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

Valdério Anselmo Reisen, Adriano Marcio Sgrancio, Céline Lévy-Leduc, Edson Zambon Monte, Higor Henrique Aranda Cotta, Pascal Bondon, Flávio Augusto Ziegelmann

To cite this version:

Valdério Anselmo Reisen, Adriano Marcio Sgrancio, Céline Lévy-Leduc, Edson Zambon Monte, Higor Henrique Aranda Cotta, et al.. Robust factor modelling for high-dimensional time series: An application to air pollution data. Applied Mathematics and Computation, Elsevier, 2019, 346, pp.842-852. 10.1016/j.amc.2018.09.062 . hal-02902032

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

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

^{a} *PhD Program in Environmental Engineering, Federal University of Espírito Santo, Esprı́to Santo, Brazil*

^{b} *Department of Statistics, Federal University of Espírito Santo, Espírito Santo, Brazil*

^{c} *AgroParisTech/UMR INRA MIA 518, France*

^{d} *Laboratoire des Signaux et Systèmes, CNRS - CentraleSupélec - Université Paris-Sud - France*

^{e} *Department of Economics, Federal University of Espírito Santo, Esprı́to Santo, Brazil*

^{f} *Department of Statistics, Ppge and Ppga, Federal University of Rio Grande do Sul, Rio Grande do Sul, Brazil*

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 [12]. The estimator of the number of factors is obtained by an eigenanalysis 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. In the 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}.

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

[*] Corresponding author. Department of Statistics, Federal University of Espírito Santo, 29075-910, 514, Vitoria, ES, Brazil. Tel.: +5502740092903.
E-mail address: valderioanselmoreisen@gmail.com, valderio.reisen@ufes.br (V. A. Reisen).

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 [33]. 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 [20], WHO [32], Curtis et al. [7] and Souza et al. [26] 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.

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 provide the identification of the pollutant emission sources from chemical characteristics of the particles on the receiver and the pollutant emission sources see, for example, Seinfeld and Pandis [25]. In the literature, the most studied receptor models are: chemical mass balance (CMB), multivariate analysis, principal component analysis techniques (PCA), factor analysis model (FA), multiple linear regression, cluster analysis and positive matrix factorization (PMF) (Watson et al. [31]). In particular, the classical factor analysis methodology 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 factor analysis techniques, the time dependence of the data is a commonly neglected feature. A basic assumption of the multivariate statistical tools is that the data are independent in time (see for example Anderson [1] and Johnson and Wichern [11]). To deal with autocorrelated data in factorial analysis, Peña and Box [22], Stock and Watson [27], Lam et al. [13] and Lam and Yao [12] studied the factor modelling for multivariate time series from a dimension-reduction point of view. Contrarily to the PCA and factor analysis 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 [12].

Since factor analysis allows to reduce the order of the estimated model, this technique has been widely used for forecasting. According to Stock and Watson [27], 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 two-step approach: 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 outlying values from an aspect of statistical analysis. 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 (Perc [23], Gosak et al. [8], Souza et al. [26]), air route network problems (Zhang et al. [35], Lordan et al. [17]), environmental engineering (Zamprogno [34]) and statistical process controls (Vanhatalo and Kulahci [30]), to name a few.

As is well known, outliers can destroy the statistical properties of the estimates such as the sample mean and sample covariance (see, for example, Chang et al. [4], Tsay [28], 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 [19] proposed a robust sample autocorrelation function (ACF) based on the robust scale estimate $Q_n(.)$ suggested in Rousseeuw and Croux [24]. This robust ACF estimator was recently studied by Lévy-Leduc et al. [14], Lévy-Leduc et al. [15] and Lévy-Leduc et al. [16].

This paper considers multivariate time series with additive outliers using the factor analysis technique for dimension reduction. In this context, a robust version of the dimension reduction estimator given in Lam and Yao [12] is proposed here. 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 Air Quality Automatic Monitoring Network (AQAMN), 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 and estimation methods

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

Let k be the number of candidate predictor series. Let $\boldsymbol{Z}_t$, $t \in \mathbb{Z}$, be a k-dimensional zero-mean vector of an observed time series. Let also $\boldsymbol{X}_t$ 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, denoted by the factor-loading matrix, and $\boldsymbol{\varepsilon}_t$ is a k-dimensional white-noise sequence with full-rank covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. When r is small relative to k, the model presented in Equation (1) is most useful, since it will result in a multivariate time series model with a reduced dimension and, consequently, can lead to a much simpler multivariate time series model for forecasting.

In the sequel, the following assumption is made.

