Applied Mathematics and Computation 346 (2019) 842–852

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Robust factor modelling for high-dimensional time series: An application to air pollution data

Valdério Anselmo Reisen[a,c,*], Adriano Marcio Sgrancio[a], Céline Lévy-Leduc[b], Pascal Bondon[c], Edson Zambon Monte[d], Higor Henrique Aranda Cotta[a,c], Flávio Augusto Ziegelmann[e]

[a] PPGEA and Department of Statistics, Federal University of Espirito Santo, Brazil
[b] AgroParisTech/UMR INRA MIA 518, France
[c] Laboratoire des Signaux et Systémes, CNRS, CentraleSupélec, Université, Paris-Sud, France
[d] Department of Economics, Federal University of Espírito Santo, Espírito Santo, Brazil
[e] Department of Statistics, Ppge and Ppga, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil

ARTICLE INFO

Keywords:
Factor analysis
Time series
Robustness
Eigenvalues
Reduced rank
Air pollution

ABSTRACT

This paper considers the factor modelling for high-dimensional time series contaminated by additive outliers. We propose a robust variant of the estimation method given in Lam and Yao [10]. The estimator of the number of factors is obtained by an eigen analysis of a robust non-negative definite covariance matrix. Asymptotic properties of the robust eigenvalues are derived and we show that the resulting estimators have the same convergence rates as those found for the standard eigenvalues estimators. Simulations are carried out to analyse the finite sample size performance of the robust estimator of the number of factors under the scenarios of multivariate time series with and without additive outliers. As an application, the robust factor analysis is performed to reduce the dimensionality of the data and, therefore, to identify the pollution behaviour of the pollutant PM_{10}.

1. Introduction

In the last fifty years, issues related to air pollution have grown into a major problem, specially in developing countries, where the air quality has been degraded as a result of industrialization, population growth, high rates of urbanization and inadequate or non-existent policies to control air pollution. The problems caused by air pollution produce local, regional and global impacts. Among different environmental problems, air pollution is reported to cause the greatest damage to health and loss of quality of life see, for example, WHO [32]. The most common health problems caused by air pollution are asthma, rhinitis, burning eyes, fatigue, dry cough, heart and lung diseases and heart failure. The main pollutants are carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen oxides (NO_x), ozone (O_3) and inhalable particles with diameter smaller than 10 μm (PM_{10}). The papers by Brunekreef and Holgate [3], Maynard [18], WHO [31], Curtis et al. [6] and Souza et al. [25] discuss the relationship between these pollutants and health problems. In addition, air pollution contributes to the degradation of the environment, the greenhouse effect among many others problems.

* Corresponding author at: Department of Statistics, Federal University of Espírito Santo, 514 – Vitoria, 29075-910 Espírito Santo, Brazil.
E-mail address: valderio.reisen@ufes.br (V.A. Reisen).

https://doi.org/10.1016/j.amc.2018.09.062
0096-3003/

In recent studies related to air pollution, much attention has been paid to mathematical receptor models with the aim to measure and analyse the pollutant concentrations at the source of emission. For this, mathematical and statistical tools are used to identify the pollutant emission sources from chemical characteristics of the particles on the receiver and the pollutant emission sources see, for example, Seinfeld and Pandis [24]. In the literature, the most studied receptor models are:chemical mass balance (CMB), multivariate analysis, principal component analysis techniques (PCA), factor analysis (FA) model, multiple linear regression, cluster analysis and positive matrix factorization (PMF) (Watson et al. [30]). In particular, the classical FA has been widely used in air pollution analysis, specially for the identification of emission sources, the management of monitoring networks, regression analysis, cluster analysis and prediction.

In many practical problems, it is quite common to have observations which accommodate the serial dependence of each component and the interdependence between different components, that is, the data are time-dependent. However, it should be noted that, among the studies that adopted the classical PCA and FA techniques, time-dependency of the data is a commonly neglected feature. A common assumption of the multivariate statistical tools is that the data are independent in time, see e.g. Anderson [1] and Johnson and Wichern [9]. To deal with autocorrelated data in FA, Pea and Box [20], Stock and Watson [26], Lam et al. [11] and Lam and Yao [10] studied the factor modelling for multivariate time series from a dimension-reduction point of view. Contrarily to PCA and FA for independent observations, these papers look for factors which drive the serial dependence of the original time series. Further discussions and additional references can be found in Lam and Yao [10].

Since FA method allows to reduce the order of the estimated model, this technique has been widely used for forecasting. According to Stock and Watson [26], the dimension reduction becomes a central concern for forecasting when the number of candidate predictor series is very large. This issue can make the forecast investigation impractical in a real application, for example in the use of vector autoregressive moving average (VARMA) models with a large number of variables. This high-dimensional problem is simplified by modelling the common dynamics in terms of a relatively small number of unobserved latent factors. Then, forecasting can be carried out in a two-step process: first, a time series of the factors is estimated from the predictors; second, the relationship between the variable to be forecast and the factors is estimated, for example, using a linear regression.

Environmental time series are often of high dimension due to the large number of measurements recorded across many different locations. These data may also present interesting phenomena to be considered from an applied and theoretical point of view. Indeed, the concentration of pollutant may present high peaks, which can be seen as aberrant values from a statistical point of view. Outliers and high dimension data are common in many areas of applied mathematics. Therefore, the methodology proposed here can be widely used in many other areas where the multivariate techniques are the main tools to describe and interpret the data. This is the case of the health science area, Gosak et al. [7], Perc [21], Souza et al. [25], air route network problems, Lordan et al. [15], Zhang et al. [34], environmental engineering, Zamprogno [33] and statistical process controls, Vanhatalo and Kulahci [29], to cite a few.

As is well known, outliers can affect the statistical properties of the estimates such as the sample mean and sample covariance, see e.g., Chang et al. [4], Tsay [27], Chen and Liu [5] and the references therein. Since the parameter estimation is connected with these sample functions, the final estimated time series model can be strongly affected by the outliers. When the series has additive outliers, one way to deal with model estimation is to use robust estimates of these statistics. For a univariate time series, Ma and Genton [17] proposed a robust sample autocorrelation function (ACF) based on the robust scale estimate $Q_n(.)$ suggested in Rousseeuw and Croux [23]. This robust ACF estimator was recently studied by Lévy-Leduc et al. [12]–[14].

This paper considers multivariate time series with additive outliers using the FA technique for dimension reduction. In this context, a robust version of the dimension reduction estimator given in Lam and Yao [10] is proposed. Some theoretical results are discussed and the method performance is investigated through Monte Carlo simulations. The proposed methodology is applied to PM_{10} concentrations measured at the Automatic Air Quality Monitoring Network (AAQMN), Vitória, Brazil.

The rest of the paper is organized as follows. In Section 2, the model and the estimation methods are presented. Section 3 discusses the asymptotic properties of the robust eigenvalues. Section 4 presents some Monte Carlo experiments. Section 5 considers an application of the proposed methodology and some concluding remarks are provided in Section 6.

