Principal component analysis with autocorrelated data

Bartolomeu Zamprogno, Valdério Reisen, Pascal Bondon, Higor Cotta, Neyval Costa Reis Júnior

Principal component analysis with autocorrelated data

Bartolomeu Zamprogno[a,b], Valdério A. Reisen[a,b,c], Pascal Bondon[c], Higor H. Aranda Cotta[a,b,c] and Neyval C. Reis, Jr[a]

[a]NuMEs - DEST - CCE, Federal University of Espírito Santo - Brazil;[b] NuMEs - DEST - PGGEA - Federal University of Espírito Santo - Brazil; [c]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France

ARTICLE HISTORY
Compiled August 10, 2021

ABSTRACT
This paper contributes to the analysis, interpretation and the use of the principal component analysis (PCA) in a multivariate time-correlated linear process. The effect of ignoring the autocorrelation structure of the vector process is investigated theoretically and empirically. The results show a spurious impact of the time-correlation on the eigenvalues. To mitigate this impact, a pre-filtering procedure to whiten the data is applied. The methodology is used to identify redundant particulate matter (PM) measurements in a densely populated region in Brazil. Among the eight considered monitoring stations, it is found that three are needed to characterize the dynamic of the pollutant in the region.

KEYWORDS
Principal component analysis; autocorrelation; cross-correlation; eigenvalue; air pollution

1. Introduction

PCA is one of the most widely used multivariate techniques to reduce the dimension of a data set while keeping most of the variability of the data. To clarify how important this technique is, Richman (1986) has shown that between 1983 and 1985 over 60 PCA applications, or similar techniques have appeared in the meteorological/climatological journals. More recently, between 1999 and 2000, 53 of the 215 articles of the International Journal of Climatology have applied PCA. This represents 25%, a rate not achieved by any other statistical technique Jollife (2002, page 71).

The use of PCA goes beyond reducing the dimension of data. For example, Karar and Gupta (2007) have used PCA as a grouping tool of pollution sources, and Romero et al. (1999), White et al. (1991) and Cohen (1983) have applied PCA to identify homogeneous sub-regions of climatic stations in a large geographical area. Besides the use of PCA as a classification tool, several studies have used the technique to extenuate the multicollinearity in a regression analysis context and to detect outliers, see e.g., Liu (2009), Wang and Pham (2011a), Souza et al. (2014), Souza et al. (2018) and Reisen et al. (2019). PCA has also been used as a step procedure in other multivariate techniques such as factor analysis, canonical correlation analysis, and discriminant analysis, see e.g., Jollife (2002, Chapter 9). For example, in the financial area, Matteson and

Tsay (2011) have proposed a PCA based approach to modeling the conditional mean vector and conditional covariance matrix of a stationary multivariate autoregressive and conditionally heteroscedastic time series. Hu and Tsay (2014) have extended the idea of PCA to principal volatility component analysis with a focus on the dynamic dependence of volatility.

In the statistical control process area, Vanhatalo and Kulahci (2016) have illustrated the impact of the autocorrelation on the descriptive ability of PCA and on the monitoring process control. Vanhatalo et al. (2017) have proposed a driven method to determine the maximum number of lags in dynamic PCA in multivariate time series analysis and a method for determining the number of principal components (PCs) to retain. In the high-dimension setting, Hellton and Thoresen (2014) have addressed the problem of the impact of measurement error on PCA.

In the domain of air quality monitoring, the identification of pollution sources using PCA has been considered by many authors. For example, in the network management context, Pires et al. (2008a,b) have used PCA with monitored pollutant concentrations to manage the monitoring network of the metropolitan area of Porto (Portugal) to reduce costs. The authors have proposed to select only one station among those belonging to a same cluster and having similar concentrations behaviours. They have concluded that six stations instead of ten are sufficient to measure the level of concentration of sulphur dioxide (SO_2), and no more than two stations are required for monitoring the PM less than 10 μm in diameter (PM_{10}). Lu et al. (2011) have evaluated the performance of PCA and cluster analysis for the management of the local air quality monitoring network of Hong Kong (China) with the aim to identify city areas with similar air pollution behaviours and to locate emission sources. They have found that the monitoring stations could be grouped into different classes based on air pollution behaviours.

One of the usual assumptions of PCA is that the data are independent in time. Nevertheless, PCA has been widely used with time series which are time-correlated, without justification. For example, the pollution data considered in the above cited papers are time-dependent. Not taking into account the time-dependent structure of the data may lead to misleading analysis and interpretations. It is essential to recognize that neglecting the required data assumption when using standard statistical methods like PCA may produce biased estimates and spurious results see e.g., Vanhatalo and Kulahci (2016).

The effect of time-correlation on model estimation using PCA is also one of the main contribution of Souza et al. (2018), where the multicollinearity issue when using pollutants as covariates in the generalized additive model is solved using PCA, and where it is suggested to use a multivariate time series model to remove the temporal correlation of the covariates. Following similar lines, Melo (2015) and Melo et al. (to appear) have considered PCA in a logistic regression model to quantify the association between the pollutants and perceived annoyance. The methodologies proposed in these three papers were mainly based on the theoretical results discussed in Zamprogno (2013). Wang and Pham (2011b) have also considered PCA in the regression model to quantify the relationship between morbidity and pollutants; however, the temporal correlation of the variables was ignored by the authors.

The purpose of this paper is to generalize the use of PCA, mainly developed for independent observations, to multivariate time series. The effect of different correlation structures of multivariate stationary processes on the interpretation and inference of the PCs is illustrated. The study is justified empirically and theoretically, and a real data set of pollutant concentrations is considered as an example of application. Due

to the serial correlation in the data, the PCs are shown to be autocorrelated and cross-correlated. Thus, this paper suggests to pre-whiten the data with a linear model to attenuate the time-correlation before applying PCA. This whitening technique has been considered by some authors in the econometric area, but without discussing the consequence of neglecting the temporal correlation. For example, Matteson and Tsay (2011) and Hu and Tsay (2014) applied vector autoregressive (VAR) models to remove the serial correlation of time series of stock returns before carrying out PCA of the residuals.

The manuscript is structured as follows: Section 2 considers the time series model and theoretical properties of PCA with autocorrelated data. Monte Carlo simulations are addressed in Section 3. Section 4 discusses the real data application and Section 5 concludes the paper.

2. PCA with time series data

Let $X_t = [X_{1t}, \ldots, X_{kt}]'$, $t \in \mathbb{Z}$, be a k-dimensional linear process defined by

$$X_t = \mu + \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j}, \tag{1}$$

where $\mu \in \mathbb{R}^k$, $\varepsilon_t = [\varepsilon_{1t}, \ldots, \varepsilon_{kt}]'$ is a vector white noise process such that $\mathrm{E}(\varepsilon_t) = 0$ and

$$\Gamma_\varepsilon(h) = \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = \mathrm{E}(\varepsilon_t \varepsilon_{t+h}') = \begin{cases} \Sigma_\varepsilon & \text{if } h = 0, \\ 0 & \text{if } h \neq 0, \end{cases} \tag{2}$$

Σ_ε is a nonsingular matrix, and the Ψ_j's are $k \times k$ matrices of real coefficients satisfying $\Psi_0 = I$, I being the identity matrix, and $\sum_{j=0}^{\infty} \mathrm{tr}(\Psi_j \Sigma_\varepsilon \Psi_j') < \infty$, where $\mathrm{tr}(A)$ denotes the trace of a square matrix A. It follows from (1) and (2) that X_t is a second-order stationary process with mean μ and covariance matrix

$$\Gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h}) = \mathrm{E}((X_t - \mu)(X_{t+h} - \mu)') = \sum_{j=0}^{\infty} \Psi_j \Sigma_\varepsilon \Psi_{j+h}', \tag{3}$$

for all $h \geq 0$. In the following, it is assumed without loss of generality that $\mu = 0$.

In the analysis of a multivariate data set, PCA looks for linear combinations of the components capturing the highest percentage of variation of the data. This technique depends exclusively on the covariance or the correlation matrix of the data, see, e.g., Jollife (2002). PCA is well suited for time-independent observations since it explains only the contemporaneous correlation of the data and does not take into account the time-correlation. Specifically, PCA calculates the characteristic roots and vectors of $\Gamma_X(0)$. Let $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$ be the non necessarily distinct eigenvalues of $\Gamma_X(0)$ with corresponding orthonormal (with respect to the usual inner product) eigenvectors $p_1, \ldots, p_k$ ($p_i' p_i = 1$ and $p_i' p_j = 0$ when $i \neq j$). Then $\Gamma_X(0) p_i = \lambda_i p_i$ for $i = 1, \ldots, k$, and $P' \Gamma_X(0) P = \Lambda$ where P is the $k \times k$ matrix whose ith column is p_i and Λ is the $k \times k$ diagonal matrix whose ith diagonal element is λ_i, i.e., $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, see e.g., Banerjee and Roy (2014, Theorem 11.27). Equivalently, $\Gamma_X(0)$ admits the

so-called spectral decomposition

$$\Gamma_X(0) = P\Lambda P' = \sum_{i=1}^{k} \lambda_i p_i p_i'. \tag{4}$$

The PC vector process is given by $Y_t = P'X_t$, i.e., $Y_t = [Y_{1t}, \ldots, Y_{kt}]'$ where $Y_{it} = p_i'X_t$ for $i = 1, \ldots, k$. The following proposition summarizes some properties of the covariance of Y_t.