(A1) $\boldsymbol{X}_t$, $t \in \mathbb{Z}$, is a multivariate stationary process and $\boldsymbol{\varepsilon}_t \sim WN(\boldsymbol{0}, \boldsymbol{\Sigma}_{\varepsilon})$. Moreover, $\boldsymbol{X}_t$ and $\boldsymbol{\varepsilon}_t$ are assumed to be uncorrelated and $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_r$, where $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix.

Note that Assumption 1 is to ensure identifiability in Equation (1); see Lam and Yao [12] and Peña and Box [22] for further details.

It follows from Equation (1) and under Assumption (A1) that the covariance matrices of $\boldsymbol{Z}_t$ are given by

$$\boldsymbol{\Gamma}_Z(0) = \boldsymbol{P}\boldsymbol{\Gamma}_X(0)\boldsymbol{P}' + \boldsymbol{\Sigma}_\varepsilon, \tag{2}$$

$$\boldsymbol{\Gamma}_Z(h) = \boldsymbol{P}\boldsymbol{\Gamma}_X(h)\boldsymbol{P}', \quad h \geq 1, \tag{3}$$

where $\boldsymbol{\Gamma}_X(h) = \mathbb{E}[\boldsymbol{X}_{t-h}\boldsymbol{X}_t']$ is the covariance matrix of $\boldsymbol{X}_t$.

Based on the observations $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$, the first step in Equation (1) is to estimate the number of factors r in order to compute the estimate of the $k \times r$ factor loading matrix $\boldsymbol{P}$. Once $\hat{\boldsymbol{P}}$ is obtained, the estimator for the factor process and the residuals are respectively given by

$$\hat{\boldsymbol{X}}_t = \hat{\boldsymbol{P}}'\boldsymbol{Z}_t, \tag{4}$$

and

$$\hat{\boldsymbol{\epsilon}}_t = (\boldsymbol{I}_d - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{Z}_t. \tag{5}$$

For further details on the estimation of $\boldsymbol{P}$, see Lam and Yao [12].

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

$$\hat{\boldsymbol{M}} = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}_Z(h)\hat{\boldsymbol{\Gamma}}_Z(h)' \tag{6}$$

for a prescribed integer $h_0 \geq 1$. Following the lines of Lam and Yao [12], the estimator for the number of factors r is given by:

$$\hat{r} = \operatorname*{argmin}_{1 \leq i \leq R} \hat{\lambda}_{i+1}/\hat{\lambda}_i, \tag{7}$$

where $r < R < k$ is a constant, $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_k$ are the eigenvalues of $\hat{\boldsymbol{M}}$. Under some assumptions, Lam and Yao [12] derive the asymptotic properties of the above results. According to Lam and Yao [12], in practice, for example, $R = p/2$ may be used. Note that, there is no need to extend the test up to p since it is expected that the minimum eigenvalue of $\hat{\boldsymbol{M}}$ be practically 0, specially for a small n and a large p. It is important to say that when p and n are of the same order, the estimators for eigenvalues are no longer consistent, although the ratio based estimator still can be used, see Lam and Yao [12].

As previously stated, the aim of this paper is to propose robust estimators of $\boldsymbol{M}$ and r which are based on a robust covariance matrix estimator for $\boldsymbol{Z}_t$. These issues are discussed in the following section.

2.1.1. The robust estimator of the number of factors

Let $(Y_i)_{i\geq 1}$ be a stationary Gaussian process. Given the observations $Y_{1:n} = (Y_1, \ldots, Y_n)$, the estimator of scale proposed by Rousseeuw and Croux [24] is defined by

$$Q_n(Y_{1:n}) = c\left\{\left|Y_i - Y_j\right|;\ 1 \leq i, j \leq n\right\}_{(\lfloor n^2/4 \rfloor)}, \tag{8}$$

where $c = 2.21914$.

Now, consider the following assumption on $\boldsymbol{X}_t$.

(**A2**) $\boldsymbol{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{r,t})'$ is a multivariate stationary zero-mean Gaussian process satisfying

$$\sum_{h\geq 1} |\gamma^X_{i,j}(h)| < \infty,\ \text{for all } i, j \in \{1, \ldots, r\},$$

where $\gamma^X_{i,j}(h) = \mathbb{C}ov[X_{i,t}, X_{j,t+h}]$.

By using Equations (1) and (3), $(\boldsymbol{Z}_t)$ is also a multivariate stationary zero-mean Gaussian process satisfying

$$\sum_{h\geq 1} |\gamma_{i,j}(h)| < \infty,\ \text{for all } i, j \in \{1, \ldots, k\}, \tag{9}$$

where $\gamma_{i,j}(h) = \mathbb{C}ov[Z_{i,t}, Z_{j,t+h}]$.