2. Factor model in time series

2.1. The factor model and the estimate of the number of factors

Let $\boldsymbol{Z}_t$, $t \in \mathbb{Z}$, be a k-dimensional zero-mean vector of an observed time series and $\boldsymbol{X}_t$ be an unobserved r-dimensional vector of common factors ($r \leq k$). It is assumed that $\boldsymbol{Z}_t$ is generated by

$$\boldsymbol{Z}_t = \boldsymbol{P}\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{P}$ is an unknown $k \times r$ matrix of parameters of rank r, denominated the factor-loading matrix, and $\boldsymbol{\varepsilon}_t$ is a k-dimensional zero-mean white-noise sequence with full-rank covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, that is, $\boldsymbol{\varepsilon}_t \sim WN(\boldsymbol{0}, \boldsymbol{\Sigma}_{\varepsilon})$. When r is small relative to k, the model presented in (1) is most useful, since it results in a multivariate time series with a reduced dimension and, consequently, leads to a much simpler multivariate time series for forecasting. The following assumption is introduced.

(A1) $\boldsymbol{X}_t$ is a zero-mean multivariate stationary process, $\boldsymbol{\varepsilon}_t \sim WN(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, $\boldsymbol{X}_t$ and $\boldsymbol{\varepsilon}_s$ are uncorrelated for any t and s, and $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_r$, where $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix.

Assumption (A1) ensures identifiability in (1), see Lam and Yao [10] and Pea and Box [20] for further details. It follows from (1) and (A1) that the covariance matrix function of $\boldsymbol{Z}_t$ satisfies

$$\boldsymbol{\Gamma}^Z(h) = \mathrm{E}[\boldsymbol{Z}_t\boldsymbol{Z}'_{t+h}] = \begin{cases} \boldsymbol{P}\boldsymbol{\Gamma}^X(0)\boldsymbol{P}' + \boldsymbol{\Sigma}_\varepsilon & \text{when } h = 0, \\ \boldsymbol{P}\boldsymbol{\Gamma}^X(h)\boldsymbol{P}' & \text{when } h \neq 0. \end{cases} \tag{2}$$

Given a sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, the first step is to estimate the number of factors r and to compute an estimate $\hat{\boldsymbol{P}}$ of the $k \times r$ factor loading matrix $\boldsymbol{P}$. Then, the estimators of the factor process and the residuals are, respectively, given by

$$\hat{\boldsymbol{X}}_t = \hat{\boldsymbol{P}}'\boldsymbol{Z}_t, \tag{3}$$

and

$$\hat{\boldsymbol{\epsilon}}_t = (\boldsymbol{I}_k - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{Z}_t. \tag{4}$$

For further details on the estimation of $\boldsymbol{P}$, see Lam and Yao [10].

Let $\hat{\boldsymbol{\Gamma}}^Z(h)$ denote the sample covariance matrix of $\boldsymbol{Z}_t$ at lag h and let

$$\hat{\boldsymbol{M}} = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}^Z(h)\hat{\boldsymbol{\Gamma}}^Z(h)', \tag{5}$$

where h_0 is a prescribed positive integer. Following the lines of Lam and Yao [10], the estimator of the number of factors r is given by

$$\hat{r} = \arg\min_{1 \le i \le R} \hat{\lambda}_{i+1}/\hat{\lambda}_i, \tag{6}$$

where $r < R < k$ is a constant and $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_k$ are the eigenvalues of $\hat{\boldsymbol{M}}$. Lam and Yao [10] derive the asymptotic properties of the eigenvalues $\hat{\lambda}_i$'s under some assumptions, and they give some practical recommendations for selecting R. In the following, we propose a robust estimator of r.

2.1.1. The robust estimator of the number of factors r

Let Y_t, $t \in \mathbb{Z}$, be a univariate stationary Gaussian process. Given the observations $Y_{1:n} = (Y_1, \ldots, Y_n)$, the $Q_n(.)$ estimator of the standard deviation of Y_1 proposed by Rousseeuw and Croux [23] is the kth order statistic defined by

$$Q_n(Y_{1:n}) = c\,\{|Y_i - Y_j|; i < j\}_{\{k\}}, \quad i, j = 1, \ldots, n, \tag{7}$$

where $c = 2.2191$ is a constant to guarantee consistency, $k = \lfloor \binom{n}{2} + 2)/4 \rfloor + 1$ and $\lfloor x \rfloor$ is the largest integer smaller than x. The asymptotic breakdown point of $Q_n(Y_{1:n})$ is 50%. Following Ma and Genton [16], from the observations $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, we propose to estimate $\boldsymbol{\Gamma}^Z_{i,j}(h) = \mathrm{Cov}(Z_{i,t}, Z_{j,t+h})$ for all $i, j = 1, \ldots, k$, by

$$\hat{\gamma}^{Q,Z}_{i,j}(h) = \frac{1}{4}\left[Q^2_{n-h}(Z_{i,1:n-h} + Z_{j,h+1:n}) - Q^2_{n-h}(Z_{i,1:n-h} - Z_{j,h+1:n})\right], \tag{8}$$

where $Z_{i,1:n-h} = (Z_{i,1}, \ldots, Z_{i,n-h})$ and $Z_{j,h+1:n} = (Z_{j,h+1}, \ldots, Z_{j,n})$. Let $\boldsymbol{\Gamma}^{Q,Z}(h)$ be the matrix with entries $\hat{\gamma}^{Q,Z}_{i,j}(h)$, we define $\hat{\boldsymbol{M}}^Q$ as

$$\hat{\boldsymbol{M}}^Q = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}^{Q,Z}(h)\hat{\boldsymbol{\Gamma}}^{Q,Z}(h)', \tag{9}$$

and the robust estimator $\hat{r}^Q$ of r is obtained from (6) where the $\hat{\lambda}_i$'s are replaced by the eigenvalues $\hat{\lambda}^Q_i$'s of $\hat{\boldsymbol{M}}^Q$.

3. Theoretical results

Here, we present some theoretical results to support the robust approach discussed in Section 2. We introduce the following assumption on $\boldsymbol{X}_t$.

(A2) $\boldsymbol{X}_t$, $t \in \mathbb{Z}$, is a zero-mean multivariate Gaussian stationary process satisfying

$$\sum_{h \ge 1} |\boldsymbol{\Gamma}^X_{i,j}(h)| < \infty, \text{ for all } i, j = 1, \ldots, r.$$

It follows from (1) and (2) that $(\boldsymbol{Z}_t)$ is also a zero-mean multivariate Gaussian stationary process satisfying

$$\sum_{h \ge 1} |\boldsymbol{\Gamma}^Z_{i,j}(h)| < \infty, \text{ for all } i, j = 1, \ldots, k. \tag{10}$$

Table 1
Relative frequency estimates of $P(\hat{r}=3)$ for the uncontaminated process.

n	50	100	200	400	800	1600
$k=0.2n$	0.170	0.585	0.870	0.995	1	1
$k=0.5n$	0.395	0.710	0.975	1	1	1
$k=0.8n$	0.435	0.785	0.960	1	1	1

Table 2
Relative frequency estimates of $P(\hat{r}^{Q}=3)$ for the uncontaminated process.

n	50	100	200	400	800	1600
$k=0.2n$	0.150	0.450	0.850	0.980	1	1
$k=0.5n$	0.320	0.680	0.950	1	1	1
$k=0.8n$	0.390	0.690	0.950	1	1	1

Table 3
Relative frequency estimates for dimensional reduction when $n=100$.