Proposition 2.1. Let X_t be defined by (1), $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$ be the eigenvalues of $\Gamma_X(0)$ with corresponding orthonormal eigenvectors $p_1, \ldots, p_k$, and $Y_{it} = p_i'X_t$ be the ith PC for $i = 1, \ldots, k$. Then,

a) $\mathrm{Var}(Y_{it}) = p_i'\Gamma_X(0)p_i = \lambda_i$,
b) $\mathrm{Cov}(Y_{it}, Y_{jt}) = p_i'\Gamma_X(0)p_j = 0$ when $i \neq j$,
c) $\mathrm{Cov}(Y_{it}, Y_{j(t+h)}) = p_i' \mathrm{Cov}(X_t, X_{t+h}')p_j = p_i'\Gamma_X(h)p_j$ for $i, j = 1, \ldots, k$ and $h \neq 0$.

Proof. a) and b) follow directly from (4), and c) results from $Y_{it} = p_i'X_t$. □

Remark 1. Propositions 2.1(a),(b) appear in Anderson (2003) and are the particular cases of an uncorrelated process, i.e, when $X_t = \varepsilon_t$ in (1). Proposition 2.1(c) shows that the autocovariances ($i = j$) and the cross-covariances ($i \neq j$) of the PCs are non-zero. This induces some issues discussed below in descriptive and inferential procedures of PCA in the case of time series.

Remark 2. If some eigenvalues are equal, the corresponding eigenvectors and PCs are not uniquely defined. Nevertheless, the vector space generated by these eigenvectors is unique, see e.g., Harville (1997, pages 537–538).

Remark 3. The properties of PCs discussed here are based on the eigenvalues and eigenvectors of the covariance matrix of X_t. These properties are still valid for the eigenvalues and eigenvectors obtained from the autocorrelation matrix. It is quite common in practice to compute PCA based on the autocorrelation matrix, especially when the unities and the variances of the variables are different. However, this is not the case in the application problem discussed in this paper. One of the advantages of using sample PCs based on the covariance matrix is that the statistical inferences for the population PCs are easier than those of the sample PCs obtained from the correlation matrix. For a discussion of the advantages and disadvantages of using sample PCA of the covariance matrix instead of the autocorrelation matrix, see Jollife (2002, Chapter 2).

Remark 4. Let X_t be defined by (1). It follows from (3) that

$$\mathrm{tr}(\Gamma_X(0)) = \mathrm{tr}(\Sigma_\varepsilon) + \mathrm{tr}(\sum_{j=1}^{\infty} \Psi_j \Sigma_\varepsilon \Psi_j'). \tag{5}$$

Let $A_n = \sum_{j=1}^{n} \Psi_j \Sigma_\varepsilon \Psi_j'$, and A_n^{ij} be the (i,j)th element of A_n for $1 \leq i, j \leq k$. Then,

$$\operatorname{tr}(\lim_{n\to\infty} A_n) = \sum_{i=1}^{k} (\lim_{n\to\infty} A_n^{ii}) = \lim_{n\to\infty} \sum_{i=1}^{k} A_n^{ii} = \lim_{n\to\infty} \operatorname{tr}(A_n). \tag{6}$$

Since A_n is a nonnegative definite matrix, $\operatorname{tr}(A_n) \geq 0$. Then, $\lim_{n\to\infty} \operatorname{tr}(A_n) \geq 0$, and we deduce from (5) and (6) that $\operatorname{tr}(\Gamma_X(0)) \geq \operatorname{tr}(\Sigma_\varepsilon)$. Now,

$$\operatorname{tr}(\Gamma_X(0)) = \operatorname{tr}(P\Lambda P') = \operatorname{tr}(\Lambda) = \operatorname{tr}(\Gamma_Y(0)) = \sum_{i=1}^{k} \lambda_i.$$

Therefore, the PCs of X_t present more variability than the ones of ε_t. This can lead to a wrong use of PCA technique if the time-correlation of X_t is ignored.

A parametric class of models satisfying (1) is the k-dimensional vector seasonal autoregressive moving average (VSARMA) process with non-seasonal orders p and q, seasonal orders P and Q, and season $s \in \mathbb{N} - \{0\}$. This process is defined by the difference equation

$$\phi(B)\Phi(B^s)X_t = \theta(B)\Theta(B^s)\varepsilon_t, \tag{7}$$

where ε_t is a vector white noise with $\mathrm{E}(\varepsilon_t) = 0$ and $\Gamma_\varepsilon(h)$ given by (2), and B is the backward operator, i.e., $BX_t = X_{t-1}$ for any process X_t. The matrix-valued polynomials $\phi(\cdot)$, $\theta(\cdot)$, $\Phi(\cdot)$ and $\Theta(\cdot)$ given by

$$\begin{aligned}
\phi(z) &= I - \phi_1 z - \cdots - \phi_p z^p, \\
\theta(z) &= I + \theta_1 z + \cdots + \theta_q z^q, \\
\Phi(z) &= I - \Phi_1 z - \cdots - \Phi_P z^P, \\
\Theta(z) &= I + \Theta_1 z + \cdots + \Theta_Q z^Q,
\end{aligned}$$

satisfy that $\det(\phi(z)\Phi(z^s)) \neq 0$ and $\det(\theta(z)\Theta(z^s)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. These two conditions are known as the causality and invertibility properties, respectively. Additional conditions have to be imposed in order to obtain an identifiable model, see e.g. Brockwell and Davis (1991, page 431) and Reinsel (1997, section 2.3). In (7), the matrix parameters ϕ_i's, θ_i's, Φ_i's and Θ_i's are unknown and have to be estimated from the observed data $X_1, \ldots, X_n$.

The VSARMA process has a short-memory correlation structure in the sense that the sequence of matrices $\Gamma_X(h)$ for $h \in \mathbb{Z}$ is summable. The vector seasonal autoregressive fractionally integrated moving average (VSARFIMA) process is a linear process defined by an extension of the difference equation (7). This process has a long-memory behaviour in the sense that the matrices $\Gamma_X(h)$ are only square summable, see Chung (2012). A VSARFIMA model is used in Section 4.

As mentioned in Remark 4, when X_t is time-correlated, the PCs of X_t have larger variances than the ones of ε_t. One way to mitigate this effect is to apply to X_t a multivariate linear filter, such as the VSARMA filter before applying PCA. In this context, PCA is applied to ε_t in place of X_t in (7).

The VAR(1) model is the particular case of (7) where X_t satisfies the difference equation $X_t = \Phi X_{t-1} + \varepsilon_t$ with Φ a matrix parameter. This model is widely used in

modelling multivariate time series. Proposition 2.2 illustrates the effect of temporal correlation on the PCs Y_t when X_t is a VAR(1) process. This result can be extended to more general processes. For example, this is well-known that the VAR(p) model can be written as a VAR(1) process see e.g., Lutkepohl (2005, page 15) and Hamilton (1994, page 259).

Proposition 2.2. Let X_t be a stationary VAR(1) process. Then $\Gamma_X(h) = \Gamma_X(0)(\Phi^h)'$ and $\Gamma_Y(h) = \Lambda P'(\Phi^h)'P$ for all $h \geq 0$.