From the estimator Q_n defined in Equation (8) and from the observations $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$, it is proposed here a robust estimator of $\gamma_{i,j}(h) = \mathbb{C}ov(Z_{i,t}, Z_{j,t+h})$ for all i, j in $\{1, \ldots, k\}$ defined as follows

$$\hat{\gamma}^Q_{i,j}(h) = \frac{1}{4}\left[Q^2_{n-h}(Z_{i,1:n-h} + Z_{j,h+1:n}) - Q^2_{n-h}(Z_{i,1:n-h} - Z_{j,h+1:n})\right], \tag{10}$$

where $Z_{i,1:n-h} = (Z_{i,1}, \ldots, Z_{i,n-h})$ and $Z_{j,h+1:n} = (Z_{j,h+1}, \ldots, Z_{j,n})$. $\hat{\gamma}^Q_{i,j}(h)$ is the multivariate estimator of the univariate case suggested by Ma and Genton [18].

From Equation (10), the robust estimator of the covariance matrix of $\boldsymbol{Z}_t$ is given by

$$\hat{\boldsymbol{\Gamma}}_Q(h) = \begin{pmatrix} \hat{\gamma}^Q_{1,1}(h) & \hat{\gamma}^Q_{1,2}(h) & \ldots & \hat{\gamma}^Q_{1,k}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}^Q_{k,1}(h) & \hat{\gamma}^Q_{k,2}(h) & \ldots & \hat{\gamma}^Q_{k,k}(h) \end{pmatrix}. \tag{11}$$

Based on Equations (6) and (11), the robust version of the estimator $\hat{\boldsymbol{M}}$ is suggested here as follows

$$\hat{\boldsymbol{M}}_Q = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}_Q(h)\hat{\boldsymbol{\Gamma}}_Q(h)'. \tag{12}$$

Therefore, the robust estimator $\hat{r}_Q$ of r is similarly obtained from Equation (7) by replacing $\hat{\lambda}_{i+1}$ and $\hat{\lambda}_i$ by $\hat{\lambda}^Q_{i+1}$ and $\hat{\lambda}^Q_i$, respectively, where $(\hat{\lambda}^Q_i)_{1\leq i\leq k}$ are the eigenvalues of $\hat{\boldsymbol{M}}_Q$.

3. Theoretical results

This section provides some analytical results to theoretically support the robust approach discussed in the previous section.

Theorem 1. *Let h be a fixed positive integer and $\left(\hat{\boldsymbol{\Gamma}}_Q(h)\right)_{1\leq i,j\leq k} = \left(\hat{\gamma}^Q_{i,j}(h)\right)_{1\leq i,j\leq k}$, where $\hat{\gamma}^Q_{i,j}(h)$ is defined in Equation (10). Assume that Assumptions (A1) and (A2) hold, then*

$$\sqrt{n} \sup_{1\leq j\leq k} \left|\hat{\lambda}^Q_j - \lambda_j\right| = O_p(1), \text{ as } n \to \infty,$$

where $(\hat{\lambda}^Q_j)_{1\leq j\leq k}$ and $(\lambda_j)_{1\leq j\leq k}$ denote the eigenvalues of $\left(\sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}_Q(h)\hat{\boldsymbol{\Gamma}}_Q(h)'\right)$ and $\left(\sum_{h=1}^{h_0} \boldsymbol{\Gamma}(h)\boldsymbol{\Gamma}(h)'\right)$, respectively, where $(\boldsymbol{\Gamma}(h))_{1\leq i,j\leq k} = \left(\gamma_{i,j}(h)\right)_{1\leq i,j\leq k}$ and h_0 is a fixed integer larger than 1.

The proof of this theorem directly follows from Lemmas 1, 2 and 3 given below and proved in Section 8.

Remark 1. *By Theorem 1 and Lam and Yao [12, Proposition 1], it can be seen that the eigenvalues of the robust estimator covariance matrix of $\mathbf{Z}_t$ have the same rate of convergence as the eigenvalues of the standard estimator of the covariance matrix of $\mathbf{Z}_t$.*

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 $u_n(\hat{A}_n - A) = O_p(1)$, where u_n is a sequence of positive numbers tending to infinity as n tends to infinity, then*

$$u_n \sup_{1\leq j\leq p} |\lambda_j(\hat{A}) - \lambda_j(A)| = O_p(1), \quad \text{as } n \to \infty,$$

where $(\lambda_j(\hat{A}))_{1\leq j\leq k}$ and $(\lambda_j(A))_{1\leq j\leq k}$ 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 $u_n(\hat{A}_n(h) - A(h)) = O_p(1)$, for each fixed $h \in \{1, \ldots, h_{max}\}$, where u_n is a sequence of positive numbers tending to infinity as n tends to infinity, then*

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

Lemma 3. *Let h be a non negative integer and i and j be two integers in $\{1, \ldots, k\}$. Assume that Assumptions (A1) and (A2) hold, then the robust autocovariance estimator $\hat{\gamma}^{Q}_{i,j}(h)$ defined in Equation (10) satisfies the following central limit theorem*

$$\sqrt{n}(\hat{\gamma}^{Q}_{i,j}(h) - \gamma_{i,j}(h)) \xrightarrow{d} \mathcal{N}(0, \widetilde{\sigma}^2_{i,j}(h)), \text{ as } n \to \infty,$$

where

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

where ψ is defined in Equation (15).

