex dependence structure, and μ_t is $\mathcal{F}_{t-1}$-measurable for all $t \in \mathbb{Z}$. One can see that the time series $\{\boldsymbol{X}_t\}$ of covariates depends on $\{\boldsymbol{Z}_t\}$ by formula $\boldsymbol{X}_t = \sum_{k=0}^{\infty} \Phi^k A \boldsymbol{Z}_{t-k}$ for all t, see [4, Example 11.3.1].

The dependence of the response time series $\{Y_t\}$ from the explanatory vector time series $\{\boldsymbol{X}_t\}$ in the GAM-PCA-VAR model can be described by three transformation steps. Clearly, by equation (2), the latent variable can be expressed as $\boldsymbol{Z}_t = A^\top \boldsymbol{U}_t$, where $\boldsymbol{U}_t := \boldsymbol{X}_t - \Phi \boldsymbol{X}_{t-1}$ for all t. Thus, as the first step, the intermediate vector times series $\{\boldsymbol{U}_t\}$ is derived from filtering $\{\boldsymbol{X}_t\}$ by a VAR(1) filter. One can see that $\{\boldsymbol{U}_t\} \sim \mathrm{GWN}(\Sigma_{\boldsymbol{U}})$ where $\Sigma_{\boldsymbol{U}} := A \Lambda A^\top$. Then, as the second step, the latent vector time series $\{\boldsymbol{Z}_t\}$ as principal component (PC) vector is derived by principal component transformation of the intermediate vector white noise $\{\boldsymbol{U}_t\}$. The transformation matrix of the PCA is given by the spectral decomposition of $\Sigma_{\boldsymbol{U}}$. Finally, as the third step, the standard GAM with link (3) is fitting for the response time series $\{Y_t\}$ using the latent vector time series $\{\boldsymbol{Z}_t\}$. The impact of the VAR(1) filter in the first step is to eliminate the serial correlation present in the original covariates. On the other hand, the impact of the PCA in the second step is to eliminate the correlation in the state space of the original covariates. Hence, the result of these two consecutive transformations is the latent vector time series $\{\boldsymbol{Z}_t\}$ whose components, Z_{it}, $i = 1, \ldots, p$, $t \in \mathbb{Z}$, are independent Gaussian variables both in space and time. In the case of logarithmic link function, large positive values in a coordinate of the latent variable indicate locally high influence according to this latent factor. On the contrary, large negative values indicate negligible influence on the response, see, for example, [20]. The order of models in the acronym GAM-PCA-VAR corresponds to these steps starting with the third one and finishing with the first one.

The GAM-PCA-VAR model contains several submodels with particular dependence structure. If $\Phi = 0$ then the VAR equation (2) is simplified to a principal component transformation. In this case, we suppose that there is no serial correlation and we only have to handle the correlation in the state space of covariates. We have two transformation steps: PCA and GAM. This kind of models is called GAM-PCA model that is intensively studied nowadays, see, e.g., [15] and [22]. Beside the full PCA when all PCs are involved into the GAM, we can fit a restricted PCA model by defining $f_{ij} = 0$ for all $i > r$ and $j \geq 0$ where $r < p$. In this case, the first rth PCs are applied as covariates in the GAM step. If the matrices in VAR(1) model (2) have the following block structures

$$\Phi = \begin{bmatrix} \Phi_q & 0 \\ 0 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} A_q & 0 \\ 0 & I_{p-q} \end{bmatrix},$$

where the eigenvalues of the $q \times q$ matrix Φ_q are less than one in modulus, A_q is an orthogonal matrix of dimension $q \times q$ ($q \leq p$), and $f_{i1}(z) = \beta_i z$ with $\beta_i \in \mathbb{R}$ for

$i = 1, \ldots, r$ ($r \leq q$), f_{i1} is a general smoothing function for $i = q+1, \ldots, p$, $f_{ij} = 0$ otherwise, then we obtain the model that was studied in [17] and applied in the data analysis of Section 5. In this model it is supposed that the set of covariates can be partitioned into two sets: $(X_1, \ldots, X_q)$ are normal covariates, e.g., the pollutant variables in the terminology of Section 5, while $(X_{q+1}, \ldots, X_p)$ are so-called confounding variables as trend, seasonality, etc. The normal covariates satisfy a q-dimensional VAR(1) model, however, instead of the all coordinates of the innovation, only its first rth PCs are involved into the GAM taking into consideration that the covariates present strong inter-correlation. Finally, we note that our model can be further generalized by replacing equation (2) by the more general VARMA or VARIMA or their seasonal variants (SVARMA or SVARIMA) models.

Since the latent variables $\{\boldsymbol{Z}_t\}$ form a Gaussian vector time series, given a sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, the log-likelihood can be expressed in an explicit form, see [17] for a particular case. Because this log-likelihood is rather complicated a three-stage estimation method is proposed. Firstly, VAR(1) model is fitted to the original covariates by applying standard time series techniques. Secondly, PCA is applied for the residuals defined by $\widehat{\boldsymbol{Z}}_t = \boldsymbol{X}_t - \widehat{\Phi}\boldsymbol{X}_{t-1}$, $t = 2, \ldots, n$, where $\widehat{\Phi}$ denotes the estimated autoregressive coefficient matrix in the fitted VAR(1) model. Thirdly, GAM model is fitted using the PCs. The approach discussed above is similar to the principal component regression, see, e.g., [12, Chapter 8], and it can be considered as a three-stage non-linear regression method.

The first two steps of the above proposed parameter estimation method for GAM-PCA-VAR model can be interpreted as consecutive orthogonalizations, firstly in time and then in the state space of covariates. In [17, Remark] we argued that the order of VAR filter and PCA can not be interchanged because the orthogonalization in the state space does not eliminate the serial correlation and, as the necessary next step, the orthogonalization in time by VAR filter bring back the inter-correlation between the covariates. In what follows, we demonstrate this phenomena by giving a simple example. Let $\{\boldsymbol{X}_t\}$ be a zero-mean causal VAR(1) process defined by

$$\boldsymbol{X}_t = \Psi \boldsymbol{X}_{t-1} + \boldsymbol{W}_t,$$

where $\{\boldsymbol{W}_t\}$ is a zero-mean vector white noise process with variance matrix Σ_W. Suppose that the variance matrix Σ_X of $\{\boldsymbol{X}_t\}$ is diagonal, i.e., the coordinates of $\{\boldsymbol{X}_t\}$ can be interpreted as PCs after PCA. Then Σ_W is not necessarily a diagonal matrix, which implies that a VAR(1) filter may result in an inter-correlated white noise. Namely, consider the following parameters $\Sigma_W = A\Lambda A^\top$ and $\Psi = ASA^\top$, where Λ and S are diagonal matrices and A is an orthogonal matrix. In other words, we suppose that the orthogonal matrix A in the spectral decomposition of Σ_W diagonalizes the autoregressive coefficient matrix as well. Then, we have, by formula (11.1.13) in [4], that

$$\Sigma_X = \sum_{j=0}^{\infty} \Psi^j \Sigma_W (\Psi^\top)^j = \sum_{j=0}^{\infty} A S^j \Lambda S^j A^\top = A \mathrm{diag}\left\{ \frac{\lambda_i}{1 - s_i^2} \right\} A^\top.$$