	$p=0$			$p=0.05$ and $\omega=15$			$p=0$			$p=0.05$ and $\omega=15$		
	$\hat{r}=1$	$\hat{r}=2$	$\hat{r}=3$	$\hat{r}=1$	$\hat{r}=2$	$\hat{r}=3$	$\hat{r}^{Q}=1$	$\hat{r}^{Q}=2$	$\hat{r}^{Q}=3$	$\hat{r}^{Q}=1$	$\hat{r}^{Q}=2$	$\hat{r}^{Q}=3$
$k=0.2n$	0.110	0.305	0.585	0.380	0.330	0.290	0.140	0.410	0.450	0.180	0.380	0.440
$k=0.5n$	0.100	0.190	0.710	0.380	0.360	0.260	0.100	0.220	0.680	0.160	0.310	0.530
$k=0.8n$	0.040	0.175	0.785	0.430	0.360	0.210	0.040	0.270	0.690	0.060	0.290	0.650

Theorem 1. *Under assumptions (A1) and (A2) and for a fixed $h_0 \geq 1$, as $n \to \infty$,*

$$|\hat{\lambda}_i^{Q} - \lambda_i| = O_p(u_n^{-1/2}), \text{ for } i = 1, \ldots, k,$$

where $\hat{\lambda}_i^{Q}$'s and λ_i's are the eigenvalues of $\hat{\boldsymbol{M}}^{Q}$ and $\sum_{h=1}^{h_0} \boldsymbol{\Gamma}^{Z}(h)\boldsymbol{\Gamma}^{Z}(h)'$, respectively.

Remark 1. Lam and Yao [10, Proposition 1] establish a similar result to Theorem 1 for the eigenvalues $\hat{\lambda}_i$'s of $\hat{\boldsymbol{M}}$.

Proof of Theorem 1 directly follows from Lemmas 1–3 given below and proved in Section 7.

Lemma 1. *Let $\hat{A}_n$ be a sequence of $k \times k$ symmetric matrices and A be a $k \times k$ symmetric matrix such that $\hat{A}_n - A = O_p(u_n^{-1})$ as $n \to \infty$, where $u_n > 0$ and $u_n \to \infty$ as $n \to \infty$. Then, as $n \to \infty$,*

$$|\lambda_i(\hat{A}_n) - \lambda_i(A)| = O_p(u_n^{-1}), \text{ for } i = 1, \ldots, k,$$

where $\lambda_i(\hat{A}_n)$'s and $\lambda_i(A)$'s are the eigenvalues of $\hat{A}_n$ and A, respectively.

Lemma 2. *Let $\hat{A}_n(h)$ be a sequence of $k \times k$ symmetric matrices and $A(h)$ be a $k \times k$ symmetric matrix such that $\hat{A}_n(h) - A(h) = O_p(u_n^{-1})$ as $n \to \infty$ for each $h = 1, \ldots, h_{\max}$, where $u_n > 0$ and $u_n \to \infty$ as $n \to \infty$. Then, as $n \to \infty$,*

$$\sum_{h=1}^{h_{\max}} \hat{A}_n(h)\hat{A}_n(h)' - \sum_{h=1}^{h_{\max}} A(h)A(h)' = O_p(u_n^{-1}).$$

Lemma 3. *Under assumptions (A1) and (A2), for all $i, j = 1, \ldots, k$ and $h \geq 0$, the robust autocovariance estimator $\hat{\gamma}_{i,j}^{Q,Z}(h)$ of $\boldsymbol{\Gamma}_{i,j}^{Z}(h)$ satisfies the central limit theorem,*

$$\sqrt{n}(\hat{\gamma}_{i,j}^{Q,Z}(h) - \boldsymbol{\Gamma}_{i,j}^{Z}(h)) \xrightarrow{d} \mathrm{N}(0, \tilde{\sigma}_{i,j}^{2}(h)),$$

as $n \to \infty$, where

$$\tilde{\sigma}_{i,j}^{2}(h) = \mathrm{E}[\psi(Z_{i,1}, Z_{j,1+h})^2] + 2\sum_{\ell \geq 1} \mathrm{E}[\psi(Z_{i,1}, Z_{j,1+h})\psi(Z_{i,\ell+1}, Z_{j,\ell+1+h})]$$

and ψ is defined by (11).

4. Simulation study

This section reports simulation results related to the performance of the proposed methodology for finite sample size. In this empirical study, $r = 3$ and $\boldsymbol{X}_t$ is the VAR(1) model defined by $\boldsymbol{X}_t = \Phi\boldsymbol{X}_{t-1} + \eta_t$, where the coefficient matrix Φ is diagonal with 0.6, −0.5 and 0.3 as the main diagonal elements, and η_t are independent zero-mean Gaussian vectors with identity covariance matrix. Since Φ and the covariance matrix of η_t are diagonal, the components of $\boldsymbol{X}_t$ are independent.

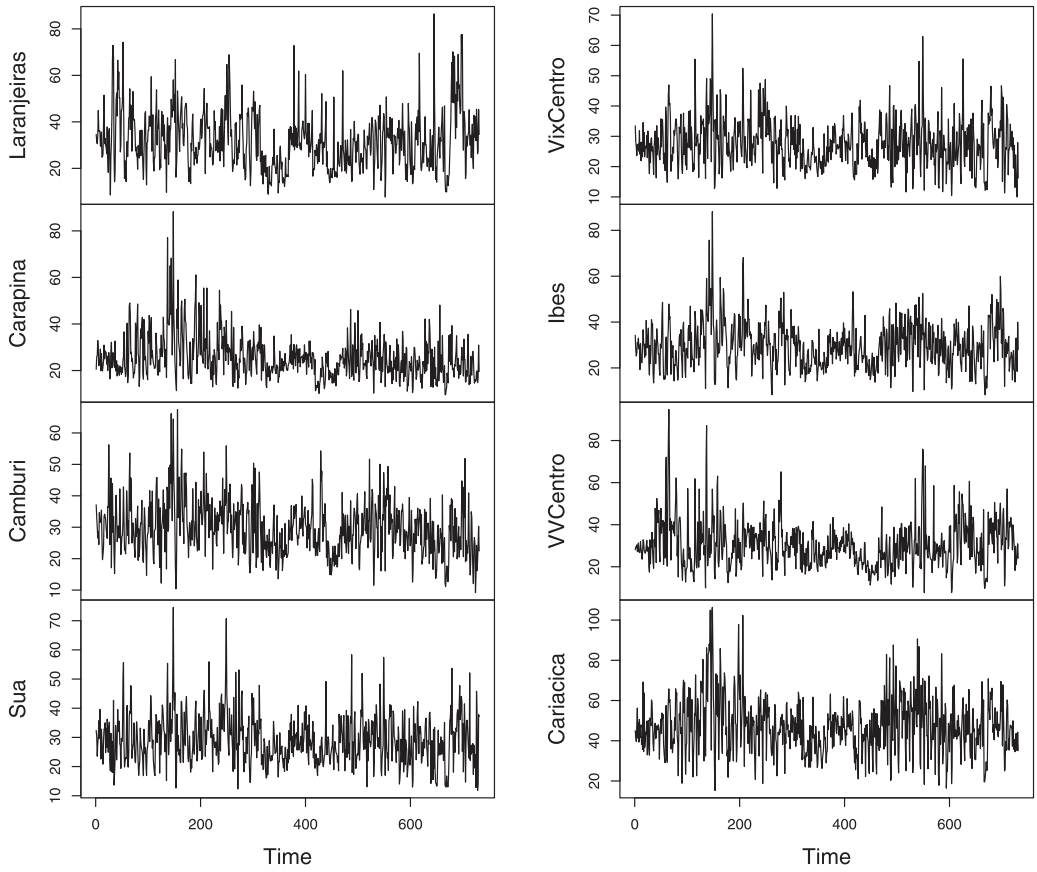


Fig. 1. Plots of the PM_{10} pollutant concentrations of the eight stations of AAQMN ($k = 8$).