4. Monte Carlo studies

This section reports simulation results related to the performance of the proposed methodology previously discussed for finite sample size. In this context, the empirical study considered $\boldsymbol{X}_t$ as a VAR(1) model with $r = 3$. The VAR(1) model was generated with independent white noise vector from $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ and the matrix of the coefficients $\boldsymbol{\Phi}_1$. This matrix is displayed in Table 1. Note that the $\boldsymbol{\Phi}$ coefficients corresponds to a process with no temporal correlation outside the main diagonal. The sample sizes are $n = 50, 100, 200, 400, 800$ and 1600, and $k = 0.2n, 0.5n, 0.8n$, and $h_0 = 1$ (these are also considered in Lam and Yao [12]). The factor model (Equation (1)) was generated as follows: first, all $k \times r$ elements of matrix $\boldsymbol{P}$ were generated as independent observations from the uniform distribution on the interval $[-1, 1]$ (see, also, Lam and Yao [12]). The process $\boldsymbol{\varepsilon}_t$ in Equation (1) consists of $\mathrm{N}(0, 1)$ components independent between each other and across t. The statistical quantities were computed based on 1000 replications. These simulations were ran using the `R` programming language and the code is available upon request.

[Table 1 about here.]

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 VAR(1) models with and without outliers. For this, the estimate of $P(\hat{r} = r)$ is obtained by computing the relative frequencies of $\hat{r} = r$, denoted here as $f_{freq.}(\hat{r} = r)$, where $\hat{r}$ is the estimator of r. The $\hat{r}_Q$ estimator was similarly computed.

Table 2 reports the relative frequency estimates ($f_{freq.}(\hat{r} = 3)$) for $P(\hat{r} = 3)$ of the model in the study. From this, it is observed that the ratio-based estimator of r improves when n is increased. Similar performance of the ratio is also observed when the dimension k increases. These are in accordance with the results given in Table 1 of Lam and Yao [12], that is, the asymptotic properties of $\hat{r}$ is corroborated with these finite sample size investigations.

The results related to the alternative method $\hat{r}_Q$ are displayed in Table 3. It shows that this estimator presents similar empirical performance of $\hat{r}$, which is an expected result based on the asymptotic theory of both estimators of the ratio (see Remark 1). Therefore, this simple empirical study together with the theoretical results discussed in the previous section give support to using $\hat{r}_Q$ as an alternative method to estimate the number of the factors in the model presented in Equation (1).

[Table 2 about here.]

[Table 3 about here.]

Now, let $\mathbf{X^*}_t, t \in \mathbb{Z}$, be a vector process contaminated by additive outliers defined as follows

$$\mathbf{X^*}_t = \boldsymbol{X}_t + \boldsymbol{\omega} \circ \boldsymbol{\delta}_t, \qquad (13)$$

where "$\circ$" is the Hadamard product (Johnson [10]). The vector $\boldsymbol{\omega} = (\omega_1, ..., \omega_k)'$ is a magnitude vector of additive outliers and the random vector $\boldsymbol{\delta}_t = (\delta_{1,t}, ..., \delta_{k,t})'$ indicates the occurrence of an outlier at time t, in variable $i, i = 1, ..., k$, such as $\mathbb{P}(\delta_{i,t} = -1) = \mathbb{P}(\delta_{i,t} = 1) = p/2$ and $\mathbb{P}(\delta_{i,t} = 0) = 1 - p$, where $\mathbb{E}[\delta_{i,t}] = 0$ and $\mathbb{E}[\delta^2_{i,t}] = \mathbb{V}ar[\delta_{i,t}] = p$. The model described above assumes that $\mathbf{X^*}_t$ and $\boldsymbol{\delta}_t$ are independent processes. Also, it is assumed that the elements of $\boldsymbol{\delta}_t$ are not correlated and temporally uncorrelated, i.e., $\mathbb{E}[\boldsymbol{\delta}_t \boldsymbol{\delta}^{'}_t] = \Sigma_\delta = diag(p, ..., p)$ and $\mathbb{E}[\boldsymbol{\delta}_t \boldsymbol{\delta}^{'}_{t+h}] = 0$ for $h \neq 0$.