Let $\sigma^2 > \max_i\{\lambda_i\}$ arbitrary and define $s_i := \sqrt{1-\lambda_i/\sigma^2}$ for all i. Clearly, Ψ is a causal matrix since all its eigenvalues are less than 1 in modulus and $\Sigma_X = \sigma^2 I$, i.e., the coordinates of $\{\boldsymbol{X}_t\}$ are uncorrelated. However, the innovation variance matrix Σ_W can be arbitrary proving that the application of VAR filter for a non-intercorrelated vector time series can give inter-correlated vector white noise in its coordinates.

Now, we present some particular examples of GAM-PCA-VAR models.

Example 1. One of the simplest GAM-PCA-VAR models is the model with dimension $p = 1$ and log-linear link function. In this case, there is only one covariate $\{X_t\}$, and the VAR equation (2) is an AR(1) model

$$X_t = \phi X_{t-1} + Z_t, \tag{4}$$

where $|\phi| < 1$ which guarantees the existence of a unique stationary causal solution, $\{Z_t\} \sim \mathrm{GWN}(\lambda)$, $\lambda > 0$. We remark that $A = 1$ in equation (2) in order for the model to be identifiable. The link is log-linear expressed as

$$\log \mu_t = \beta_0 + \beta_1 Z_t. \tag{5}$$

The parameter set of this model is $(\beta_0, \beta_1, \lambda, \phi)$ with parameter space $\mathbb{R}^2 \times \mathbb{R}_+ \times (-1, 1)$. In this model, there is no dimension reduction. Clearly, $Z_t = X_t - \phi X_{t-1}$, thus the response depends on the covariate through the link

$$\log \mu_t = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1}, \tag{6}$$

where there is a one-to-one correspondence between the parameter sets (β_0, β_1, ϕ) and $(\gamma_0, \gamma_1, \gamma_2)$ defined by the equations $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1$ and $\gamma_2 = -\phi\beta_1$ provided $\phi \neq 0$. However, if we fit the standard GAM by using the link (6) with covariates X_t and X_{t-1} at time t, we take no count of the interdependence in time series $\{X_t\}$ which can result in biased and inconsistent estimators of the GAM parameters.

Example 2. Define a particular two-dimensional ($p = 2$) GAM-PCA-VAR model with logarithmic link function in the following way. The two-dimensional covariate vector process $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t = (X_{1t}, X_{2t})^\top$, satisfies the VAR(1) model

$$\begin{bmatrix} X_{1t} \\ X_{2t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix} \begin{bmatrix} X_{1(t-1)} \\ X_{2(t-1)} \end{bmatrix} + \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} Z_{1t} \\ Z_{2t} \end{bmatrix},$$

where $|\phi_1| < 1$, $|\phi_2| < 1$ and $\{Z_{it}\} \sim \mathrm{GWN}(\lambda_i)$ with $\lambda_i > 0$, $i = 1, 2$, which are independent from each other. Note that the set of two-dimensional orthogonal matrices, A, can be parametrized by an angle parameter $\varphi \in [0, 2\pi)$. We assume that the link is

$$\log \mu_t = \beta_0 + \beta_1 Z_{1t}.$$

The parameter set of this model is $(\beta_0, \beta_1, \varphi, \lambda_1, \lambda_2, \phi_1, \phi_2)$ and the parameter space is $\mathbb{R}^2 \times [0, 2\pi) \times \mathbb{R}_+^2 \times (-1, 1)^2$. Note that, in this model, there is a PCA step as a dimension reduction since only the first coordinate $\{Z_{1t}\}$ of the vector

innovation is involved into the GAM as covariate. One can see that the response depends on the covariates through the link

$$\log \mu_t = \gamma_0 + \gamma_1 X_{1t} + \gamma_2 X_{2t} + \gamma_3 X_{1(t-1)} + \gamma_4 X_{2(t-1)},$$

where $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1 \cos\varphi$, $\gamma_2 = \beta_1 \sin\varphi$, $\gamma_3 = -\beta_1\phi_1 \cos\varphi$ and $\gamma_4 = -\beta_1\phi_2 \sin\varphi$. Thus, the intensity process $\{\mu_t\}$ depends on all coordinates of $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-1}$. Clearly, there is a one-to-one correspondence between the two parameter sets $(\beta_0, \beta_1, \varphi, \phi_1, \phi_2)$ and $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

Example 3. A seasonal one-dimensional GAM-PCA-VAR model with linear link function can be defined in the following way. Suppose that the one-dimensional covariate process $\{X_t\}$ satisfies the $\mathrm{SAR}_s(1)$ model:

$$X_t = \phi X_{t-s} + Z_t,$$

where $|\phi| < 1$, $\{Z_t\} \sim \mathrm{GWN}(\lambda)$ with $\lambda > 0$ and $s \in \mathbb{Z}_+$ denotes the seasonal period. The link is linear and is given by

$$\mu_t = \beta_0 + \beta_1 f(Z_t),$$

where $f : \mathbb{R} \to \mathbb{R}_+$ is a known function and $\beta_0, \beta_1 \in \mathbb{R}_+$ are parameters. The parameter set of this model is $(\beta_0, \beta_1, \lambda, \phi)$ with parameter space $\mathbb{R}_+^3 \times (-1, 1)$. The response variable depends on the original covariates through the link

$$\mu_t = \beta_0 + \beta_1 f(X_t - \phi X_{t-s}).$$

If the function f is sufficiently smooth we have by approximation $f(X_t - \phi X_{t-s}) \approx f(X_t) - \phi f'(X_t) X_{t-s}$, and then

$$\mu_t = \gamma_0 + \gamma_1 f_1(X_t) + \gamma_2 f_2(X_t, X_{t-s}), \tag{7}$$

where f_1, f_2 are known functions and $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1$ and $\gamma_2 = -\beta_1\phi$. Thus, the response depends on the original covariate and its s-step lagged series through the standard GAM. However, the covariates in equation (7) are clearly dependent.