Proof. It follows from Brockwell and Davis (1991, Example 11.3.1) that $\Psi_j = \Phi^j$ in (1). Then (3) implies that $\Gamma_X(h) = \Gamma_X(0)(\Phi^h)'$ for all $h \geq 0$. Since $Y_t = P'X_t$, $\Gamma_Y(h) = P'\Gamma_X(h)P = P'\Gamma_X(0)(\Phi^h)'P = P'P\Lambda P'(\Phi^h)'P = \Lambda P'(\Phi^h)'P$ for all $h \geq 0$. □

Remark 5. Consider the particular VAR(1) process where $\Phi = \text{diag}(\phi_1, \ldots, \phi_k)$ with $|\phi_i| < 1$ for $i = 1, \ldots, k$. Then, it results from (3) that the (i,j)th element of $\Gamma_X(h)$, $\Gamma_X^{ij}(h)$, is given by

$$\Gamma_X^{ij}(h) = \sum_{l=0}^{\infty} \phi_i^l \, \Sigma_\varepsilon^{ij} \, \phi_j^{l+h} = \phi_j^h/(1 - \phi_i\phi_j)\Sigma_\varepsilon^{ij}, \tag{8}$$

for all $h \geq 0$. Therefore,

$$\text{tr}(\Gamma_Y(0)) = \text{tr}(\Gamma_X(0)) = \text{tr}(\Sigma_\varepsilon) + \sum_{i=1}^{k} \phi_i^2/(1 - \phi_i^2)\Sigma_\varepsilon^{ii}. \tag{9}$$

It follows from (9) that the variability of the PCs of X_t increases as $|\phi_i|$ increases, and may be much larger than the one of the PCs of ε_t. Furthermore, since $\Gamma_Y(h) = \Lambda P'(\Phi^h)'P$, its (i,j)th element, $\Gamma_Y^{ij}(h)$, is given by

$$\Gamma_Y^{ij}(h) = \lambda_i \sum_{l=1}^{k} \phi_l^h \, p_{li} \, p_{lj}, \tag{10}$$

for all $h \geq 0$, where $p_i = [p_{1i}, \ldots, p_{ki}]'$.

Suppose that $\phi_i = \phi$ for $i = 1, \ldots, k$. Then, $\Gamma_X(h) = \phi^h\Gamma_X(0) = \phi^h/(1 - \phi^2)\Sigma_\varepsilon$ for all $h \geq 0$. The eigenvectors of $\Gamma_X(h)$ and Σ_ε are the same, while the eigenvalues of $\Gamma_X(h)$ are the ones of Σ_ε multiplied by $\phi^h/(1 - \phi^2)$. We have $\Gamma_Y(h) = \Lambda P'(\Phi^h)'P = \phi^h\Lambda$. Then, when Σ_ε is not diagonal, the components of X_t are cross-correlated, while the components of the PCs are not, for all $h \geq 0$. Observe, using (10), that the components of the PCs are generally cross-correlated when the parameters ϕ_i's are not all equal.

The VMA(1) model is the particular case of (7) where X_t satisfies the difference equation $X_t = \varepsilon_t + \Theta\varepsilon_{t-1}$ with Θ a matrix parameter. Proposition 2.3 gives the expressions of $\Gamma_X(h)$ and $\Gamma_Y(h)$ when X_t is a VMA(1) process. As for the VAR(1) model, this result can be extended to more complicated processes.

Proposition 2.3. Let X_t be a VMA(1) process where all the eigenvalues of Θ are

less than one in modulus. Then

$$\Gamma_X(h) = \begin{cases} \Sigma_\varepsilon + \Theta\Sigma_\varepsilon\Theta' & \text{if } h = 0, \\ \Sigma_\varepsilon\Theta' & \text{if } h = 1, \\ 0 & \text{if } h > 1, \end{cases} \quad \text{and} \quad \Gamma_Y(h) = \begin{cases} \Lambda & \text{if } h = 0, \\ P'\Sigma_\varepsilon\Theta'P & \text{if } h = 1, \\ 0 & \text{if } h > 1. \end{cases}$$

Proof. The expression of $\Gamma_X(h)$ follows from the difference equation $X_t = \varepsilon_t + \Theta\varepsilon_{t-1}$. On the other hand, $\Gamma_Y(h) = P'\Gamma_X(h)P$ for $h \in \mathbb{Z}$. □

Remark 6. Consider the particular VMA(1) process where $\Theta = \text{diag}(\theta_1, \ldots, \theta_k)$ with $|\theta_i| < 1$ for $i = 1, \ldots, k$. We deduce from Proposition 2.3 that

$$\Gamma_X^{ij}(h) = \begin{cases} (1 + \theta_i\theta_j)\Sigma_\varepsilon^{ij} & \text{if } h = 0, \\ \theta_j\Sigma_\varepsilon^{ij} & \text{if } h = 1, \\ 0 & \text{if } h > 1. \end{cases}$$

Therefore,

$$\text{tr}(\Gamma_Y(0)) = \text{tr}(\Gamma_X(0)) = \text{tr}(\Sigma_\varepsilon) + \sum_{i=1}^{k} \theta_i^2\Sigma_\varepsilon^{ii}. \tag{11}$$

It follows from (11) that $\text{tr}(\Sigma_\varepsilon) \leq \text{tr}(\Gamma_Y(0)) \leq 2\,\text{tr}(\Sigma_\varepsilon)$.

If, $\theta_i = \theta$ for $i = 1, \ldots, k$, Σ_ε, $\Gamma_X(0)$ and $\Gamma_X(1)$ have the same eigenvectors, while the eigenvalues of $\Gamma_X(0)$ and $\Gamma_X(1)$ are the ones of Σ_ε multiplied by $1 + \theta^2$ and θ, respectively. Furthermore, in this case, we deduce from Proposition 2.3 that $\Gamma_Y(1) = \theta P'\Sigma_\varepsilon P = \theta/(1 + \theta^2)P'\Gamma_X(0)P = \theta/(1 + \theta^2)\Lambda$. Then, the components of the PCs are not cross-correlated for all $h \geq 0$.

In practice, $\Gamma_X(0)$ is unknown and must be estimated from a set of observations $X_1, \ldots, X_n$ of X_t. The sample estimate of $\Gamma_X(0)$ is

$$\hat{\Gamma}_X(0) = \frac{1}{n}\sum_{t=1}^{n} X_t X_t', \tag{12}$$

$\hat{\Gamma}_X(0)$ is symmetric and non-negative definite with spectral decomposition

$$\hat{\Gamma}_X(0) = BLB', \tag{13}$$

where $L = \text{diag}(l_1, \ldots, l_k)$, $l_1 \geq \ldots \geq l_k \geq 0$ are the eigenvalues of $\hat{\Gamma}_X(0)$, and B is an orthonormal matrix whose ith column b_i is an eigenvector associated to l_i for $i = 1, \ldots, k$. Each eigenvalue l_i is an estimate of λ_i. Suppose that the eigenvalues of $\Gamma_X(0)$ are distinct, i.e., $\lambda_1 > \ldots > \lambda_k$. In this case, P is unique in (4). Let $D = \sqrt{n}(L - \Lambda)$ and $G = \sqrt{n}(B - P)$. Under additional assumptions, Taniguchi and Krishnaiah (1987, Theorem 1) have shown that for model (1), the joint distribution of D and G converges as n tends to infinity. If X_t is Gaussian, then the limiting joint distribution of D and G is normal with D and G independent and the diagonal elements of D are independent.

A major concern about using PCA is how many PCs should be selected. Several criteria have been proposed in the literature such as the eigenvalues plot of Jollife (2002)

and the mean eigenvalue test of Perez-Neto et al. (2005). Assume that the random variables X_t are mutually independent and identically distributed with finite moments and $\lambda_1 > \ldots > \lambda_k > 0$. Fujikoshi (1980, Theorem 1) has generalized Anderson (2003, Theorem 13.5.1) to non Gaussian data and has shown that $\sqrt{n}(l_i - \lambda_i)$ has the limiting normal distribution $\mathrm{N}(0, 2\lambda_i^2 + \kappa_4^i)$, where κ_4^i is the fourth-order cumulant of the ith component X_{it} of X_t for all $i = 1, \ldots, k$. Therefore, an asymptotic confidence interval (ACI) of significance level α for λ_i is given by

$$l_i - \sqrt{\frac{2l_i^2 + \hat{\kappa}_4^i}{n}}\, z_{\frac{\alpha}{2}} \leq \lambda_i \leq l_i + \sqrt{\frac{2l_i^2 + \hat{\kappa}_4^i}{n}}\, z_{\frac{\alpha}{2}}, \tag{14}$$

where $\hat{\kappa}_4^i$ is the sample estimate of κ_4^i, $F(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$ and F is the cumulative distribution function of the $\mathrm{N}(0, 1)$ random variable. Now, let $\tau_m = (\lambda_1 + \cdots + \lambda_m)/(\lambda_1 + \cdots + \lambda_k)$ be the fraction of the variance explained by the first m PCs, where $1 \leq m < k$, and $R_m = (l_1 + \cdots + l_m)/(l_1 + \cdots + l_k)$ be an estimate of τ_m. Fujikoshi (1980, Theorem 3) implies that $\sqrt{n}(R_m - \tau_m)$ has the limiting normal distribution $\mathrm{N}(0, \sum_{i=1}^k T_i^2(2\lambda_i^2 + \kappa_4^i))$, where $T_i = (c_i - \tau_m)/(\lambda_1 + \cdots + \lambda_k)$, and $c_i = 1$ for $i = 1, \ldots, m$, $c_i = 0$ for $i = m + 1, \ldots, k$. Therefore, an ACI of significance level α for τ_m is

$$R_m - \sqrt{\frac{\sum_{i=1}^k \hat{T}_i^2(2l_i^2 + \hat{\kappa}_4^i)}{n}}\, z_{\frac{\alpha}{2}} \leq \tau_m \leq \; R_m + \sqrt{\frac{\sum_{i=1}^k \hat{T}_i^2(2l_i^2 + \hat{\kappa}_4^i)}{n}}\, z_{\frac{\alpha}{2}}, \tag{15}$$

where $\hat{T}_i = (c_i - R_m)/(l_1 + \cdots + l_k)$.