Remark 2. *$\delta_{i,t}, i = 1, ..., k$, is the product of Bernoulli(p) random variable with Rademacher random variable, the latter equals 1 or -1, both with probability 1/2.*

In this empirical investigation, the probability of an outlier occurring at time t is $p = 0.05$ and, without loss of generality, it is also assumed that $\boldsymbol{\omega} = [\omega_1, 0, 0]'$; that is, $X^*_{1,t}$, $t = 1, ..., n$, is the only process in $\mathbf{X^*}_t = (X^*_{1,t}, X^*_{2,t}, X^*_{3,t})^{'}$ contaminated with outliers and $\omega_1 = 15$.

Table 4 shows the relative frequency estimates, $\hat{r}$ and $\hat{r}_Q$, for the dimensional reduction for Model 1, when $r = 3$ and in the presence of outliers. The standard case, that is, $\hat{\boldsymbol{\Gamma}}_{X^*}$ and $p = 0$, is in accordance with the results given by Tables 2 and 3 for $\hat{r}$ and $\hat{r}_Q$, respectively.

From the cases where there are outliers (column of Table 4 in which $p \neq 0$), it is clearly perceived that the standard method is destroyed by a single outlier at the time t by presenting a substantial reduction of the correct estimated ratio. This is an expected result since $\hat{r}$ is based on the standard sample matrix of autocorrelation. Note that, according the simulation plan, the expected number of outliers in the whole series $X^*_{1,t}$ is 5. The conclusions of the empirical study in this section corroborate the Proposition 1 in Cotta et al. [6], which demonstrates theoretically that one outlier is enough to destroy the properties of the sample autocorrelation function of multivariate processes.

In contrast with the empirical performance of $\hat{r}$, $\hat{r}_Q$ keeps almost unchanged with the proportion of the outliers in the present study. The estimated ratios are very close in the contaminated and uncontaminated series. Therefore, the percentage of outliers in this empirical example seems to be, in general, not strong enough to destroy the robustness property of $\hat{r}_Q$ and this indicates that the methodology proposed here may be used when the presence of outliers in the series is uncertain. Other simulation cases with different degrees of contamination present similar conclusions and are available upon request.

[Table 4 about here.]

5. Application to the pollutant PM$_{10}$

This section presents an application of the methodology discussed previously for PM$_{10}$ pollutant concentrations measured at the Automatic Quality Monitoring Network (AAQMN) in the Greater Vitória Area (GVR), Espírito Santo, Brazil. GVR is comprised of seven cities with a population of approximately 1.9 million inhabitants in the area of 2,319 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 $\mu g/m^3$ and was hourly measured from January 2008 to December 2009, $k = 8$, though the daily average values ($n = 731$) are used in this empirical study. This follows the same lines of the application given in Lam and Yao [12]. Now let $\boldsymbol{Z}_t$, $t = 1, ..., 731$, be the vector of the PM$_{10}$ concentrations, that is, $\boldsymbol{Z}_t = (Z_{1,t}, ..., Z_{8,t})^{'}]$ where $Z_{i,t}$ corresponds to PM$_{10}$ concentration at location i.

Figure 1 shows the plots of the PM$_{10}$ concentrations for the eight stations. Based on this figure, the series indicated that they present high levels of pollutant concentrations which can be identified, from the statistical point of view, as additive outliers. This is justified by the fact that they produce similar impact on the sample ACF caused by additive outliers, that is, they lead to a reduction of the sample autocorrelation values. Therefore, the high values can be seen as (additive) outliers. Hence, both methods, the robust and non-robust approaches discussed previously, are used here to verify whether or not these high levels make any impact on the factor model estimation.

The sample ACFs displayed in Figures 2 (the classical ACF estimator) and 3 (the robust ACF estimator) show possible seasonal pattern of period $s = 7$ which is an expected evidence since the data are daily averaged. In terms of the comparison between the sample ACF estimates, as a simple case, the classical sample ACF values at VVCentro station are 0.47, 0.12, 0.15 and 0.13 for lags $h = 1, 3, 5, 10$, respectively, while the ACF based on the Q_n function are 0.54, 0.25, 0.20 and 0.19. This simple case shows that the high levels of PM$_{10}$ at VVCentro station indicated an reduction of the sample ACFs values of the classical autocorrelation estimator.

The effect of atypical observations on the estimation of the ACF function is well discussed in Molinares et al. [21] for a single time series. The comparison between the sample ACFs of PM_{10} concentrations from the other stations presented similar conclusions.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

From the above discussion, it is expected that the FA estimated model will show different performance for the two methodologies, that is, for the standard and robust ones. The estimates of the number of factors r were computed by performing an eigenanalysis on $\hat{\boldsymbol{M}}$ and on $\hat{\boldsymbol{M}}_Q$ of Equations (6) and (12), respectively, with $h_0 = 7$, to capture the seasonality feature of the data set. The eigenvalues (the scree plot), in decreasing order, and their ratios obtained using $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{Z}}$ are shown in Figure 4 (a) and (b), respectively. The corresponding robust versions, i.e., using $\hat{\boldsymbol{\Gamma}}_Q$, are shown in Figure 5 (a) and (b), respectively. As can be seen, the robust and the classical ACF estimators suggest $\hat{r} = 2$ and $\hat{r} = 1$, respectively. This confirms the expected result previously stated. The reduction was not affected when varying the value of h_0.