3 Theoretical results

In this section, we prove some theoretical results for particular classes of GAM-PCA-VAR models. Consider the log-linear model defined by the link

$$\log \mu_t = \beta_0 + \sum_{i=1}^{p} \sum_{j=0}^{\infty} \beta_{ij} Z_{i(t-j)}, \tag{8}$$

where $\beta_0, \beta_{ij} \in \mathbb{R}$, $i = 1, \ldots, p$, $j \in \mathbb{Z}_+$. The first proposition is about the existence of log-linear GAM-PCA-VAR models.

Proposition 1. *Suppose that* $\sigma^2 := \sum_{i=1}^{p} \lambda_i \sum_{j=0}^{\infty} \beta_{ij}^2$ *is finite. Then the GAM-PCA-VAR model with log-linear link* (8) *has solution* $\{(Y_t, \boldsymbol{X}_t)\}$ *which is a strictly stationary process and* $\mathsf{E}(Y_t) = \mathsf{E}(\mu_t) = \exp(\beta_0 + \sigma^2/2)$ *for all* $t \in \mathbb{Z}$.

Proof. By conditioning we have that

$$\mathsf{E}(Y_t) = \mathsf{E}(\mathsf{E}(Y_t \,|\, \mathcal{F}_{t-1})) = \mathsf{E}(\mu_t) = \mathsf{E}(\exp(\log \mu_t)) = \exp(\beta_0 + \sigma^2/2) \tag{9}$$

is finite since, by equation (8), $\log \mu_t \sim \mathcal{N}(\beta_0, \sigma^2)$, i.e., μ_t has a lognormal distribution, and the moment generating function of $\xi \sim \mathcal{N}(\beta_0, \sigma^2)$ is given by $M_\xi(t) := \mathsf{E}(\exp(t\xi)) = \exp(\beta_0 t + (\sigma t)^2/2)$. Thus, the non-negative integer valued random variable Y_t is finite with probability one for all $t \in \mathbb{Z}$. The vector time series $\{\boldsymbol{Z}_t\}$ forms a Gaussian white noise. Hence it is strictly stationary process with backshift operator $B(\boldsymbol{Z}_t) = \boldsymbol{Z}_{t-1}$ for all $t \in \mathbb{Z}$. Since both stochastic processes $\{Y_t\}$ and $\{\boldsymbol{X}_t\}$ depend on $\{\boldsymbol{Z}_t\}$ through time-invariant functionals, we have the strict stationarity of $\{(Y_t, \boldsymbol{X}_t)\}$ and $B(\boldsymbol{X}_t) = \boldsymbol{X}_{t-1}$, $B(Y_t) = Y_{t-1}$ for all $t \in \mathbb{Z}$. □

In the next proposition, we prove that all moments of the log-linear GAM-PCA-VAR model are finite.

Proposition 2. *Suppose that* σ^2 *defined in Proposition 1 is finite. Then all moments of the stochastic process* $\{(Y_t, \boldsymbol{X}_t)\}$ *are finite. In particular, we have, for all* $t \in \mathbb{Z}$,

$$\begin{aligned}
\mathrm{Var}(Y_t) &= \exp(2\beta_0 + \sigma^2)(\exp(\sigma^2) - 1 + \exp(-\beta_0 - \sigma^2/2)), \\
\mathrm{Var}(\mu_t) &= \exp(2\beta_0 + \sigma^2)(\exp(\sigma^2) - 1).
\end{aligned}$$

Proof. Let $r \in \mathbb{N}$. Define the rth factorial of a non-negative integer k as $k^{[r]} := k(k-1)\cdots(k-r+1)$ and let $k^{[0]} := 1$. For the rth factorial moment of Y_t we have by conditioning that

$$\begin{aligned}
\mathsf{E}(Y_t^{[r]}) &= \sum_{k=0}^{\infty} k^{[r]} \mathsf{P}(Y_t = k) = \mathsf{E} \sum_{k=0}^{\infty} k^{[r]} \mathsf{P}(Y_t = k \,|\, \mathcal{F}_{t-1}) \\
&= \mathsf{E} \sum_{k=r}^{\infty} \frac{\mu_t^k}{(k-r)!} e^{-\mu_t} = \mathsf{E}(\mu_t^r)
\end{aligned}$$

for all $t \in \mathbb{Z}$. Similarly to (9), we have that the factorial moments are finite, since

$$\mathsf{E}(Y_t^{[r]}) = \mathsf{E}(\mu_t^r) = \mathsf{E}(\exp(r \log \mu_t)) = \exp\{\beta_0 r + (\sigma r)^2/2\}. \tag{10}$$

Since the higher order moments can be expressed by the factorial moment via the formula

$$\mathsf{E}(Y^r) = \sum_{j=0}^{r} S(r, j) \mathsf{E}(Y^{[j]}),$$

where $S(r, j)$'s denotes Stirling numbers of the second kind, the finiteness of all higher order moments follows easily. Since $\{\boldsymbol{X}_t\}$ is a Gaussian process all

its moments are finite. Finally, the existence of mixed moments follows by the Cauchy-Schwarz inequality.

From Equation (10), we have

$$\begin{aligned}\operatorname{Var}(\mu_t) =& \mathsf{E}(\mu_t^2) - \mathsf{E}^2(\mu_t) = \exp(2\beta_0 + (2\sigma)^2/2) - \exp(2\beta_0 + \sigma^2) \\ &= \exp(2\beta_0 + \sigma^2)(\exp(\sigma^2) - 1).\end{aligned}$$

Finally, the formula for $\operatorname{Var}(Y_t)$ can be derived by

$$\operatorname{Var}(Y_t) = \mathsf{E}(\operatorname{Var}(Y_t \mid \mathcal{F}_{t-1})) + \operatorname{Var}(\mathsf{E}(Y_t \mid \mathcal{F}_{t-1})) = \mathsf{E}(\mu_t) + \operatorname{Var}(\mu_t). \qquad \square$$

The existence of all moments for the log-linear GAM-PCA-VAR process is to be compared with the same result for the integer valued GARCH, so-called INGARCH, process, see [9, Proposition 6]. This implies that the log-linear GAM-PCA-VAR process possesses second and higher order structures, e.g., the autocorrelation function, the spectral density function, the cumulants and the higher order spectra exist. Let ρ_Y denotes the autocorrelation function of the time series $\{Y_t\}$.