3. Numerical experiments

This section presents finite sample size studies to illustrate the effect of time-correlation on the eigenvalues of $\Gamma_X(0)$ and on the interpretation of PCA. For this purpose, we consider VAR(1) processes with different correlation structures. The calculus and simulations were coded with R Core Team (2019, Version 3.6.2) and are available upon request.

Let $X_t = \Phi X_{t-1} + \varepsilon_t$, where matrix Σ_ε is given by

$$\Sigma_\varepsilon = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{16}$$

and the matrix parameters Φ are displayed in Table 1. The correlation structures of X_t depend on Φ. The white noise model $X_t = \varepsilon_t$ is denoted by Model 1. Since the matrices of the parameters of Models 2 and 3 are diagonal and Σ_ε is also diagonal, it follows from (8) that the covariance matrices of these models are also diagonal and have the same eigenvectors which correspond to the natural basis of $\mathbb{R}^4$. Since all ϕ_i's are equal in Model 2, the eigenvalues of $\Gamma_X(0)$ in Models 1 and 2 are proportional, which is not the case in Models 1 and 3.

Contrarily to the three first models, Models 4 and 5 present cross-correlations between the components of X_t at different lags h. According to Proposition 2.2,

$\Gamma_X(h) = \Gamma_X(0)(\Phi^h)'$ for all $h \geq 0$. Therefore, if the entries of $\Gamma_X(0)$ are nonnegative, large positive entries of Φ implies large positive cross-covariances. In this sense, Model 5 presents stronger cross-covariances than Model 4. These correlation structures may seriously affect the analysis and interpretation of the PCA. In particular, a significant impact occurs when using Models 4 and 5, which have large positive degrees of the correlations. These issues are discussed as follows.

The covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 2 to 5 are displayed in Table 2. As expected, Model 5 displays the largest covariances. For each model, it can be seen that $\text{tr}(\Gamma_X(0)) \geq \text{tr}(\Sigma_\varepsilon)$, as mentioned in Remark 4.

Table 1.: Matrix parameters Φ of VAR(1) in Models 2 to 5.

Model 2				Model 3			
0.3	0.0	0.0	0.0	0.8	0.0	0.0	0.0
0.0	0.3	0.0	0.0	0.0	0.5	0.0	0.0
0.0	0.0	0.3	0.0	0.0	0.0	0.3	0.0
0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.3
Model 4				Model 5			
0.3	0.0	0.1	0.1	0.3	0.5	0.7	0.4
0.0	0.3	0.0	0.0	0.0	0.3	0.0	0.0
0.2	0.0	0.3	0.0	0.0	0.0	0.3	0.0
0.0	0.1	0.0	0.3	0.8	0.6	0.0	0.3

Table 2.: Covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 2 to 5.

Model 2				Model 3			
10.99	0.00	0.00	0.00	27.78	0.00	0.00	0.00
0.00	5.49	0.00	0.00	0.00	6.67	0.00	0.00
0.00	0.00	3.30	0.00	0.00	0.00	3.30	0.00
0.00	0.00	0.00	1.10	0.00	0.00	0.00	1.01
Model 4				Model 5			
11.11	0.01	0.88	0.04	29.29	1.09	0.79	25.43
0.01	5.49	0.00	0.18	1.09	5.49	0.00	1.37
0.88	0.00	3.90	0.00	0.79	0.00	3.30	0.21
0.04	0.18	0.00	1.17	25.43	1.37	0.21	38.98

Table 3 shows, for each VAR(1) model, the eigenvalues λ_i's of $\Gamma_X(0)$ with their respective percentage of variability $\lambda_i/(\lambda_1 + \cdots + \lambda_4)$. As expected, Models 1 and 2 display the same percentages since the λ_i's are proportional. Model 3 presents more variability than Models 1 and 2 because λ_1 is much larger than the other eigenvalues. Since the parameters Φ of Models 2 and 4 are close, the associated eigenvalues of $\Gamma_X(0)$ and their percentages of variability are similar. A very distorted case of the percentages is observed between Model 2 and Model 5. The large positive cross-covariance in Model 5 drastically increases the variability of the eigenvalues of $\Gamma_X(0)$, and the first PC captures almost all the variability. This is a problem of high practical relevance, for example in the context of reducing the data dimension.

Now, more general VAR(1) models are considered in the study. The matrix Σ_ε

Table 3.: Eigenvalues of $\Gamma_X(0)$ of the VAR(1) Models 1 to 5 with their percentages of variability.

Model	λ_1	λ_2	λ_3	λ_4	% λ_1	% λ_2	% λ_3	% λ_4
1	10.00	5.00	3.00	1.00	52.63	26.32	15.79	5.26
2	10.00	5.49	3.30	1.10	52.63	26.32	15.79	5.26
3	27.78	6.67	3.30	1.01	71.68	17.20	8.51	2.61
4	11.21	5.50	3.79	1.16	51.73	25.39	17.51	5.37
5	60.09	8.29	5.44	3.24	77.98	10.75	7.06	4.21

becomes

$$\Sigma_\varepsilon = \begin{bmatrix} 127 & 30 & 47 & 62 \\ 30 & 58 & 33 & 70 \\ 47 & 33 & 64 & 58 \\ 62 & 70 & 58 & 172 \end{bmatrix},$$

and the white noise model $X_t = \varepsilon_t$ is denoted by Model 6. The matrix parameters Φ are displayed in Table 4. Note that some autoregressive parameters are negative, which implies that the models may produce negative autocorrelations. These negative correlations may lead to different impacts on the inferential analysis compared to the previous cases. The covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 7 to 10 are

Table 4.: Matrix parameters Φ of VAR(1) Models 7 to 10.

Model 7				Model 8			
0.2	0.0	0.0	0.0	-0.5	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
0.0	0.0	-0.5	0.0	0.0	0.0	-0.1	0.0
0.0	0.0	0.0	-0.3	0.0	0.0	0.0	0.9
Model 9				Model 10			
0.4	0.1	0.3	0.1	0.6	0.3	0.6	0.03
0.0	0.8	0.4	0.0	-0.1	0.2	-0.1	0.2
0.2	0.0	0.3	0.0	0.1	-0.8	0.4	0.5
0.0	0.0	0.6	-0.4	0.2	0.0	0.1	-0.5

presented in Table 5. For each model, we have $\text{tr}(\Gamma_X(0)) \geq \text{tr}(\Sigma_\varepsilon)$ in agreement with Remark 4. The trace of $\Gamma_X(0)$ represents the total variability of the PCs of X_t and increases from Model 6 to Model 10.

Table 6 shows the eigenvalues λ_i's of the matrices $\Gamma_X(0)$ for each VAR(1) model and their respective percentage of variability $\lambda_i/(\lambda_1 + \cdots + \lambda_4)$. Comparing with Table 3, it can be seen that the cross-covariances in Models 7 to 10 do not have drastic effects in the interpretation of PCA compared to Model 6. On the contrary, the percentages of variability are very stable across different correlation structures.

Samples of size $n = 1000$ of Models 6, 8, 9 and 10 with Gaussian innovations, were generated and the sample autocorrelation and cross-correlation functions (ACF and CCF) of the PCs were computed. The number of replications was 500. The mean of some of these quantities are displayed in Figures 1 and 2 for Models 6 and 8 and Models 9 and 10, respectively.

Table 5.: Covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 7 to 10.

Model 7				Model 8			
132.29	30.00	42.73	54.39	169.33	24.00	49.47	44.29
30.00	58.00	33.00	70.00	24.00	77.33	31.43	116.67
42.73	33.00	85.33	89.23	49.47	31.43	64.65	53.70
54.39	70.00	89.23	337.25	44.29	116.67	53.70	477.78
Model 9				Model 10			
240.04	193.95	104.52	81.18	575.20	44.82	183.86	120.35
193.95	399.10	110.20	101.54	44.82	74.72	43.80	46.26
104.52	110.20	94.66	72.40	183.86	43.80	175.62	42.00
81.18	101.54	72.40	203.96	120.35	46.26	42.00	234.47

Table 6.: Eigenvalues of $\Gamma_X(0)$ of the VAR(1) Models 6 to 10 with their percentages of variability.