The sample sizes are $n = 50, 100, 200, 400, 800$ and 1600, $k = 0.2n, 0.5n, 0.8n$, and $h_0 = 1$. Factor model (1) is obtained as follows. The elements of $\boldsymbol{P}$ are realizations of independent random variables with a uniform distribution on $[-1, 1]$. The random variables $\boldsymbol{\varepsilon}_t$ are independent zero-mean Gaussian vectors with identity covariance matrix. The same simulation process is considered by Lam and Yao [10]. The empirical results are based on 1000 replications. The simulations were ran using the R programming language.

The main interest in this empirical study is to verify the performance of the statistics $\hat{r}$ and $\hat{r}^{Q}$ in the context of a VAR(1) model with and without outliers. For this, the relative frequency estimates for the probabilities $P(\hat{r} = r)$ and $P(\hat{r}^{Q} = r)$ are reported in Tables 1 and 2, respectively. The results in Table 1 are similar to the ones in Table 1 of Lam and Yao [10], i.e., $\hat{r}$ performs better as n and k increase. Table 2 shows that $\hat{r}^{Q}$ slightly under performs $\hat{r}$ which indicates that $\hat{r}^{Q}$ can also be used to estimate r.

Now, let $\boldsymbol{X}_t^*$ be the contaminated version of $\boldsymbol{X}_t$ defined by $\boldsymbol{X}_{i,t}^* = \boldsymbol{X}_{i,t} + \omega_i \delta_{i,t}$ for all $i = 1, \ldots, r$, where $\omega_i \geq 0$ is the magnitude of the outlier which impacts $\boldsymbol{X}_{i,\,t}$ and $\delta_{i,\,t}$ indicates the presence or not of this outlier and its sign at time t. The random variable $\delta_{i,\,t}$ takes the values $-1, 1, 0$ with the respective probabilities $p/2, p/2, 1 - p$ where $0 < p < 1$ is the probability of occurrence of the outlier. We assume that $\boldsymbol{X}_{i,\,t}$ and $\delta_{i,\,t}$ are independent and that $\mathrm{E}(\delta_{i,t}\delta_{j,t+h}) \neq 0$ only when $i = j$ and $h = 0$. Here, we take $p = 0.05$, $\omega_1 = 15$ and $\omega_2 = \omega_3 = 0$. Table 3 shows the relative frequency estimates for $P(\hat{r} = 3)$ and $P(\hat{r}^{Q} = 3)$. We see that $P(\hat{r} = 3)$ decreases substantially with respect to the case $p = 0$ presented in Table 1. This shows that $\hat{r}$ which is based on $\hat{\boldsymbol{M}}$ in (5) is not robust to additive outliers, and this is not surprising since the sample covariance matrix $\hat{\boldsymbol{\Gamma}}^{Z}(h)$ is not robust. On the other hand, we see that $P(\hat{r}^{Q} = 3)$ is almost similar in Tables 2 and 3 which shows the good robustness of $\hat{r}^{Q}$ to additive outliers and indicates that the methodology proposed in this paper may be used when the presence of outliers in the series is uncertain. Table 3 also shows the estimated probability of the test to indicate $\hat{r} = 1$ or $\hat{r} = 2$. In the outliers case, the non-robust test has the tendency to increase the relative frequency estimates for $P(\hat{r} = 1)$. This spurious result is caused by the fact that outliers lead to an underestimation of the true ACF see, for example, Reisen et al. [22]. Other simulations with different degrees of contamination present similar conclusions and are available upon request.

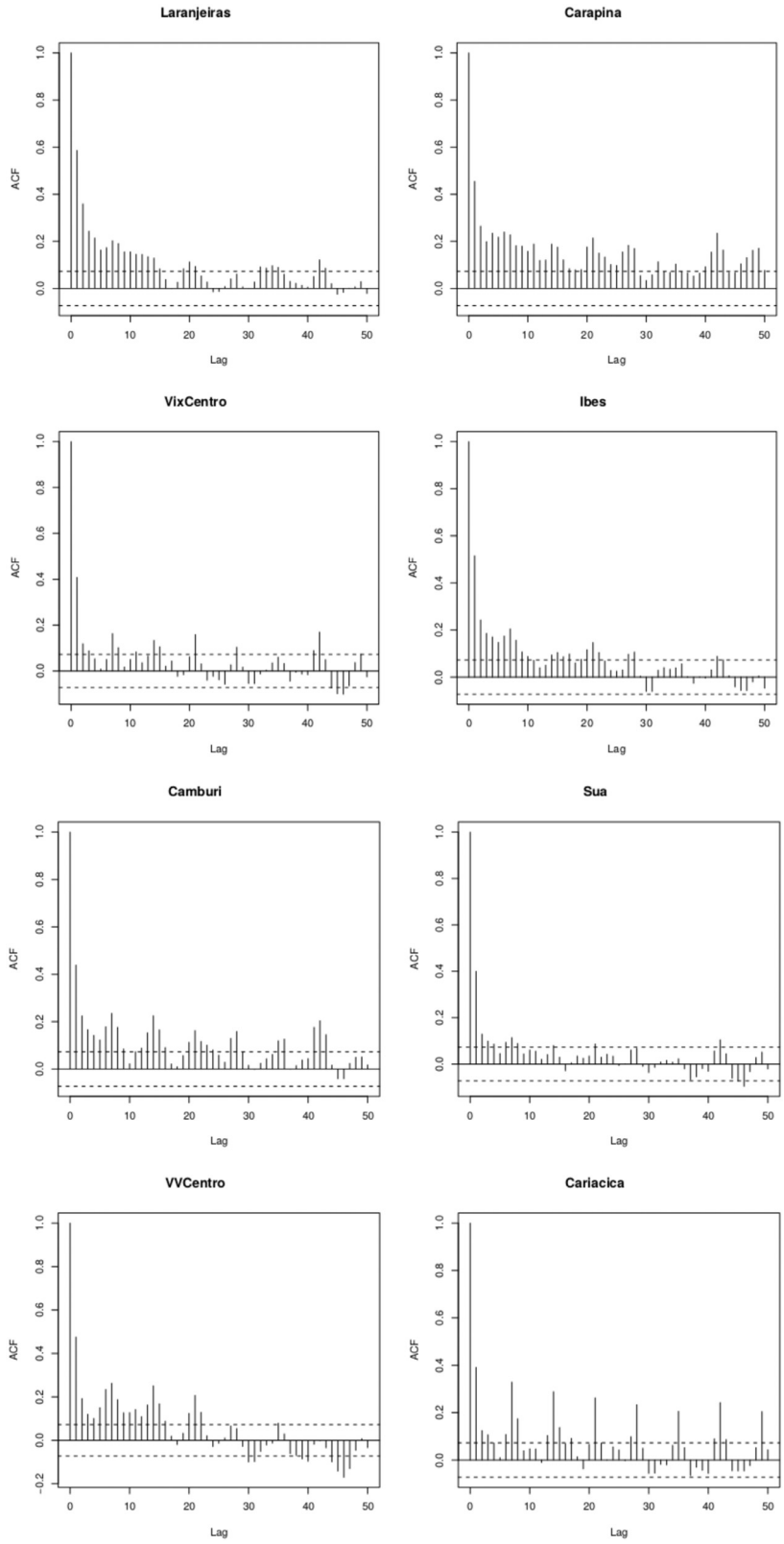


Fig. 2. Classical ACF estimates of the PM_{10} pollutant concentrations.