Proposition 3. *For the auto- and cross-correlation functions of the GAM-PCA-VAR process $\{(Y_t, \boldsymbol{X}_t)\}$ with intensity process $\{\mu_t\}$, we have $\rho_Y(h) = c_Y \rho(h)$, $\rho_\mu(h) = c_\mu \rho(h)$ and $\rho_{Y\mu}(h) = c_{Y\mu}\rho(h)$ where*

$$\rho(h) := \exp\left(\sum_{i=1}^{p} \lambda_i \sum_{j=0}^{\infty} \beta_{i(j+|h|)}\beta_{ij}\right) - 1, \qquad h \in \mathbb{Z} \setminus \{0\},$$

and the constants $c_Y, c_\mu, c_{Y\mu}$ are defined by

$$c_Y := (\exp(\sigma^2) - 1 + \exp(-\beta_0 - \sigma^2/2))^{-1}, \quad c_\mu := (\exp(\sigma^2) - 1)^{-1}, \quad c_{Y\mu} := \sqrt{c_Y c_\mu}.$$

Moreover, $\operatorname{Cov}(Y_{t+h}, \boldsymbol{X}_t) = \operatorname{Cov}(\mu_{t+h}, \boldsymbol{X}_t) = \mathsf{E}(Y_{t+h}\boldsymbol{X}_t) = \mathsf{E}(\mu_{t+h}\boldsymbol{X}_t) = C(h)$ *with*

$$C(h) := \exp(\beta_0 + \sigma^2/2) \times \begin{cases} \sum_{k=0}^{\infty} \boldsymbol{\Phi}^k A(\boldsymbol{\lambda} \circ \boldsymbol{\beta}_{h+k}) & \text{if} \quad h \geq 0, \\ \sum_{k=0}^{\infty} \boldsymbol{\Phi}^{k-h} A(\boldsymbol{\lambda} \circ \boldsymbol{\beta}_k) & \text{if} \quad h \leq 0, \end{cases} \tag{11}$$

where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_p)^\top$, $\boldsymbol{\beta}_j := (\beta_{1j}, \ldots, \beta_{pj})^\top$, $j \in \mathbb{Z}_+$, and $\circ$ denotes the entrywise (Hadamard) product.

Proof. Let $h \in \mathbb{N}$. One can see that for the intensity process we have $\mu_{t+h} = \mu_{th}^{(1)} \mu_{th}^{(2)}$ where

$$\log \mu_{th}^{(1)} := \beta_0 + \sum_{i=1}^{p} \sum_{j=1}^{h} \beta_{i(h-j)} Z_{i(t+j)}, \qquad \log \mu_{th}^{(2)} := \sum_{i=1}^{p} \sum_{j=0}^{\infty} \beta_{i(j+h)} Z_{i(t-j)}.$$

Clearly, $\mu_{th}^{(1)}$ is independent of $\mathcal{F}_{t-1}$ and Y_t, while $\mu_{th}^{(2)}$ is $\mathcal{F}_{t-1}$-measurable. Hence, we have by conditioning that

$$\begin{aligned}\mathsf{E}(Y_{t+h}Y_t) =&\mathsf{E}(Y_t\mathsf{E}(Y_{t+h}\,|\,\mathcal{F}_{t+h-1})) = \mathsf{E}(\mu_{t+h}Y_t) = \mathsf{E}(\mu_{th}^{(1)}\mu_{th}^{(2)}Y_t)\\ =&\mathsf{E}(\mu_{th}^{(1)})\mathsf{E}(\mu_{th}^{(2)}\mathsf{E}(Y_t\,|\,\mathcal{F}_{t-1})) = \mathsf{E}(\mu_{th}^{(1)})\mathsf{E}(\mu_{th}^{(2)}\mu_t) = \mathsf{E}(\mu_{t+h}\mu_t)\end{aligned}$$

since μ_t is independent of $\mu_{th}^{(1)}$. This gives the result for $h > 0$. On the other hand, for all $h > 0$, again by conditioning, $\mathsf{E}(Y_{t+h}\mu_t) = \mathsf{E}(\mu_{t+h}\mu_t)$. Thus

$$\mathrm{Cov}(Y_{t+h}, Y_t) = \mathrm{Cov}(\mu_{t+h}, \mu_t) = \mathrm{Cov}(Y_{t+h}, \mu_t), \quad h \in \mathbb{Z}\setminus\{0\}.$$

Since

$$\mathsf{E}(\mu_{t+h}\mu_t) = \mathsf{E}(\mu_{th}^{(1)}\mu_{th}^{(2)}\mu_t) = \mathsf{E}(\mu_{th}^{(1)})\mathsf{E}(\mu_{th}^{(2)}\mu_t)$$

similarly to equation (9) we have

$$\begin{aligned}\mathsf{E}(\mu_{t+h}\mu_t) =& \exp\left(2\beta_0 + \frac{1}{2}\sum_{i=1}^{p}\lambda_i\left(\sum_{j=0}^{h-1}\beta_{ij}^2 + \sum_{j=0}^{\infty}(\beta_{i(j+h)} + \beta_{ij})^2\right)\right)\\ =& \exp\left(\sum_{i=1}^{p}\lambda_i\sum_{j=0}^{\infty}\beta_{i(j+h)}\beta_{ij}\right)\mathsf{E}(\mu_{t+h})\mathsf{E}(\mu_t).\end{aligned}$$

Thus, the first part of the proposition follows by Proposition 2.

Next we prove the formula (11) for the cross-correlations of response and covariate variables. Clearly, by conditioning, $\mathsf{E}(Y_{t+h}\boldsymbol{X}_t) = \mathsf{E}(\mu_{t+h}\boldsymbol{X}_t)$ for all $h \in \mathbb{Z}_+$. On the other hand, for all $t \in \mathbb{Z}$, $h \in \mathbb{Z}_+$, we have $\boldsymbol{X}_{t+h} = \boldsymbol{X}_{th}^{(1)} + \boldsymbol{X}_{th}^{(2)}$ where

$$\boldsymbol{X}_{th}^{(1)} := \sum_{k=1}^{h}\Phi^{h-k}A\boldsymbol{Z}_{t+k}, \qquad \boldsymbol{X}_{th}^{(2)} := \sum_{k=0}^{\infty}\Phi^{h+k}A\boldsymbol{Z}_{t-k}.$$