Model	λ_1	λ_2	λ_3	λ_4	% λ_1	% λ_2	% λ_3	% λ_4
6	276.42	87.71	34.22	22.65	65.66	20.83	8.13	5.38
7	402.11	125.27	50.32	35.17	65.61	20.44	8.21	5.74
8	525.90	177.62	54.26	31.32	66.65	22.51	6.88	3.97
9	626.19	164.90	112.36	34.31	66.78	17.58	11.98	3.66
10	690.65	204.24	115.34	49.78	65.16	19.27	10.88	4.70

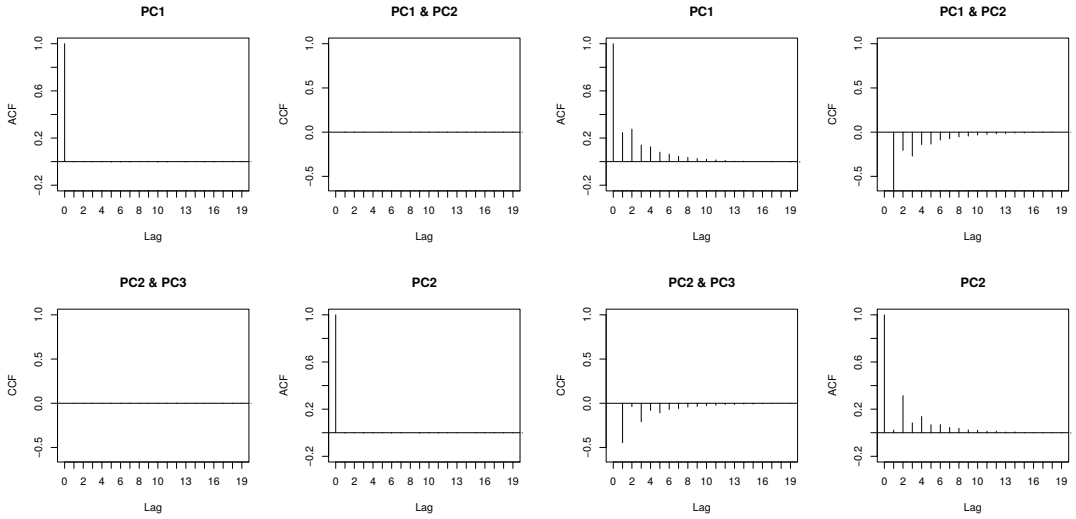
Figure 1a) shows that the PCs are neither autocorrelated nor cross-correlated in the case of a white noise. Figure 1b) shows that the PCs may be autocorrelated and cross-correlated when the matrix parameter Φ is diagonal but the diagonal elements are not all equal. These features become more clear for Models 9 and 10. Indeed, Figure 2 shows that the full correlation structure of the data is transferred to the PCs in the case of general matrices Φ and Σ_ε. These empirical evidences corroborate and illustrate Proposition 2.1.

The numerical experiments discussed in this section confirm that time-correlations in the vector X_t have impacts on PCA. Therefore, it is necessary to introduce procedures that allow the use of PCA with multivariate time-correlated data. This paper suggest to pre-processing the data with a multivariate linear filter in order to whiten the data before applying PCA. This is explored in the application Section. Note that transforming the data with linear filters to attenuate the temporal structure in multivariate techniques has been also addressed in the recent work of Jaimungal and Ng (2007), Greenaway-McGrevy et al. (2012) and Hu and Tsay (2014).

4. Application to PM_{10} data

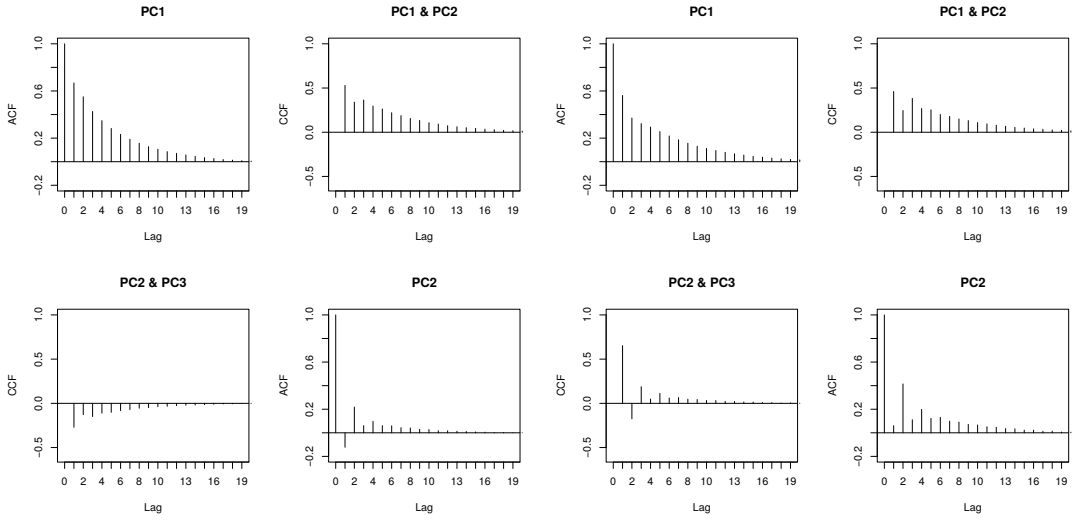
PCA is used here to identify cities areas with similar PM_{10} concentrations, without ignoring the time dependence of the data. We investigate whether or not the temporal correlation of the variables affects PCA and its interpretation. In general, this issue is not addressed in applied works, see e.g., Pires et al. (2008a,b). All the results in this section were also obtained with R Core Team (2019, Version 3.6.2).

The data set was collected at the automatic air quality monitoring network



(a) Model 6. (b) Model 8.

Figure 1.: ACF and CCF plots of some sample PCs of Models 6 and 8.



(a) Model 9. (b) Model 10.

Figure 2.: ACF and CCF plots of some sample PCs of Models 9 and 10.

(AAQMN) in the Greater Vitória Region (GVR) in Brazil. The eight monitoring stations are located at urban sites of four cities in the GVR. Additionally to PM_{10} concentrations, the AAQMN monitors the total suspended particles (TSP), ozone (O_3), nitrogen oxides (NO_x), carbon monoxide (CO), hydrocarbons (HC) and meteorological variables. The PM_{10} concentrations ($\mu g/m^3$) were measured in eight stations, from January 2005 to December 2009. The daily averages at the eight stations constitute the time series X_t which are plotted in Figure 3.

The sample ACF of each component of X_t are plotted in Figure 4. This figure shows a strong weekly seasonal behaviour which is expected with daily pollution data. In addition, the sample autocorrelations are positive and decrease slowly, which is typical of a long memory seasonal time series.

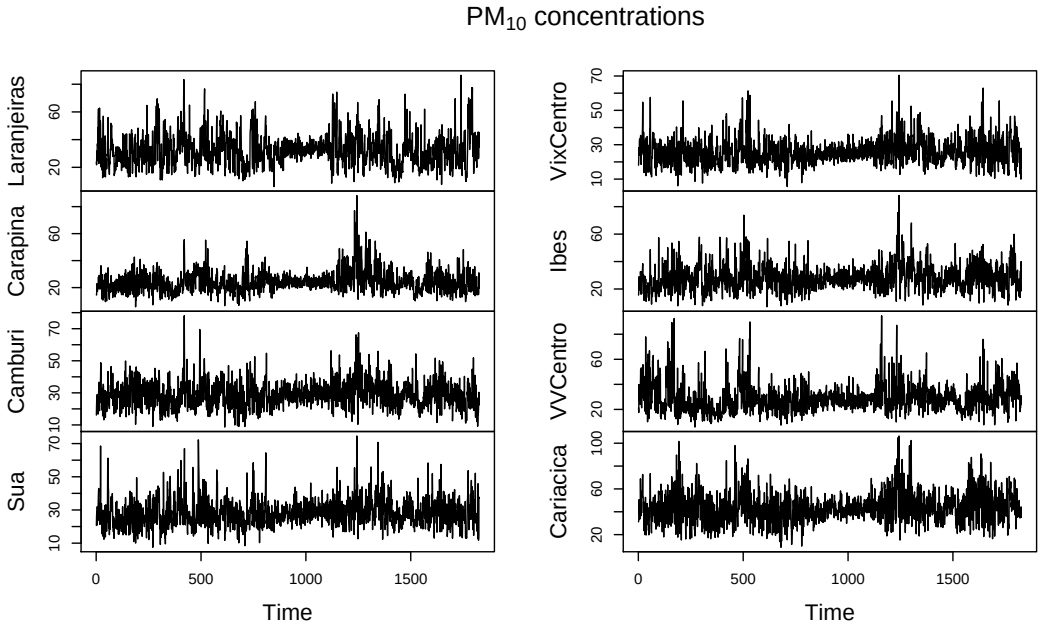


Figure 3.: Plots of the daily averages of the PM_{10} concentrations of the AAQMN.