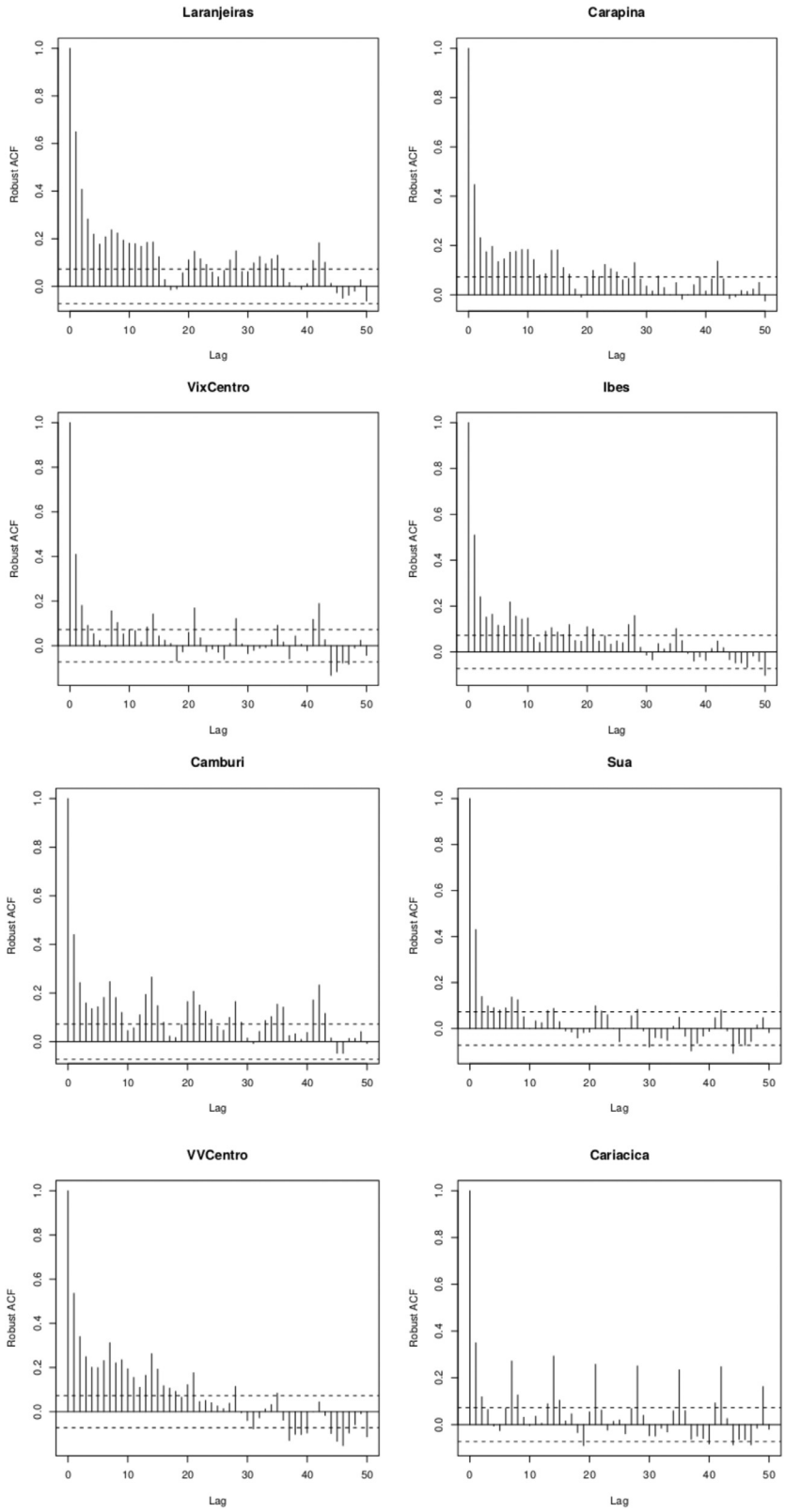


Fig. 3. Robust ACF estimates of the PM_{10} pollutant concentrations.

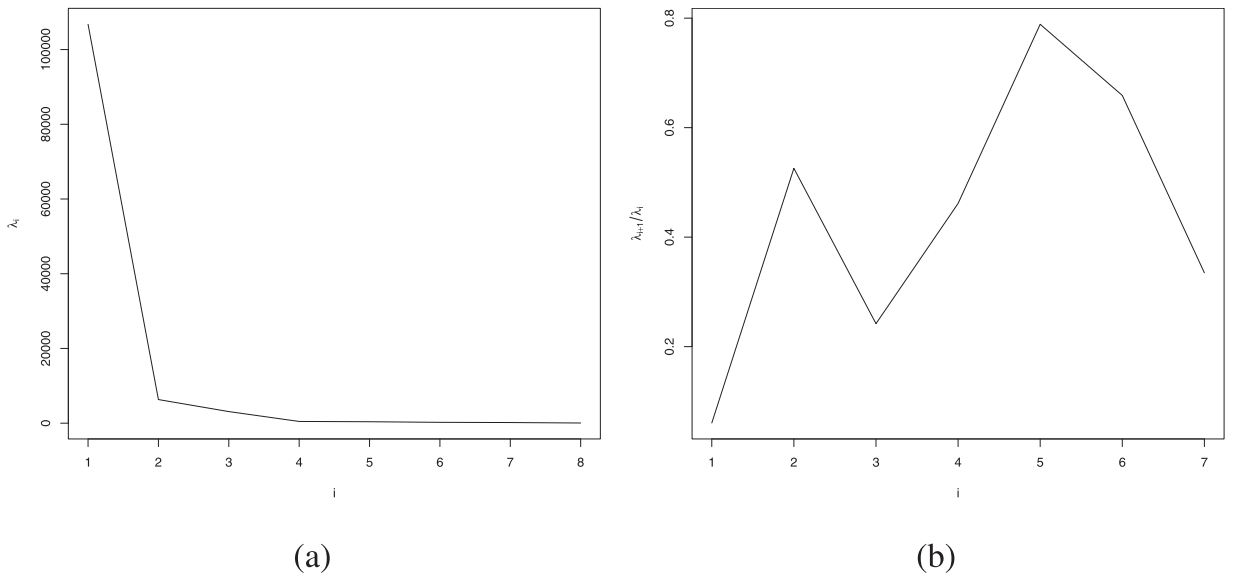


Fig. 4. A scree plot (a) and the plot of the ratios (b) of the eigenvalues of $\hat{\boldsymbol{M}}$.