One can see that $\boldsymbol{X}_{th}^{(1)}$ is independent of $\mathcal{F}_{t-1}$ and Y_t, while $\boldsymbol{X}_{th}^{(2)}$ is $\mathcal{F}_{t-1}$-measurable. Thus, we have that

$$\begin{aligned}\mathsf{E}(\boldsymbol{X}_{t+h}Y_t) =&\mathsf{E}((\boldsymbol{X}_{th}^{(1)} + \boldsymbol{X}_{th}^{(2)})Y_t) = \mathsf{E}(\boldsymbol{X}_{th}^{(1)})\mathsf{E}(Y_t) + \mathsf{E}(\boldsymbol{X}_{th}^{(2)}\mathsf{E}(Y_t\,|\,\mathcal{F}_{t-1}))\\ =&\mathsf{E}(\boldsymbol{X}_{th}^{(1)})\mathsf{E}(\mu_t) + \mathsf{E}(\boldsymbol{X}_{th}^{(2)}\mu_t) = \mathsf{E}(\boldsymbol{X}_{t+h}\mu_t).\end{aligned}$$

Hence $\mathsf{E}(Y_{t+h}\boldsymbol{X}_t) = \mathsf{E}(\mu_{t+h}\boldsymbol{X}_t)$ for all $h \in \mathbb{Z}$ and it is enough to compute the cross-correlation between $\{\boldsymbol{X}_t\}$ and $\{\mu_t\}$. Let $h \geq 0$. For all $\ell \in \{1,\dots,p\}$, $k \in \mathbb{Z}_+$ let $\mathcal{I}_{\ell k}^h := \{1,\dots,p\}\times\mathbb{Z}_+\setminus(\ell, k+h)$ and define the random variables

$$\log\xi_{\ell k}^{th} := \beta_0 + \sum_{(i,j)\in\mathcal{I}_{\ell k}^h}\beta_{ij}Z_{i(t+h-j)}, \qquad \log\eta_{\ell k}^{th} := \beta_{\ell(k+h)}Z_{\ell(t-k)}.$$

Then $\mu_{t+h} = \xi_{\ell k}^{th}\eta_{\ell k}^{th}$, where the factors in this decompostion are independent. Since $\mathsf{E}(\mu_{t+h}\boldsymbol{X}_t) = \sum_{k=0}^{\infty}\Phi^k A\mathsf{E}(\mu_{t+h}\boldsymbol{Z}_{t-k})$ and, using the fact that for $Z \sim \mathcal{N}(0,\lambda)$ and $\beta \in \mathbb{R}$ we have $\mathsf{E}(Z\exp(\beta Z)) = \beta\lambda\exp(\lambda\beta^2/2)$,

$$\mathsf{E}(\mu_{t+h}Z_{\ell(t-k)}) = \mathsf{E}(\xi_{\ell k}^{th}\eta_{\ell k}^{th}Z_{\ell(t-k)}) = \mathsf{E}(\xi_{\ell k}^{th})\mathsf{E}(\eta_{\ell k}^{th}Z_{\ell(t-k)}) = \mathsf{E}(\mu_{t+h})\beta_{\ell(k+h)}\lambda_\ell,$$

we obtain the formula (11). The proof is similar in the case of $h < 0$. □

Remark 1. It is easy to see that if $\beta_{ij} = \beta_i^j$ for all i, j, then the function ρ is given by $\rho(h) = \exp(\sum_{i=1}^p \lambda_i \beta_i^{|h|}/(1-\beta_i^2)) - 1$, $h \in \mathbb{Z}$. If β_i's are all positive then ρ is positive everywhere and we have autocorrelation functions which are similar to what is displayed in Figure 1. For the one-dimensional model in Example 1 we have the cross-correlation function (CCF) $C(h) = \exp(\beta_0 + \lambda\beta_1^2/2)\lambda\beta_1\phi^{-h}$ for $h \leq 0$ and $C(h) = 0$ for $h > 0$. If $\phi > 0$ then, according to positive or negative β_1, we obtain everywhere positive or negative CCFs. For example, see the CCFs in Figure 2 between the response (Admissions) and pollutants CO, NO_2 that are positive and the CCFs between the response (Admissions) and O_3, SO_2 that are negative at every lag, respectively.

Consider another widely used link function, the linear one, and define the linear GAM-PCA-VAR model by the link

$$\mu_t = \beta_0 + \sum_{i=1}^{p} \sum_{j=0}^{\infty} \beta_{ij} f(Z_{i(t-j)}), \tag{12}$$

where $\beta_0, \beta_{ij} \in \mathbb{R}_+$, $i = 1, \ldots, p$, $j \in \mathbb{Z}_+$ are parameters and $f : \mathbb{R} \to \mathbb{R}_+$ is a known function, e.g., $f(z) = \exp(z)$. Let $\varphi(x \,|\, \lambda)$ denote the probability density function of the normal distribution with mean 0 and variance λ.

Proposition 4. *Suppose that, for all* $i = 1, \ldots, p$, $\sum_{j=0}^{\infty} \beta_{ij} < \infty$ *and* $\tau_i := \int_{-\infty}^{\infty} f(x)\varphi(x \,|\, \lambda_i)\mathsf{d}x < \infty$. *Then the GAM-PCA-VAR model with linear link* (12) *has a strictly stationary solution* $\{(Y_t, \boldsymbol{X}_t)\}$. *Moreover,* $\mathsf{E}(Y_t) = \mathsf{E}(\mu_t) = \beta_0 + \sum_{i=1}^p \tau_i \sum_{j=0}^{\infty} \beta_{ij}$.

Proof. The proof is similar to the proof of Proposition 1. □

Clearly, the assumptions of Proposition 4 do not necessarily garantee the existence of higher order moments of linear GAM-PCA-VAR process. Indeed, the rth order moment $\mathsf{E}(Y_t^r)$ is finite if and only $\int_{-\infty}^{\infty} f^r(x)\varphi(x \,|\, \lambda_i)\mathsf{d}x < \infty$ for all i where $r \geq 1$.

4 Simulation study

In order to evaluate the effect on the parameter estimation of a GAM model in the presence of temporal correlation in the covariate $\{X_t\}$, a simulation study was conducted. The data were generated according to the model discussed in Example 1. Three estimation methods were considered: the standard GAM with only one covariate where the estimated parameters were β_0 and β_1 (M1); the standard GAM with two covariates, the original one and its 1-step lagged series, where the estimated parameters were $\beta_0, \beta_1, \beta_2$ and $\phi = -\beta_2/\beta_1$ (M2); the full GAM-PCA-VAR model by the procedure described in Section 2 where all parameters $\beta_0, \beta_1, \phi, \lambda$ were estimated (M3).