We fit a VSARFIMA model with season $s = 7$ to X_t. The estimator suggested by Reisen et al. (2014) is used to estimate the fractional parameters at the lung-run (d) and at the seasonal period s=7 (D) with the bandwidth $m = n^{0.5}$. The estimates $(\hat{d}, \hat{D})$ and their standard deviation $(\hat{\sigma}(\hat{d}), \hat{\sigma}(\hat{D}))$ are displayed in Table 7. We see that these fractional parameters are significant for each station.

Table 7.: Fractional parameters estimates for PM_{10} data.

Station	$\hat{d}$	$\hat{\sigma}(\hat{d})$	$\hat{D}$	$\hat{\sigma}(\hat{D})$
Laranjeiras	0.2588	0.0019	0.1170	0.0093
Carapina	0.2792	0.0022	0.1787	0.0107
Camburi	0.2377	0.0079	0.2282	0.0393
Sua	0.2339	0.0048	0.0694	0.0240
VixCentro	0.2194	0.0027	0.1052	0.0132
Ibes	0.2801	0.0022	0.0512	0.0112
VVCentro	0.2832	0.0029	0.1270	0.0144
Cariacica	0.1992	0.0026	0.0844	0.0128

For each $i = 1, \ldots, 8$, we build the series $\hat{Z}_{it} = (1 - B)^{\hat{d}_i}(1 - B^s)^{\hat{D}_i} X_{it}$ and we fit a VSARMA model (7) to $\hat{Z}_t$. Following the standard methodology, we choose the orders (p, q, P, Q) with an information criterion, namely the bias-corrected Akaike information criterion (AICC), see Brockwell and Davis (1991, Section 9.2). This criterion selects a

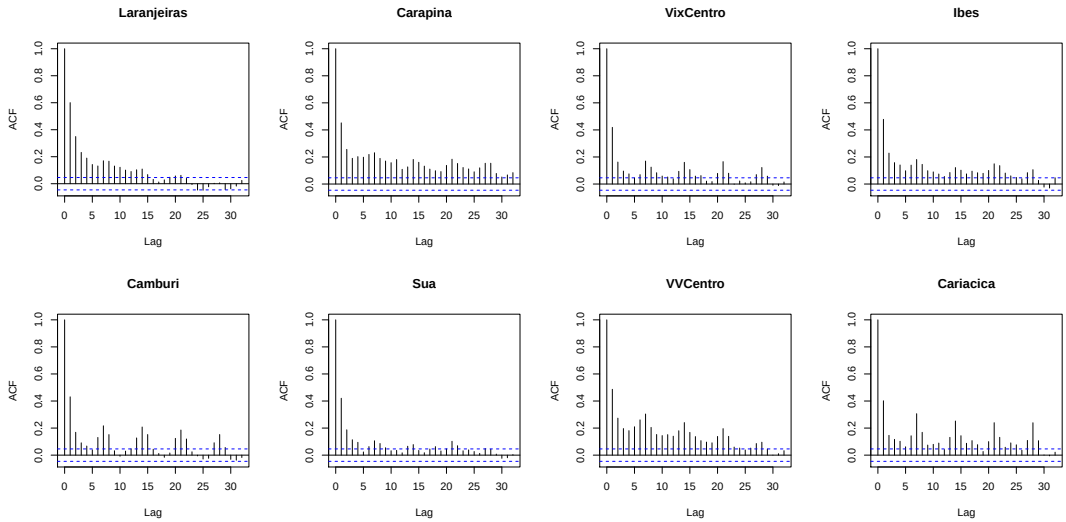


Figure 4.: Sample ACF of the daily average of the PM_{10} concentrations.

simple VAR(1) model with the following matrix parameter

$$\hat{\Phi} = \begin{bmatrix} 0.27 & -0.13 & 0.17 & -0.06 & -0.03 & 0.13 & -0.01 & 0.02 \\ 0.02 & -0.05 & 0.10 & 0.01 & -0.05 & -0.01 & 0.08 & 0.12 \\ 0.08 & -0.05 & 0.07 & -0.04 & 0.07 & 0.07 & -0.05 & 0.09 \\ 0.18 & -0.07 & 0.04 & 0.06 & 0.01 & 0.02 & 0.01 & 0.06 \\ 0.09 & -0.01 & 0.04 & 0.01 & 0.04 & -0.03 & 0.07 & 0.09 \\ 0.08 & -0.01 & 0.09 & -0.02 & -0.06 & 0.09 & 0.00 & 0.08 \\ 0.06 & 0.02 & 0.02 & -0.05 & 0.02 & -0.03 & 0.09 & 0.06 \\ 0.04 & 0.00 & 0.06 & 0.00 & -0.08 & 0.06 & 0.05 & 0.06 \end{bmatrix}.$$

Apart from the first diagonal element, all the coefficients of $\hat{\Phi}$ are quite small, which indicates that the fractional filtering giving $\hat{Z}_t$ extracts almost all the temporal correlation of X_t. Figure 5 plots the sample ACF of each component of the residual $\hat{\varepsilon}_t = \hat{Z}_t - \hat{\Phi}\hat{Z}_{t-1}$ and clearly shows that these components are white noises. Note that, even if the nondiagonal autoregressive parameters are very small, they should not be ignored in the use of PCA.

Now, we investigate the temporal correlation effect in the analysis and interpretation of PCA applied to PM_{10} data. The sample estimate $\hat{\Gamma}_X(0)$ of $\Gamma_X(0)$ is given by (12) and its spectral decomposition is (13). Let $\hat{\Gamma}_{\hat{\varepsilon}}(0) = (1/n)\sum_{t=1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_t'$ with the spectral decomposition $\hat{\Gamma}_{\hat{\varepsilon}}(0) = CMC'$, where $M = \text{diag}(m_1, \ldots, m_k)$, $m_1 \geq \ldots \geq m_k \geq 0$ are the eigenvalues of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, and C is an orthonormal matrix whose ith column c_i is an eigenvector associated to m_i for $i = 1, \ldots, k$.

As addressed in Remark 3, in practice, it is more common to compute the PCs based on the eigenvectors and eigenvalues derived from standardized variables, i.e, from the correlation matrix. This is the cases when the components of the vector X_t have distinct units and very different variances. The PM_{10} concentrations are measured with the same unit and have similar standard deviations; the minimum and maximum standard deviations are $7.4\,\mu g/m^3$ (Cariacica) and $13.12\,\mu g/m^3$ (Laranjeiras). In addition, the percentages of cumulative variation explained by the PCs obtained from the autocorrelation and autocovariance matrices are very close. For example, the cumulative

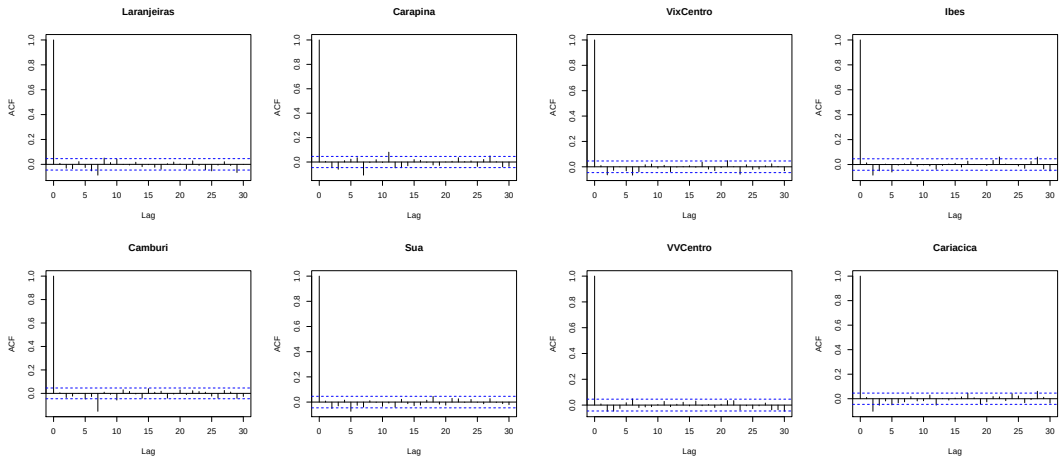


Figure 5.: Sample ACF of the residuals of the fitted VSARFIMA model to PM_{10} concentrations.

percentages for the first three PCs are 60.40, 72.24, 82.94, and 61.21 70.89 78.75 in the first and second cases, respectively. Thus, there do not seem to be any noticeable differences between the PCs from the sample correlation and the autocovariance matrices.