[Figure 4 about here.]

[Figure 5 about here.]

Figure 6 (a) and (b) display the time series plots of the two components series $\hat{X}_{1,t}$ and $\hat{X}_{2,t}$, respectively, of the estimated factors $\hat{X}_t = (\hat{X}_{1,t}, \hat{X}_{2,t})'$ for $\hat{\boldsymbol{\Gamma}}_{Z,Q}$ (see Equation (4)).

Following similar lines as in the application section of Lam and Yao [12], it is now addressed the study to verify the percentage of the variability of the pollutant vector $\boldsymbol{Z}_t$ explained by $\hat{\boldsymbol{P}}\hat{\boldsymbol{X}}_t$. For this, the PM_{10} concentration measured at the Laranjeiras station is used and the original data and the estimated one are displayed in Figures 6 (c) and (d), respectively. From the plots of Figure 6, it is possible to note that the behaviour of the two factors is similar to the ones of the Laranjeiras station, including the high volatility and the periods with peaks of PM_{10} concentrations. The estimated PM_{10} concentrations (Figure 6(d)) are a linear combination of the estimated factor coefficients with values 0.4753 and

0.8231, for the first and second factor, respectively. This estimated series behaves similarly to the observed values (Figure 6 (c)).

The quantity $||\boldsymbol{Bu}||^2/||\boldsymbol{u}||^2 = 0.0015$ where $\boldsymbol{u}$ is the 731×1 vector consisting of the PM_{10} concentrations of Laranjeiras station index over the same period of time, and $\boldsymbol{B}$ denotes the projection matrix onto the orthogonal complement of the linear space spanned by the two components series $\hat{X}_{1,t}$ and $\hat{X}_{2,t}$. From this, 99,85% of the PM_{10} concentrations of Laranjeiras station can be explained by a linear combination $\hat{X}_{1,t}$ and $\hat{X}_{2,t}$. Therefore, these analyses suggest the following factor model for the PM_{10} concentrations vector

$$\boldsymbol{Z}_t = \boldsymbol{p}_1 X_{1,t} + \boldsymbol{p}_2 X_{2,t} + \boldsymbol{\varepsilon}_t, \tag{14}$$

where $X_{1,t}$ denotes the first factor, $X_{2,t}$ is the second factor, and $\boldsymbol{\varepsilon}_t$ is a vector white-noise process.

Finally, for forecasting purpose, Equation (14) is a simpler model than a K-multivariate stationary time series model with dimension $k = 8$, that is, the h-step ahead forecast for the $\boldsymbol{Z}_t$ series is simplified using the formula $\hat{\boldsymbol{Z}}^{(h)}_{T+h} = \hat{\boldsymbol{P}}\hat{\boldsymbol{X}}^{(h)}_{T+h}$, where $\hat{\boldsymbol{X}}^{(h)}_{T+h}$ is an h-step ahead forecast for $\{\boldsymbol{X}_t\}$, based on the estimated past values $\hat{X}_1, \ldots, \hat{X}_T$ (see, Lam et al. [13]).

[Figure 6 about here.]

6. Conclusions

In this paper, a robust method for high-dimensional time series with additive outliers is proposed. Some theoretical results are discussed and verified through Monte Carlo experiments under different scenarios of outliers contamination. The simulations illustrated the effect of the additive outliers on the reduction of the factor dimension. The empirical investigation showed that the robust method presented better performance compared to the classic procedure of identifying the number of factors being 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. In addition, 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. The results in this paper will hopefully stimulate further research on this topic.

7. 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, CAPES and FAPES 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 a 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.

8. Proofs

Proof of Lemma 1. By Weyl's Theorem, see Horn and Johnson [9, p. 239], for all $j \in \{1, \ldots, k\}$, 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 $j \in \{1, \ldots, k\}$, 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 [29, Theorem 2.3]. □