For the model discussed in Example 1 the data were generated under $\beta_0 = 0.2$, $\beta_1 = 1$, $\lambda = 2$ and three scenarios were considered as $\phi = -0.7, 0.3, 0.9$ to model strong negative, small positive and strong positive correlations, respectively. In order to model the impact due to some unobservable variables, e.g., environmental ones in the context of the next section, independent $\mathcal{N}(0, 0.1)$ distributed random variables were added to the predictor of $\log \mu_t$ for all $t \in \mathbb{Z}$. The sample size $n = 1000$ and the number of Monte Carlo simulations was equal to 100. The empirical values of mean, bias and mean square error (MSE) are displayed in Table 1. All results were obtained by using R-code.

Table 1. Simulation results for model in Example 1

Estimation method	ϕ	Parameter	Mean	Bias	MSE
M1: GAM with X_t	-0.7	$\beta_0 = 0.2$	0.699	0.499	0.253
		$\beta_1 = 1$	0.507	-0.492	0.244
M2: GAM with X_t, X_{t-1}		$\beta_0 = 0.2$	0.204	0.004	0.001
		$\beta_1 = 1$	0.999	-0.001	0.0002
		$\phi = -0.7$	-0.7	0	0.0001
M3: GAM-PCA-VAR		$\beta_0 = 0.2$	0.205	0.005	0.001
		$\beta_1 = 1$	0.999	-0.001	0.0002
		$\phi = -0.7$	-0.695	0.004	0.0005
		$\lambda = 2$	2.003	0.003	0.008
M1: GAM with X_t	0.3	$\beta_0 = 0.2$	0.302	0.102	0.012
		$\beta_1 = 1$	0.905	-0.095	0.009
M2: GAM with X_t, X_{t-1}		$\beta_0 = 0.2$	0.209	0.009	0.001
		$\beta_1 = 1$	0.998	-0.002	0.0002
		$\phi = 0.3$	0.3	0	0.0002
M3: GAM-PCA-VAR		$\beta_0 = 0.2$	0.209	0.009	0.001
		$\beta_1 = 1$	0.999	-0.001	0.0002
		$\phi = 0.3$	0.306	0.006	0.0008
		$\lambda = 2$	1.995	-0.005	0.009
M1: GAM with X_t	0.9	$\beta_0 = 0.2$	1.002	0.802	0.651
		$\beta_1 = 1$	0.191	-0.809	0.655
M2: GAM with X_t, X_{t-1}		$\beta_0 = 0.2$	0.2	0	0.001
		$\beta_1 = 1$	1	0	0.0002
		$\phi = 0.9$	0.899	-0.001	0
M3: GAM-PCA-VAR		$\beta_0 = 0.2$	0.203	0.003	0.001
		$\beta_1 = 1$	1	0	0.0002
		$\phi = 0.9$	0.899	-0.001	0.0001
		$\lambda = 2$	2.007	0.007	0.0086

In the case of standard GAM estimation (M1) it can be seen that the estimate of β_1 is heavily affected by the autocorrelation structure present in the covariate, by presenting a negative bias which increases in absolute value as $|\varphi|$ increases. The estimated MSE also increases substantially with $|\varphi|$. On the other hand, it can also be seen that the fitted standard GAM model tends to severely

overestimate β_0. Contrarily, the estimation methods M2 and M3 work equally well, the estimates of the parameters are very close to the true values with noticeably small MSE. The undoubted advantage of method M3 against M2 is that an AR(1) model is also fitted for the covariate where the innovation variance λ is estimated and which can be applied later in the prediction. In this procedure firstly the covariate variable is predicted by equation (4) and then the response variable is predicted by the GAM using the link (5).

5 Application to air pollution data

In this study, the number of hospital admissions (Admissions) for respiratory diseases (RD) as response variable was obtained from the main childrens emergency department in the Vitória Metropolitan Area (called Hospital Infantil Nossa Senhora da Gloria), ES, Brazil. The following atmospheric pollutants as covariates were studied: particulate material (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). For details, e.g., descriptive statistics and basic time series plots, see [17]. The data analysed in this section can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

The graphs of the sampling functions of the autocorrelations and partial autocorrelations in Figure 1 show that the series of the number of hospital admissions for RD possesses seasonal behaviour, which was to be expected for this phenomena. Another characteristic observed in the series was an apparently weak stationarity. Similar graphs for the pollutant series can be found in [17].

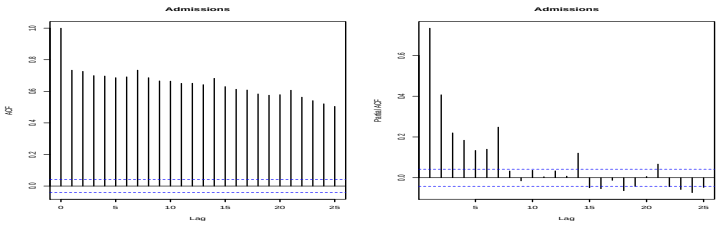

Fig. 1. Sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the response variable.

Figure 2 shows the sample cross-correlation functions (CCF) between the response and pollutant covariates. As we discussed in Remark 1 four CCF's among them present similar behaviour: the impact of pollutants CO and NO_2 is positive while the impact of SO_2 and O_3 are negative to the response variable at every lag. This observation is consistent with the PCA result presented in [17], see Table 5, where CO and NO_2 form a joint cluster for PC1. On the other hand, all CCF's possess seasonal behaviour as well.

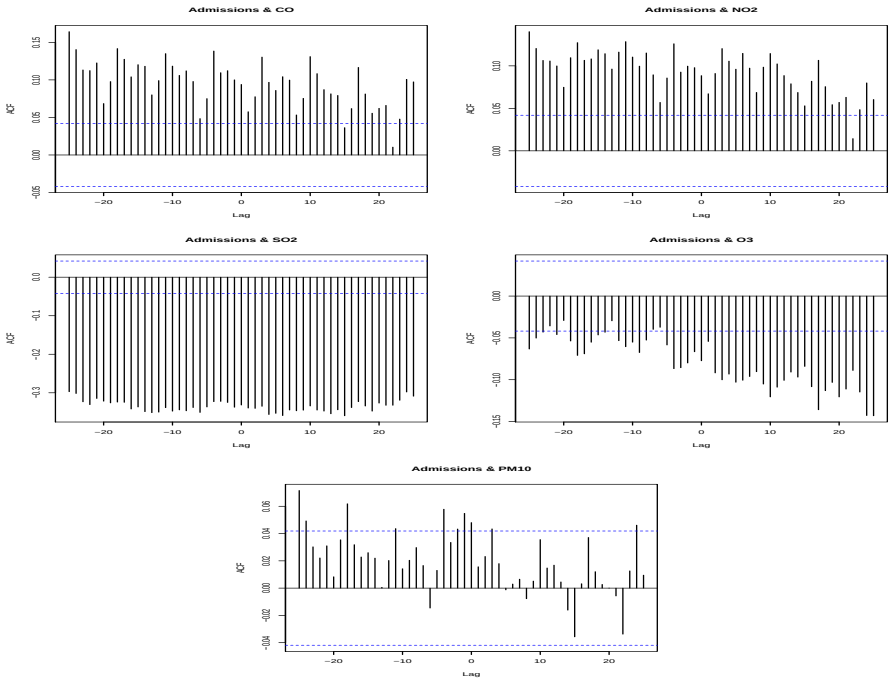

Fig. 2. Sample cross-correlation function (CCF) of the response and pollutant variables.