In Table 8, the four columns corresponding to the PCA of $\hat{\Gamma}_X(0)$ display the eigenvectors b_i's, the eigenvalues l_i's, the proportions $l_i/(l_1+\cdots+l_8)$'s and the cumulative proportions $(l_1+\cdots+l_i)/(l_1+\cdots+l_8)$'s for $i=1,\ldots,4$. The four columns corresponding to the PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$ display the eigenvectors c_i's, the eigenvalues m_i's, the proportions $m_i/(m_1+\cdots+m_8)$'s and the cumulative proportions $(m_1+\cdots+m_i)/(m_1+\cdots+m_8)$'s for $i=1,\ldots,4$. For both PCA, the main part of the variability is captured by the first PC, namely 61% for the PCA of $\hat{\Gamma}_X(0)$ and 57% for the PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$. The proportions for the other PCs are quite similar for both PCA. To group the monitoring stations in classes, we select for each PC the stations with the highest factor loading in absolute value. The coefficients in bold are larger than 0.37 in absolute value. Selecting these coefficients, we retain the class CL1 : VixCentro, Ibes and Cariacica for the 1st PC of $\hat{\Gamma}_X(0)$, the class CL2 : Laranjeiras and Carapina for the 2nd PC of $\hat{\Gamma}_X(0)$, the class CL3 : VVCentro for the 3rd PC of $\hat{\Gamma}_X(0)$, the class CL4 : Camburi and Sua for the 4th PC of $\hat{\Gamma}_X(0)$, and the class CL1 : Sua, VixCentro and Ibes for the 1st PC of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, the class CL2 : Laranjeiras, Carapina and Cariacica for the 2nd PC of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, and the classes CL3 : CL4 : Camburi and VVCentro for the 3rd and the 4th PC of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, respectively. Note that four PCs are necessary in the PCA of $\hat{\Gamma}_X(0)$ to encompass the eight stations, while three PCs are enough in the PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$.

Figure 6 shows the average daily profile of daily average PM_{10} concentrations at the monitoring stations, grouped by the correspondent PC/CL category. Similar profiles of PM_{10} concentrations are observed in all sites belonging to the same PC/CL category. However, it is clear that the associations PC/CL obtained with $\hat{\Gamma}_{\hat{\varepsilon}}(0)$ are better balanced and discriminate the data more clearly.

Following the same approach as Pires et al. (2008a,b), the number of monitoring stations that should be maintained among the eight corresponds to the maximum number of selected PCs. Based on the PCs of $\hat{\Gamma}_X(0)$, the four stations Ibes, Laranjeiras, VVCentro and Camburi are maintained, while the analysis of the PCs of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$ leads

Table 8.: PCA of original and filtered PM_{10} concentrations.

Station	PCA of $\hat{\Gamma}_X(0)$				PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$			
	1	2	3	4	1	2	3	4
Laranjeiras	-0.3002	**0.7193**	-0.1756	0.1460	-0.3067	**0.7090**	-0.0529	0.1606
Carapina	-0.3554	**-0.4004**	0.2628	0.1750	-0.3536	**-0.5233**	0.0368	0.0669
Camburi	-0.3472	0.1700	0.0502	**0.7019**	-0.3166	0.0560	**0.7079**	**0.5055**
Sua	-0.3632	0.2163	0.0406	**-0.6118**	**-0.3722**	0.2283	-0.3546	-0.1360
VixCentro	**-0.3864**	-0.2265	-0.1026	-0.1629	**-0.3856**	-0.0222	-0.2168	-0.2125
Ibes	**-0.3869**	0.1787	0.2359	-0.2271	**-0.3935**	0.0625	-0.1563	0.1426
VVCentro	-0.3055	-0.2942	**-0.8391**	0.0141	-0.3222	-0.0087	**-0.4764**	**-0.7571**
Cariacica	**-0.3721**	-0.2766	0.3542	0.0507	-0.3669	**-0.4044**	-0.2652	0.2383
Eigenvalue	4.8971	0.7744	0.6282	0.4973	4.5586	0.7462	0.6412	0.6050
Proportion	61.22	9.68	7.85	6.22	56.98	9.32	8.01	7.56
Cumulative	61.22	70.90	78.75	84.97	56.98	66.30	74.31	81.87

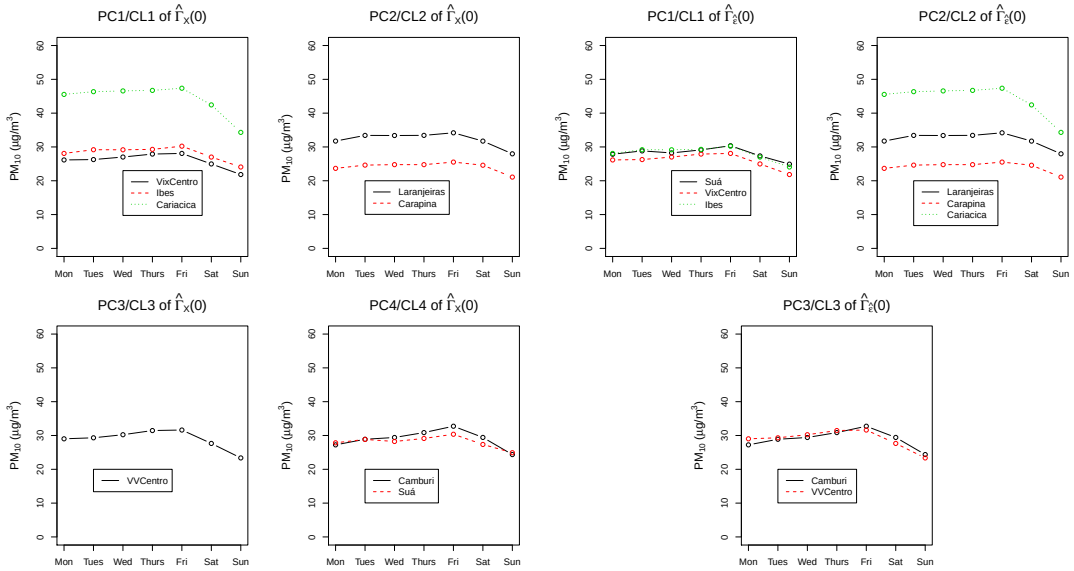


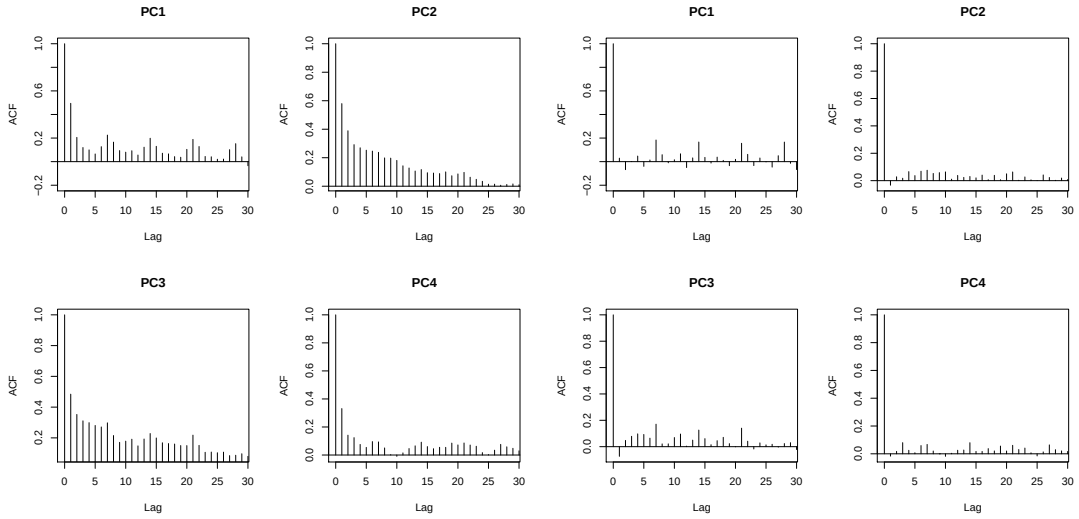
Figure 6.: Average daily profile of PM_{10} concentrations grouped by the PC/CL category.

to retain only the three stations, Ibes, Laranjeiras and Camburi. The equipment of the others stations may be moved to alternative areas of interest to cover a larger area of the GVR.