Proof of Lemma 3. Observe that the autocovariance of the process $(Z_{i,t} + Z_{j,t+h})_{t \geq 1}$ at lag ℓ is equal to

$$\gamma_{i,j}^{(+)}(\ell) = \mathbb{C}ov[Z_{i,t} + Z_{j,t+h}, Z_{i,t+\ell} + Z_{j,t+h+\ell}] = \gamma_{i,i}(\ell) + \gamma_{i,j}(h+\ell) + \gamma_{j,i}(\ell - h) + \gamma_{j,j}(\ell),$$

and that the autocovariance of the process $(Z_{i,t} - Z_{j,t+h})_{t \geq 1}$ at lag ℓ is equal to

$$\gamma_{i,j}^{(-)}(\ell) = \mathbb{C}ov[Z_{i,t} - Z_{j,t+h}, Z_{i,t+\ell} - Z_{j,t+h+\ell}] = \gamma_{i,i}(\ell) - \gamma_{i,j}(h+\ell) - \gamma_{j,i}(\ell - h) + \gamma_{j,j}(\ell).$$

By A2 and Equation (9), $\sum_{\ell\geq 1}|\gamma_{i,j}^{(+)}(\ell)| < \infty$ and $\sum_{\ell\geq 1}|\gamma_{i,j}^{(-)}(\ell)| < \infty$. The proof of this lemma, thus, follows the same lines as the ones of Lévy-Leduc et al. [16, 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

$$\sqrt{n-h}\left(\hat{\gamma}_{i,j}^{Q}(h) - \gamma_{i,j}(h)\right) = \frac{1}{\sqrt{n-h}}\sum_{t=1}^{n-h}\psi(Z_{i,t}, Z_{j,t+h}) + o_P(1),$$

where

$$\begin{aligned}&\psi(x,y) = \\ &\frac{1}{2}\left(\gamma_{i,i}(0)+\gamma_{j,j}(0)+\gamma_{i,j}(h)+\gamma_{j,i}(-h)\right)\mathrm{IF}\left(\frac{x+y}{\sqrt{\gamma_{i,i}(0)+\gamma_{j,j}(0)+\gamma_{i,j}(h)+\gamma_{j,i}(-h)}}, Q, \Phi\right) \\ &-\frac{1}{2}\left(\gamma_{i,i}(0)+\gamma_{j,j}(0)-\gamma_{i,j}(h)-\gamma_{j,i}(-h)\right)\mathrm{IF}\left(\frac{x-y}{\sqrt{\gamma_{i,i}(0)+\gamma_{j,j}(0)-\gamma_{i,j}(h)-\gamma_{j,i}(-h)}}, Q, \Phi\right),\end{aligned} \tag{15}$$

where IF is defined in Equation (20) of Lévy-Leduc et al. [16]. By applying Arcones [2, Theorem 4], the result is obtained. □

References

References

[1] T.W. Anderson, An introduction to multivariate statistical analysis, John Wiley & Sons, New Jersey, 3rd edition, 2003.

[2] M.A. Arcones, Limit theorems for nonlinear functionals of a stationary gaussian sequence of vectors, Ann. Probab. 22 (1994) 2242–2274.

[3] B. Brunekreef, S.T. Holgate, Air pollution and health, The Lancet 360 (2002) 1233–1242.

[4] I. Chang, G.C. Tiao, C. Chen, Estimation of time series parameters in the presence of outliers, Technometrics 30 (1988) 193–204.

[5] C. Chen, L.M. Liu, Joint estimation of model parameters and outlier effects in time series, Journal of the American Statistical Association 88 (1993) 284–297.

[6] H.H.A. Cotta, V.A. Reisen, P. Bondon, W. Stummer, C. Lévy-Leduc, Robust estimation of covariance and correlation functions of a stationary multivariate process, in: Proceedings ITISE 2017, International work- conference on Time Series, 2017, Universidad de Granada.

[7] L. Curtis, W. Rea, P. Smith-Willis, E. Fenyves, Y. Pan, Adverse health effects of outdoor air pollutants, Environment International 32 (2006) 815–830.

[8] 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: An Interdisciplinary Journal of Nonlinear Science 25 (2015) 073115.

[9] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 1985. Cambridge Books Online.

[10] C. Johnson, Matrix theory and applications, American Mathematical Society, 1989.

[11] R. Johnson, D. Wichern, Applied multivariate statistical analysis, Prentice Hall, New Jersey, 6rd edition, 2007.

[12] C. Lam, Q. Yao, Factor modeling for high-dimensional time series: inference for the number of factors, Ann. Statist. 40 (2012) 694–726.

[13] C. Lam, Q. Yao, N. Bathia, Estimation of latent factors for high-dimensional time series, Biometrika 98 (2011) 901–918.

[14] C. Lévy-Leduc, H. Boistard, E. Moulines, M.S. Taqqu, V.A. Reisen, Asymptotic properties of U-processes under long-range dependence, The Annals of Statistics 39 (2011) 1399–1426.

[15] 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 (2011) 59–71.

[16] 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, Journal of Time Series Analysis 32 (2011) 135–156.

[17] O. Lordan, J.M. Sallan, N. Escorihuela, D. Gonzalez-Prieto, Robustness of airline route networks, Physica A: Statistical Mechanics and its Applications 445 (2016) 18 – 26.