Figure 3 shows the sample cross-correlation functions (CCF) between the response variable and the first three PCs derived from applying PCA for the vector of pollutants. In Section 3.2 of [17], see Table 5 there, one can see that the first three components correspond to 83.2% of the total variability. The temporal behaviour of the PCs is also presented in the autocorrelation plots of [17, Figure 4]. The autocorrelations and the cross-correlations displayed here presented heavy seasonality as well. On the other hand, the shape of the CCFs for the response and PCs can also be classifed into similar groups to the CCFs in Figure 2. The CCF of PC1 is similar to the one of the PM_{10}. The CCF of PC2 displays only negative correlations similar to SO_2 and O_3, while the CCF of PC3 (Figure 3) displays only positive correlations, see CO and NO_2 in Figure 2.

In order to filter the vigorous seasonality both in the response and pollutant variables, seasonal ARMA filters with a 7-day period were applied. The pollutant vector time series and the one-dimensional response time series were filtered by $\text{SVAR}_7(1)$ and $\text{SARMA}_7(1,1)$ processes, respectively. The residuals obtained by these filters indicate remaining significant correlations, see the CCFs between these residuals in Figure 4. The significant cross-correlations and their respective lags are presented in Table 2. Clearly, the correlations which belong

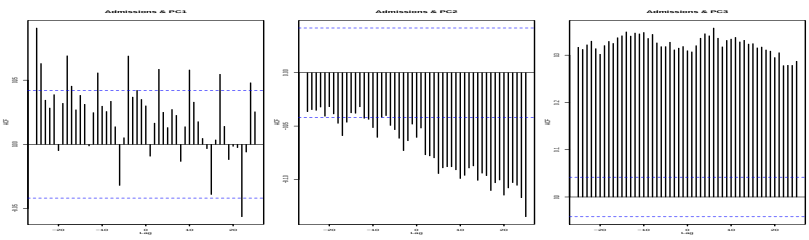

Fig. 3. Sample cross-correlation function (CCF) of the response and first three PCs.

to the negative lags are spurious. However, the correlations which belong to the positive lags measure the true impact of a covariate. For example, there are significant correlations at lag 2 for pollutants PM_{10}, NO_2 and CO equally which could mean that the influence of these pollutants to the response indicates 2 days delay. Contrarily, the influence of the pollutants SO_2 and O_3 presents far delays.

Table 2. Significant cross-correlations and their respective lags between the response and pollutants after the filtering

	RD×SO_2					RD×NO_2				
Lag	-19	-14	-6	12	23	-12	2	4	14	22
Value	-0.063	-0.062	-0.042	-0.047	-0.051	-0.044	-0.050	0.048	0.053	-0.044

	RD×PM10		RD×CO			RD×O_3	
Lag	2	23	-12	2	6	9	25
Value	-0.044	-0.043	-0.053	-0.048	0.045	0.054	-0.055

Figure 5 shows the sample CCF between the residuals of the response variable and the first three PCs after the filtering. The significant cross-correlations and its respective lags are presented in Table 3. It should be emphasized that there are strong coincidences in the lags between Table 2 and 3. For example, the lag 2 in PC1 corresponds to the pollutants PM_{10}, NO_2 and CO, the lag 6 in PC1 corresponds to the pollutant CO, while lag 25 in PC1 corresponds to the pollutant O_3. The lag 12 in PC2 corresponds to the pollutant SO_2. Finally, the lag 14 corresponds to the pollutant NO_2 and the lag 23 to the pollutants SO_2 and NO_2. These correspondences are compatible with the clustering derived in [17, Table 7]. The fitted GAM-PCA-VAR model with its goodness-of-fit measures are reported in [17] as well. We note that in this fitted model $f_{ij} = 0$ was chosen for all $j > 0$. In view of the above results the GAM-PCA-VAR model with link

$$\log \mu_t = \beta_0 + \sum_{i=1}^{p} \sum_{j \in \mathcal{I}_i} f_{ij}(Z_{i(t-j)})$$

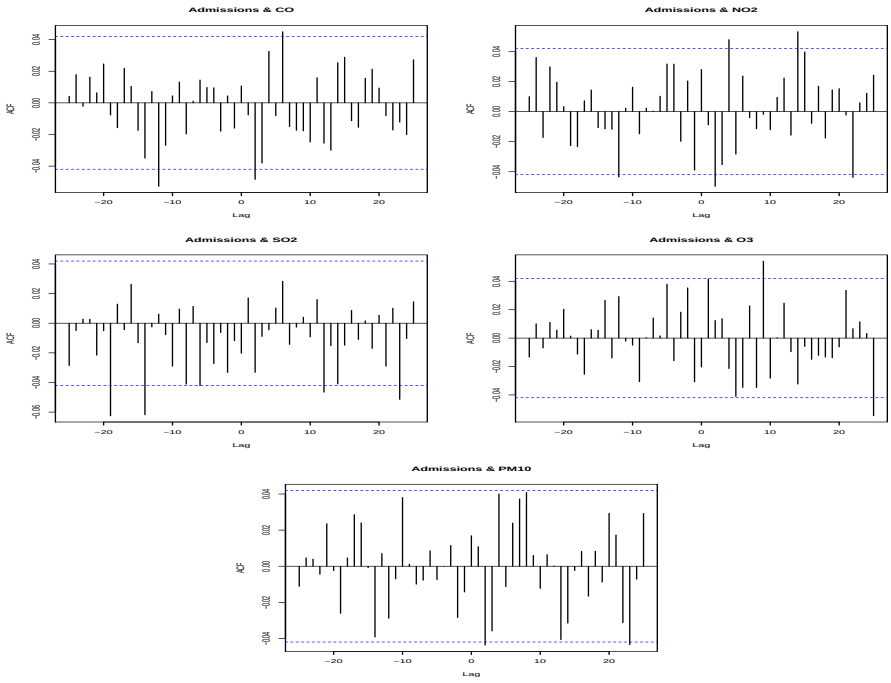

Fig. 4. Sample cross-correlation function (CCF) between the response and pollutant variables after the filtering.

can also be a possible candidate, where $\mathcal{I}_i$ denotes the set of lags which belong to the significant cross-correlation between the residuals of the response and the ith PC. This model can be fitted by using the procedure described in Section 2.