Figure 7 plots the sample ACF of the PCs of original and filtered PM_{10} concentrations. Figure 7a) shows that the PCs are autocorrelated in the case of a correlated time series. Since the filtered time series $\hat{\varepsilon}_t$ is almost a white noise, the autocorrelations in Figure 7b) are very small.

5. Conclusion

This paper has investigated the effect of time-correlation on the PCA technique. It was shown that the PCs are generally cross-correlated and present more variability compared to the case of time uncorrelated data. Explicit calculations have illustrated the effect of time-correlation on the PCs when the data follow a VAR(1) and a VMA(1)



(a) PCs of $\hat{\Gamma}_X(0)$. (b) PCs of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$.

Figure 7.: Sample ACF of the PCs of original and filtered PM_{10} concentrations.

model. The theoretical results were illustrated empirically through Monte Carlo simulations. It was found that large positive cross-covariance radically increases the variability of the PCs, and the first PC captures almost all the variability. Therefore, when the data are strongly time-correlated, it is recommended to apply a linear filter for whitening the data before PCA. The proposed methodology was applied to PM_{10} concentrations to identify redundant air quality measurements. The PCA of the filtered data is more parsimonious and leads to retaining fewer monitoring stations.

6. Acknowledgements

The authors are indebted to an anonymous reviewer for providing insightful comments that led to improve substantially the manuscript. Some results of this paper appear in the PhD thesis of Zamprogno (2013) under the supervision of Prof. V. A. Reisen in the Environmental Engineering PhD program at UFES (PPGEA), Brazil. The authors would like to thank CNPq, CAPES and FAPES for their financial support. Part of this paper was revised when Prof. V. A. Reisen was visiting CentraleSupélec (12/2018 to 01/2019). This author is indebted to CentraleSupĺec for its financial support. This research was also partially supported by DATAIA Convergence Institute as part of the "Programme d'Investissement d'Avenir, (ANR17-CONV-0003) operated by CentraleSupélec", and by the iCODE Institute, research project of the IDEX Paris-Saclay, and by the Hadamard Mathematics LabEx (LMH) through the grant number ANR-11-LABX-0056-LMH in the "Programme d'Investissement d'Avenir".

References

Anderson, T. W. (2003), An Introduction to Multivariate Statistical Analysis, 3rd edn, John Wiley & Sons.

Banerjee, S. and Roy, A. (2014), Linear algebra and matrix analysis for statistics,

Chapman & Hall/CRC Texts in Statistical Science Series, CRC Press, Boca Raton, FL.

Brockwell, P. J. and Davis, R. A. (1991), Time Series: Theory and Methods, Springer Series in Statistics, 2nd edn, Springer Science, New York, NY.

Chung, C.-F. (2012), 'Sample means, sample autocovariances, and linear regression of stationary multivariate long memory processes', Econometric Theory **18**(1), 51–58.

Cohen, S. J. (1983), 'Classification of 500 mb height anomalies using obliquely rotated principal components', J. Climate Appl. Meteorol. **22**, 1975–1988.

Fujikoshi, Y. (1980), 'Asymptotic expansions for the distributions of the sample roots under nonnormality', Biometrika **67**, 45–51.

Greenaway-McGrevy, R., Han, C. and Sul, D. (2012), 'Estimating the number of common factors in serially dependent appoximate factor models', Economics Letters **116**, 531–534.

Hamilton, J. D. (1994), Time Series Analysis, Princeton University Press.

Harville, D. A. (1997), Matrix algebra from a statistician's perspective, Springer-Verlag, New York.

Hellton, H. K. and Thoresen, M. (2014), 'The impact of measurement error on principal component analysis', Scandinavian Journal of Statistics **41**, 1051–1063.

Hu, Y.-P. and Tsay, R. S. (2014), 'Principal volatility component analysis', Journal of Business & Economic Statistics **32**(2), 153–164.

Jaimungal, S. and Ng, E. K. H. (2007), Consistent functional PCA for financial time-series, in 'Proceedings of the Fourth IASTED International Conference on Financial Engineering and Applications', FEA'07, Berkeley, USA.

Jollife, I. T. (2002), Principal component analysis, 2th edn, Prentice Hall.

Karar, K. and Gupta, A. (2007), 'Source apportionment of PM_{10} at residential and industrial sites of an urban region of Kolkata, India', Atmospheric Research **84**, 30–41.

Liu, P.-W. G. (2009), 'Simulation of the daily average PM_{10} concentrations at Ta-Liao with Box-Jenkins time series models and multivariate analysis', Atmospheric Environment **43**, 2101–2113.

Lu, W.-Z., He, H.-D. and Dong, L.-y. (2011), 'Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis', Building and Environment **46**(3), 577–583.

Lutkepohl, H. (2005), New Introduction to Multiple Time Series Analysis, Springer-Verlag.

Matteson, D. S. and Tsay, R. S. (2011), 'Dynamic orthogonal components for multivariate time series', Journal of the American Statistical Association **106**(496), 1450–1463.

Melo, M. M. (2015), Correlação entre percepção do incômodo e exposição ao material particulado presente na atmosfera e sedimentado, PhD thesis, Federal University of Espírito Santo, Brazil.

Melo, M. M., Reisen, V. A., Santos, J. M., Reis Junior, N. C., Frère, S., Bondon, P., Ispány, M. and Cotta, H. H. A. (to appear), 'The use of multivariate time series techniques to estimate the impact of particulate matter on the perceived annoyance', Atmospheric environment .

Perez-Neto, P. R., Jackson, D. A. and Somers, K. M. (2005), 'How many principal components? Stopping rules for determining the number of non-trivial axes revisited', Computacional Statistics & Data Analysis **49**(4), 974–997.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. and Martins, F. G. (2008a), 'Management of air quality monitoring using principal component

and cluster analysis — part I: SO_2 and PM_{10}', Atmospheric Environment **42**, 1249–1260.

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. and Martins, F. G. (2008b), 'Management of air quality monitoring using principal component and cluster analysis — part II: CO, NO_2 and O_3', Atmospheric Environment **42**, 1261–1274.

R Core Team (2019), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org*

Reinsel, G. C. (1997), Elements of Multivariate Time Series Analysis, second edn, Springer Series in Statistics.

Reisen, V. A., Sgrâncio, A. M., Lévy-Leduc, C., Bondon, P., Ziegelmann, F., Monte, E. Z. and Cotta, H. H. A. (2019), 'Robust factor modeling for high-dimensional time series: an application to air pollution data.', Applied Mathematics and Computation **346**, 842–852.

Reisen, V. A., Zamprogno, B., Palma, W. and Arteche, J. (2014), 'A semiparametric approach to estimate two seasonal fractional parameters in the SARFIMA model', Mathematics and Computers in Simulation **98**, 1 – 17.

Richman, M. B. (1986), 'A principal component analysis of sulphur concentrations in the western United States', Atmospheric Environment **20**, 606–607.

Romero, R., Ramis, C., Guijarro, J. A. and Sumner, G. (1999), 'Daily rainfall affinity areas in Mediterranean Spain', Int. J. Climatol **19**, 557–578.

Souza, J. B., Reisen, V. A., Franco, G. C., Ispany, M., Bondon, P. and Santos, J. M. (2018), 'Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data', Journal of the Royal Statistical Society: Series C (Applied Statistics) **67**(2), 453–480.

Souza, J. B., Reisen, V. A., Santos, J. M. and Franco, G. C. (2014), 'Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution', Revista de saude publica **48**(3), 451–458.

Taniguchi, M. and Krishnaiah, P. R. (1987), 'Asymptotic distributions of functions of the eigenvalues of sample covariance matrix and canonical correlation matrix in multivariate time series', Journal of Multivariate Analysis **22**, 156–176.

Vanhatalo, E. and Kulahci, M. (2016), 'Impact of autocorrelation on principal components and their use in statistical process control', Quality and Reliability Engineering International **32**(4), 1483–1500. QRE-15-0259.

Vanhatalo, E., Kulahci, M. and Bergquist, B. (2017), 'On the structure of dynamic principal component analysis used in statistical process monitoring', Chemometrics and Intelligent Laboratory Systems **167**, 1 – 11.

Wang, Y. and Pham, H. (2011a), 'Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components', International Journal of System Assurance Engineering and Management **2**(3), 253–259.

Wang, Y. and Pham, H. (2011b), '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.

White, D., Richman, M. B. and Yarnal, B. (1991), 'Climate regionalization and rotation of principal components', Int. J. Climatol. **11**, 1–25.

Zamprogno, B. (2013), O uso e interpretação de análise de componentes principais, em séries temporais, com enfoque no gerenciamento da qualidade do ar, PhD thesis, Federal University of Espírito Santo, Brazil.