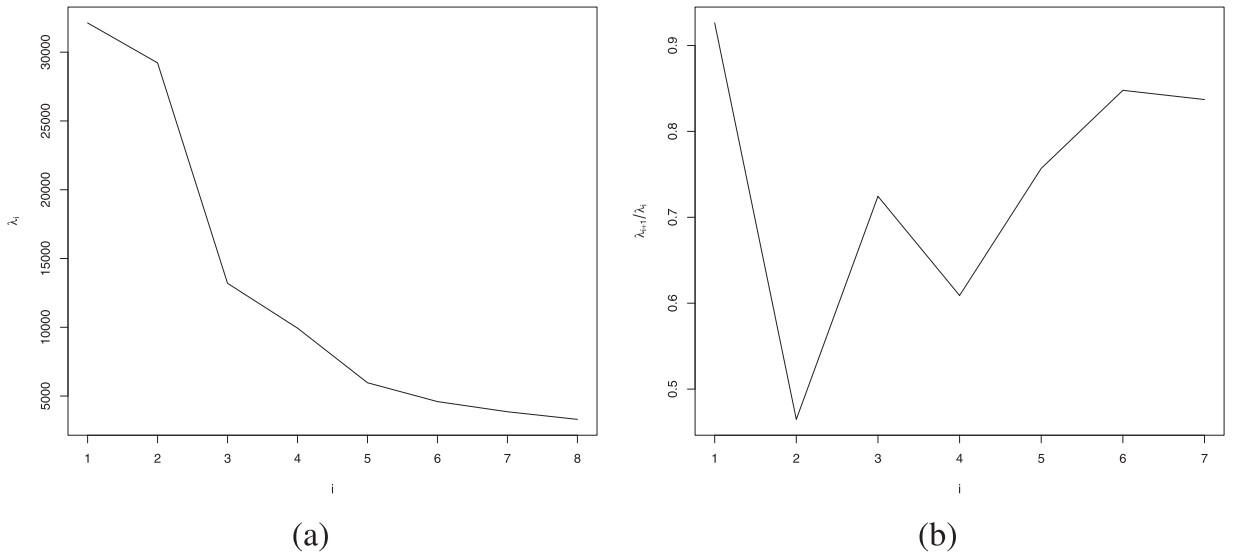
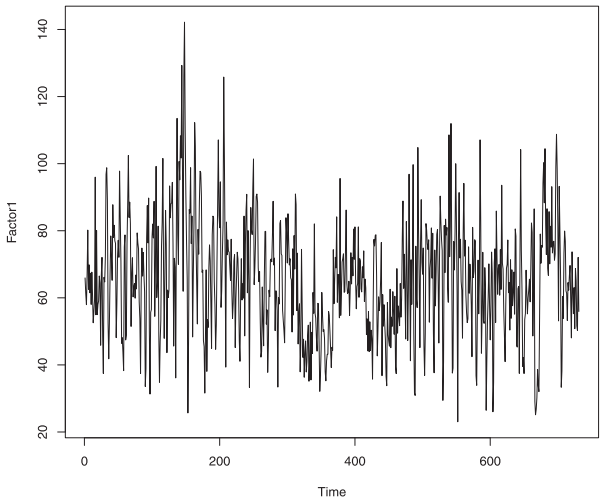


Fig. 5. A scree plot (a) and the plot of the ratios (b) of the eigenvalues of $\hat{\boldsymbol{M}}^{Q}$.

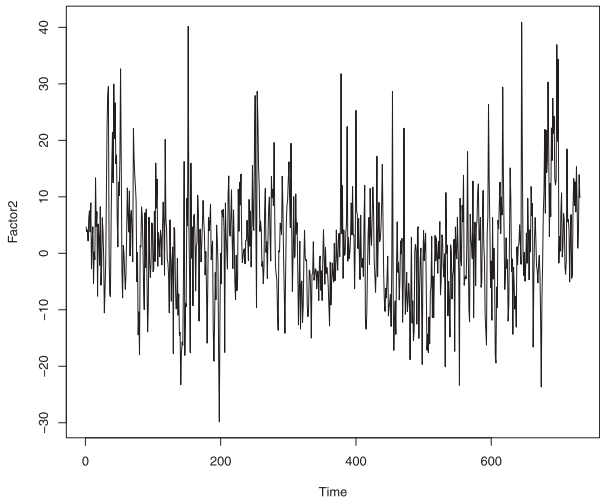
5. Application to the pollutant PM_{10}

Here, we present an application of our methodology for the PM_{10} pollutant concentrations measured at the AAQMN in the Greater Vitória Region (GVR), Espírito Santo, Brazil. GVR is comprised of seven cities with a population of approximately 1.9 million inhabitants in an area of 2319 km^2. The AAQMN consists of eight monitoring stations distributed in the cities of GVR; Laranjeiras, Carapina, Camburi, Suá, Vitória (Center), Vila Velha (center), Ibes and Cariacica. The pollutant PM_{10}, expressed in $\mu g/m^3$ was hourly measured from January 2008 to December 2009, $k = 8$, and the daily average values ($n = 731$) are used in this study. This follows the same lines as the application considered by Lam and Yao [10]. Let $\boldsymbol{Z}_t = (Z_{1,t}, \ldots, Z_{8,t})'$, $t = 1, \ldots, 731$, be the vector of the PM_{10} concentrations, where $Z_{i,\,t}$ corresponds to PM_{10} concentration at ith location.

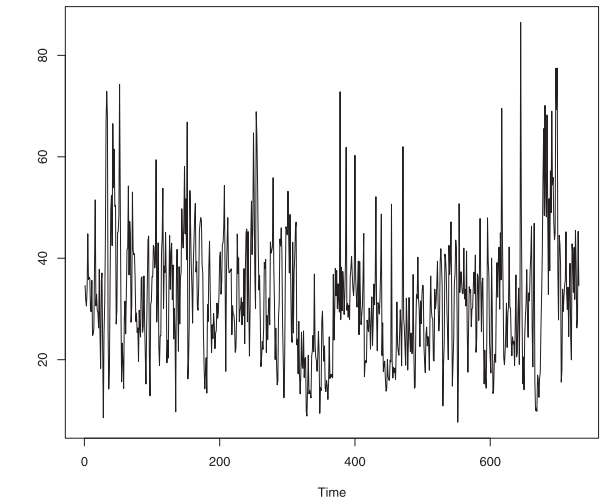
Fig. 1 shows the plots of the PM_{10} concentrations for the eight stations. We see that the series present high levels of pollutant concentrations which can be identified, from a statistical point of view, as additive outliers. This is justified by the fact that these values produce a similar reduction of the sample autocorrelations as additive outliers do. The robust and non-robust approaches discussed previously, are used here to verify whether these high levels influence the factor model estimation or not.



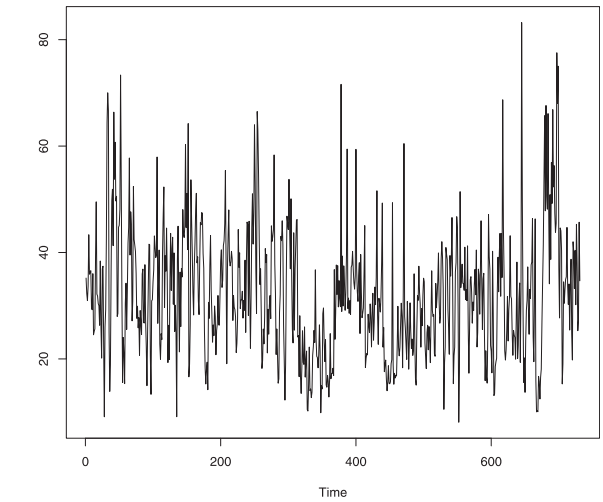
(a) Estimated factor 1 ($\hat{X}_{1,t}$).



(b) Estimated factor 2 ($\hat{X}_{2,t}$).



(c) Observed Laranjeiras series.



(d) Estimated Laranjeiras series.

Fig. 6. The time series plots of the two estimated factors by means of the robust method, (a) and (b), respectively. The observed concentrations of Laranjeiras station (c) and the estimated concentrations of Laranjeiras station (d), in the same time period.