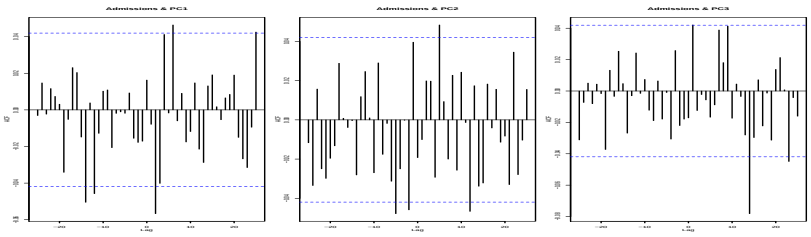

Fig. 5. Sample cross-correlation function (CCF) between the response and PCs after the filtering.

Table 3. Significant cross-correlations and their respective lags between the response variable RD and PCs after the filtering

	RD×PC1					RD×PC2				RD×PC3		
Lag	-14	-12	2	6	25	-5	-2	5	12	1	14	23
Value	-0.051	-0.046	-0.057	0.046	0.043	-0.048	-0.046	0.048	-0.047	0.042	-0.078	-0.045

6 Conclusions

A hybrid called GAM-PCA-VAR model composed by three statistical tools, the VAR model, PCA and the GAM, with Poisson marginal distribution, was developed in a more general framework than in [17]. A three-stage estimation method was proposed and studied by simulation for some examples. Some theoretical properties were also proved. The model was applied to describe the dependence between the number of hospital admissions for respiratory diseases and air pollutant covariates.

An extension of the proposed estimation method for the GAM-PCA-VAR model by a variable selection procedure which ensures that only the significant PCs with their respective lags are involved into the model will be pursed in future works.

Acknowledgments

The authors thank the following agencies for their support: the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq), the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES), Espírito Santo State Research Foundation (Fundação de Amparo à Pesquisa do Espírito Santo - FAPES) and Minas Gerais State Research Foundation (Fundação de Amparo à Pesquisa do estado de Minas Gerais - FAPEMIG). Márton Ispány was supported by the EFOP-3.6.1-16-2016-00022 project.The project is co-financed by the European Union and the European Social Fund. Pascal Bondon thanks to the Institute for Control and Decision of the Université Paris-Saclay.

References

1. Al-Osh M. A., Alzaid A. A.: First-order integer valued autoregressive (INAR(1)) process. J. Time Ser. Anal. 8, 261–275 (1987)
2. Barczy M., Ispány M., Pap G., Scotto M. G., Silva M. E.: Additive outliers in INAR(1) models. Stat. Pap. 53, 935–949 (2012)
3. Benjamin, M. A., Rigby, R. A., Stasinopoulos, D. M.: Generalized autoregressive moving average models. J. Amer. Statist. Assoc. 98, 214–223 (2003)
4. Brockwell, P. J., Davis, R. A.: Time Series: Theory and Methods. Springer Series in Statistics. New York, Springer-Verlag (1991)

5. Chen, R. J., Chu C., Tan, J., Cao, J., Song, W., Xu, X., Jiang, C., Ma W., Yang, C., Chen, B., Gui, Y., Kan, H.: Ambient air pollution and hospital admission in Shanghai, China. J. Hazard. Mater. 181, 234–240 (2010)
6. Davis, R. A., Dunsmuir, W. T. M., Streett, S. B.: Observation-driven models for Poisson counts. Biometrika 90, 777–790 (2003)
7. Dionisio, K. L., Chang, H. H., Baxter, L. K.: A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. Environ. Health 15:114 (2016)
8. Durbin, J., Koopman, S. J.: Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. J. Roy. Stat. Soc. B 62, 3–56. (2000)
9. Ferland, R., Latour, A., Oraichi, D.: Integer-valued GARCH process. J. Time Ser. Anal. 27(6), 923–942 (2006)
10. Gamerman, D., Santos, T. R., Franco, G. C.: A non-Gaussian family of state-space models with exact marginal likelihood. J. Time Ser. Anal. 34, 625–645 (2013)
11. Hastie, T. J., Tibshirani, R. J.: Generalized Additive Models. London, Chapman and Hall (1990)
12. Jolliffe, I. T.: Principal Component Analysis. 2nd edn. New York, Springer (2002)
13. Nascimento, A. P., Santos, J. M., Mil, J. G., de Souza, J. B., Reis Júnior, N. C., Reisen, V. A.: Association between the concentration of fine particles in the atmosphere and acute respiratory diseases in children. Rev. Saude Publ. 51:3 (2017)
14. Ostro, B. D., Eskeland, G. S., Sánchez, J. M., Feyzioglu, T.: Air pollution and health effects: A study of medical visits among children in Santiago, Chile. Environ. Health Persp. 107, 69–73 (1999)
15. Roberts, S., Martin, M.: Using supervised principal components analysis to assess multiple pollutant effects. Environ. Health Persp. 114(12), 1877–1882 (2006)
16. Schwartz, J.: Harvesting and long term exposure effects in the relationship between air pollution and mortality. Am. J. Epidemiol. 151, 440–448 (2000)
17. de Souza, J. B., Reisen, V. A., Franco, G. C., Ispány, M., Bondon, P., Santos, J. M.: Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. J. Roy. Stat. Soc. C-App., DOI: 10.1111/rssc.12239, (2017)
18. Souza, J. B., Reisen, V. A, Santos, J. M., Franco, G. C.: Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. Rev. Saude Publ. 48(3), 451–8 (2014)
19. Wood, S. N.: Generalized Additive Models: An Introduction with R. 2nd edn. Chapman and Hall/CRC (2017)
20. Zamprogno, B.: PCA in time series with short and long-memory time series. PhD Thesis at the Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico, UFES, Vitória, Brazil. (2013)
21. Wang, Y., Pham, H.: Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. Int. J. Syst. Assur. Eng. Manag. 2, 253–259 (2011)
22. Zhao, J., Cao, J., Tian, S., Chen, Y., Zhang, Sh., Wang, Zh., Zhou, X.: A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices. Aquat. Ecol. 48, 297–312 (2014)