The classical and robust ACF estimators displayed in Figs. 2 and 3, respectively, exhibit a possible seasonal pattern of period $s = 7$, which is not surprising since the data are daily. In terms of magnitude, the classical ACF estimator values at Vila Velha (center) station for example are 0.47, 0.12, 0.15 and 0.13 for lags $h = 1, 3, 5, 10$, respectively, while the ACF values based on the Q_n function are 0.54, 0.25, 0.20 and 0.19. This shows that the high levels of PM_{10} at Vila Velha (center) station reduce the sample ACF estimator values. Similar results are observed at the other stations. The effect of atypical observations on the estimation of the ACF function is discussed in Molinares et al. [19] for a univariate time series.

From the above discussion, it is expected that the standard and robust FA estimated models present different conclusions. The estimates of the number of factors r are computed by performing an eigenanalysis of $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{M}}^{Q}$ given by (5) and (9), respectively, with $h_0 = 7$ to capture the seasonality of the data set. The eigenvalues of (5) (the scree plot), in decreasing order, and their ratios are shown in Fig. 4(a) and (b), respectively. The robust versions obtained from $\hat{\boldsymbol{M}}^{Q}$ are shown in Fig. 5(a) and (b), respectively. We see that $\hat{r} = 1$ while $\hat{r}^{Q} = 2$. This confirms the expected result previously stated. The results are insensitive to the choice of h_0 as already noticed by Lam et al. [11].

Fig. 6(a) and (b) plot the two estimated factor time series $\hat{X}_{1,t}$ and $\hat{X}_{2,t}$, respectively, given by (3) where the columns of the estimated factor loading matrix $\hat{\boldsymbol{P}}$ are the $\hat{r}^{\mathrm{Q}}=2$ orthonormal eigenvectors of $\hat{\boldsymbol{M}}^{\mathrm{Q}}$ corresponding to its $\hat{r}^{\mathrm{Q}}=2$ largest eigenvalues.

Following similar lines as in Lam and Yao [10, Section 5], we calculate the percentage of the variability of the pollutant $\boldsymbol{Z}_t$ explained by $\hat{\boldsymbol{P}}\hat{\boldsymbol{X}}_t$. For this, the PM_{10} concentration at Laranjeiras station is used. The measured data and the estimated one by the linear combination of the two estimated factors are displayed in Fig. 6(c) and (d), respectively. There is no apparent difference between these two plots, including during the high volatility and large peaks periods of PM_{10} concentrations. The quantity $\|\boldsymbol{B}\boldsymbol{u}\|^2/\|\boldsymbol{u}\|^2=0.0015$, where $\boldsymbol{u}$ is the vector of the 731 observations at Laranjeiras station and $\boldsymbol{B}$ is the projection matrix onto the orthogonal complement of the linear space spanned by the two vectors $(\hat{X}_{1,1},\ldots,\hat{X}_{1,731})$ and $(\hat{X}_{2,1},\ldots,\hat{X}_{2,731})$. Then, 99.85% of the PM_{10} concentrations of Laranjeiras station can be explained linearly by the two estimated factors. Finally, for forecasting purpose, this is simpler to use (1) than fitting a multivariate stationary time series model with dimension $k=8$ to $\boldsymbol{Z}_t$. The h-step ahead forecast $\hat{\boldsymbol{Z}}_{n+h}^{(h)}$ of $\boldsymbol{Z}_n$ is obtained by $\hat{\boldsymbol{Z}}_{n+h}^{(h)}=\hat{\boldsymbol{P}}\hat{\boldsymbol{X}}_{n+h}^{(h)}$, where $\hat{\boldsymbol{X}}_{n+h}^{(h)}$ is the h-step ahead forecast for $\boldsymbol{X}_n$, based on the estimated past values $\hat{X}_1,\ldots,\hat{X}_n$, see Lam et al. [11].

6. Conclusions

In this paper, a robust FA method for high-dimensional time series with additive outliers is proposed. Some theoretical results are discussed and verified through Monte Carlo experiments. The simulations show that additive outliers reduce the classical estimated factor dimension. The robust method presents better performance and appears as an alternative method when there is any evidence of atypical observations in the multivariate time series data, such as high levels of the pollutants in the pollution area. The proposed methodology was used to identify pollution behaviour of the pollutant PM_{10}, which can be very useful for the management of the air quality network.

7. Proofs

Proof of Lemma 1. By Weyl's Theorem, see Horn and Johnson [8, p. 239], for all $j=1,\ldots,k$, it follows that

$$\lambda_j(\hat{A})-\lambda_j(A)\leq\lambda_k(\hat{A}-A)\leq\sup_{1\leq\ell\leq k}|\lambda_\ell(\hat{A}-A)|.$$

By exchanging the role of $\hat{A}$ and A, for all $j=1,\ldots,k$, it follows that

$$\lambda_j(A)-\lambda_j(\hat{A})\leq\sup_{1\leq\ell\leq k}|\lambda_\ell(\hat{A}-A)|.$$

Hence,

$$\sup_{1\leq j\leq k}|\lambda_j(\hat{A})-\lambda_j(A)|\leq\sup_{1\leq\ell\leq k}|\lambda_\ell(\hat{A}-A)|=\|\hat{A}-A\|_2,$$

where $\|X\|_2$ denotes the largest absolute value of the eigenvalues of a matrix X. Since $u_n(\hat{A}_n-A)=O_p(1)$, the result follows. □

Proof of Lemma 2. The proof of this lemma directly follows from the application of the continuous mapping theorem; see van der Vaart [28, Theorem 2.3]. □

Proof of Lemma 3. Observe that the autocovariance of the process $(Z_{i,t}+Z_{j,t+h})_{t\geq1}$ at lag ℓ is equal to

$$\gamma_{i,j}^{(+)}(\ell)=\mathrm{Cov}[Z_{i,t}+Z_{j,t+h},Z_{i,t+\ell}+Z_{j,t+h+\ell}]=\boldsymbol{\Gamma}_{i,i}^{Z}(\ell)+\boldsymbol{\Gamma}_{i,j}^{Z}(h+\ell)+\boldsymbol{\Gamma}_{j,i}^{Z}(\ell-h)+\boldsymbol{\Gamma}_{j,j}^{Z}(\ell),$$

and that the autocovariance of the process $(Z_{i,t}-Z_{j,t+h})_{t\geq1}$ at lag ℓ is equal to

$$\gamma_{i,j}^{(-)}(\ell)=\mathrm{Cov}[Z_{i,t}-Z_{j,t+h},Z_{i,t+\ell}-Z_{j,t+h+\ell}]=\boldsymbol{\Gamma}_{i,i}^{Z}(\ell)-\boldsymbol{\Gamma}_{i,j}^{Z}(h+\ell)-\boldsymbol{\Gamma}_{j,i}^{Z}(\ell-h)+\boldsymbol{\Gamma}_{j,j}^{Z}(\ell).$$

By (A2) and (10), $\sum_{\ell\geq1}|\gamma_{i,j}^{(+)}(\ell)|<\infty$ and $\sum_{\ell\geq1}|\gamma_{i,j}^{(-)}(\ell)|<\infty$. The proof of this lemma, thus, follows the same lines as the ones of Lévy-Leduc et al. [14, Theorem 2] by replacing X_i and X_{i+h} by $Z_{i,t}$ and $Z_{j,t+h}$, respectively, and the summations on i by summations on t which leads to

$$\sqrt{n-h}\big(\hat{\gamma}_{i,j}^{\mathrm{Q}}(h)-\boldsymbol{\Gamma}_{i,j}^{Z}(h)\big)=\frac{1}{\sqrt{n-h}}\sum_{t=1}^{n-h}\psi(Z_{i,t},Z_{j,t+h})+o_P(1),$$

where

$$\psi(x,y)=\frac{1}{2}\big(\boldsymbol{\Gamma}_{i,i}^{Z}(0)+\boldsymbol{\Gamma}_{j,j}^{Z}(0)+\boldsymbol{\Gamma}_{i,j}^{Z}(h)+\boldsymbol{\Gamma}_{j,i}^{Z}(-h)\big)\mathrm{IF}\left(\frac{x+y}{\sqrt{\boldsymbol{\Gamma}_{i,i}^{Z}(0)+\boldsymbol{\Gamma}_{j,j}^{Z}(0)+\boldsymbol{\Gamma}_{i,j}^{Z}(h)+\boldsymbol{\Gamma}_{j,i}^{Z}(-h)}},\mathrm{Q},\Phi\right)$$

$$-\frac{1}{2}\left(\mathbf{\Gamma}^Z_{i,i}(0)+\mathbf{\Gamma}^Z_{j,j}(0)-\mathbf{\Gamma}^Z_{i,j}(h)-\mathbf{\Gamma}^Z_{j,i}(-h)\right)\mathrm{IF}\left(\frac{x-y}{\sqrt{\mathbf{\Gamma}^Z_{i,i}(0)+\mathbf{\Gamma}^Z_{j,j}(0)-\mathbf{\Gamma}^Z_{i,j}(h)-\mathbf{\Gamma}^Z_{j,i}(-h)}},Q,\Phi\right), \quad (11)$$

and IF is defined in Equation (20) of Lévy-Leduc et al. [14]. By applying Arcones [2, Theorem 4], the result is obtained. □

Acknowledgements

Part of the simulation and application results in this paper are in chapters of the Ph.D. thesis of the second author under the supervision Prof. V. A. Reisen. The authors would like to thank CNPq (grant no. 504726/2007-2) and FAPES (grant no. 007/2014) for their financial support. Part of this paper was revised when Prof. Valdério Reisen was visiting CentraleSupélec (from December 2016 to January 2017 and in July 2018). This author is indebted to CentraleSupélec for its financial support. The authors are grateful to the referee for the time and efforts in providing helpful comments and additional references that have led to clarify and substantially improve the quality of the paper.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd ed., John Wiley & Sons, New Jersey, 2003.
[2] M.A. Arcones, Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors, Ann. Probab. 22 (4) (1994) 2242–2274.
[3] B. Brunekreef, S.T. Holgate, Air pollution and health, Lancet 360 (9341) (2002) 1233–1242.
[4] I. Chang, G.C. Tiao, C. Chen, Estimation of time series parameters in the presence of outliers, Technometrics 30 (2) (1988) 193–204.
[5] C. Chen, L.-M. Liu, Joint estimation of model parameters and outlier effects in time series, J. Am. Stat. Assoc. 88 (421) (1993) 284–297.
[6] L. Curtis, W. Rea, P. Smith-Willis, E. Fenyves, Y. Pan, Adverse health effects of outdoor air pollutants, Environ. Int. 32 (6) (2006) 815–830.
[7] M. Gosak, A. Stožer, R. Markovič, J. Dolenšek, M. Marhl, M. Slak Rupnik, M. Perc, The relationship between node degree and dissipation rate in networks of diffusively coupled oscillators and its significance for pancreatic beta cells, Chaos: Interdiscipl. J. Nonlinear Sci. 25 (7) (2015) 073115.
[8] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 1985. Cambridge Books Online
[9] R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, 6th ed., Prentice Hall, New Jersey, 2007.
[10] C. Lam, Q. Yao, Factor modeling for high-dimensional time series: inference for the number of factors, Ann. Stat. 40 (2) (2012) 694–726.
[11] C. Lam, Q. Yao, N. Bathia, Estimation of latent factors for high-dimensional time series, Biometrika 98 (2011) 901–918.
[12] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Asymptotic properties of U-processes under long-range dependence, Ann. Stat. 39 (3) (2011a) 1399–1426.
[13] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Large sample behaviour of some well-known robust estimators under long-range dependence, Statistics 45 (1) (2011b) 59–71.
[14] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Robust estimation of the scale and the autocovariance function of Gaussian short and long-range dependent processes, J. Time Ser. Anal. 32 (2) (2011c) 135–156.
[15] O. Lordan, J.M. Sallan, N. Escorihuela, D. Gonzalez-Prieto, Robustness of airline route networks, Phys. A: Stat. Mech. Appl. 445 (2016) 18–26, doi:10.1016/j.physa.2015.10.053.
[16] Y. Ma, M.G. Genton, Highly robust estimation of the autocovariance function, J. Time Ser. Anal. 21 (2000) 663–684.
[17] Y. Ma, M.G. Genton, Highly robust estimation of dispersion matrices, J. Multivar. Anal. 78 (2001) 11–36.
[18] R. Maynard, Key airborne pollutants: the impact on health, Sci. Total Environ. 334–335 (0) (2004) 9–13.
[19] F.F. Molinares, V.A. Reisen, F. Cribari-Neto, Robust estimation in long-memory processes under additive outliers, J. Stat. Plann. Inference 139 (8) (2009) 2511–2525.
[20] D. Pea, G.E.P. Box, Identifying a simplifying structure in time series, J. Am. Stat. Assoc. 82 (399) (1987) 836–843.
[21] M. Perc, Nonlinear time series analysis of the human electrocardiogram, Eur. J. Phys. 26 (5) (2005) 757.
[22] V.A. Reisen, C. Lévy-Leduc, M. Bourguignon, H. Boistard, Robust Dickey–Fuller tests based on ranks for time series with additive outliers, Metrika 80 (1) (2017) 115–131.
[23] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, J. Am. Stat. Assoc. 88 (424) (1993) 1273–1283.
[24] J.H. Seinfeld, S.N. Pandis, Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, John Wiley, New York, 2006.
[25] J.B. Souza, V.A. Reisen, G.C. Franco, M. Spány, P. Bondon, J.M. Santos, Generalized additive model with principal component analysis: an application to time series of respiratory disease and air pollution data, J. R. Stat. Soc. Ser. C – Appl. Stat. 67 (2018) 453–480.
[26] J.H. Stock, M.W. Watson, Forecasting using principal components from a large number of predictors, J. Am. Stat. Assoc. 97 (460) (2002) 1167–1179.
[27] R.S. Tsay, Outliers, level shifts, and variance changes in time series, J. Forecast. 7 (1) (1988) 1–20.
[28] A.W. van der Vaart, Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.
[29] E. Vanhatalo, M. Kulahci, Impact of autocorrelation on principal components and their use in statistical process control, Qual. Reliab. Eng. Int. 32 (4) (2016) 1483–1500.
[30] J.G. Watson, T. Zhu, J.C. Chow, J. Engelbrecht, E.M. Fujita, W.E. Wilson, Receptor modeling application framework for particle source apportionment, Chemosphere 49 (9) (2002) 1093–1136.
[31] WHO, Air Quality Guidelines: global update 2005, WHO – World Health Organization, 2006.
[32] WHO, Air Pollution Estimates, WHO – World Health Organization, 2014.
[33] B. Zamprogno, PCA applied in time series data with applications to air quality data, PPGEA – Universidade Federal do Espírito Santo, 2013 (Ph.D. thesis). In press
[34] M. Zhang, B. Liang, S. Wang, M. Perc, W. Du, X. Cao, Analysis of flight conflicts in the chinese air route network, Chaos Solitons Fractals 112 (2018) 97–102, doi:10.1016/j.chaos.2018.04.041.