[18] Y. Ma, M.G. Genton, Highly robust estimation of the autocovariance function, Journal of Time Series Analysis 21 (2000) 663–684.

[19] Y. Ma, M.G. Genton, Highly robust estimation of dispersion matrices, Journal of Multivariate Analysis 78 (2001) 11–36.

[20] R. Maynard, Key airborne pollutants: the impact on health, Science of The Total Environment 334-335 (2004) 9–13.

[21] F.F. Molinares, V.A. Reisen, F. Cribari-Neto, Robust estimation in long-memory processes under additive outliers, Journal of Statistical Planning and Inference 139 (2009) 2511–2525.

[22] D. Peña, G.E.P. Box, Identifying a simplifying structure in time series, Journal of the American Statistical Association 82 (1987) 836–843.

[23] M. Perc, Nonlinear time series analysis of the human electrocardiogram, European Journal of Physics 26 (2005) 757.

[24] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, Journal of the American Statistical association 88 (1993) 1273–1283.

[25] J.H. Seinfeld, S.N. Pandis, Atmospheric chemistry and physics: from air pollution to climate change, J. Wiley, New York, 2006.

[26] 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, Journal of the Royal Statistical Society Series C-Applied Statistics. 67 (2018) 453–480.

[27] J.H. Stock, M.W. Watson, Forecasting using principal components from a large number of predictors, Journal of the American Statistical Association 97 (2002) 1167–1179.

[28] R.S. Tsay, Outliers, level shifts, and variance changes in time series, Journal of Forecasting 7 (1988) 1–20.

[29] A.W. van der Vaart, Asymptotic statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.

[30] E. Vanhatalo, M. Kulahci, Impact of autocorrelation on principal components and their use in statistical process control, Quality and Reliability Engineering International 32 (2016) 1483–1500.

[31] 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 (2002) 1093–1136.

[32] WHO, Air quality guidelines: global update 2005, WHO - World Health Organization, 2006.

[33] WHO, Air pollution estimates, WHO - World Health Organization, 2014.

[34] B. Zamprogno, PCA applied in time series data with applications to air quality data, Ph.D. thesis, PPGEA - Universidade Federal do Espírito Santo, 2013. In press.

[35] 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.

List of Figures

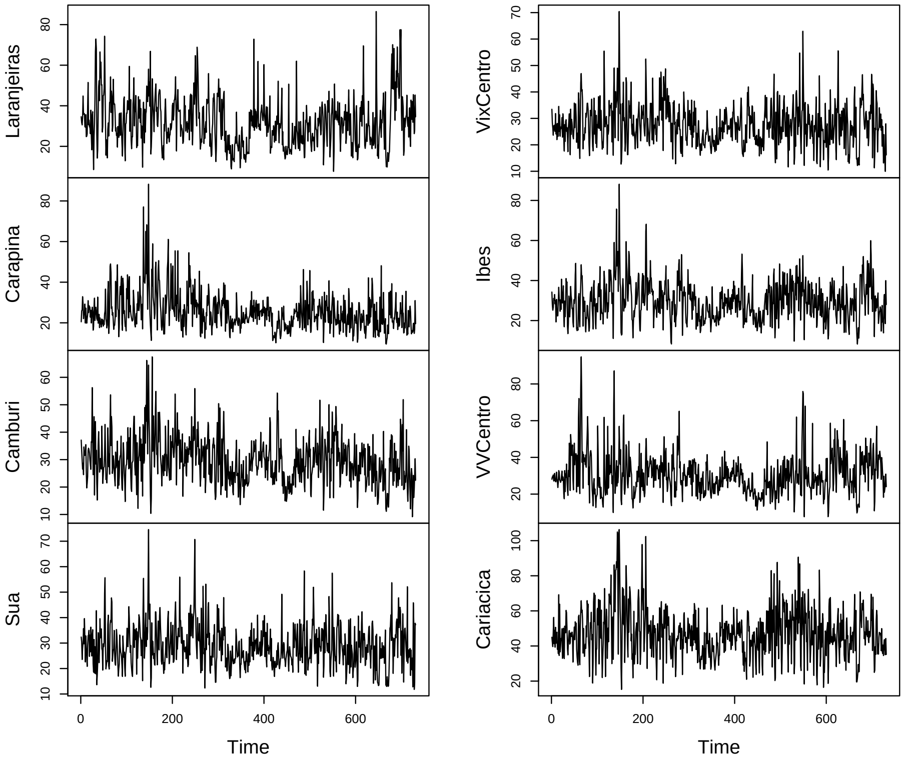


Figure 1: PM_{10} pollutant concentrations of the eight stations of AQAMN ($k = 8$).

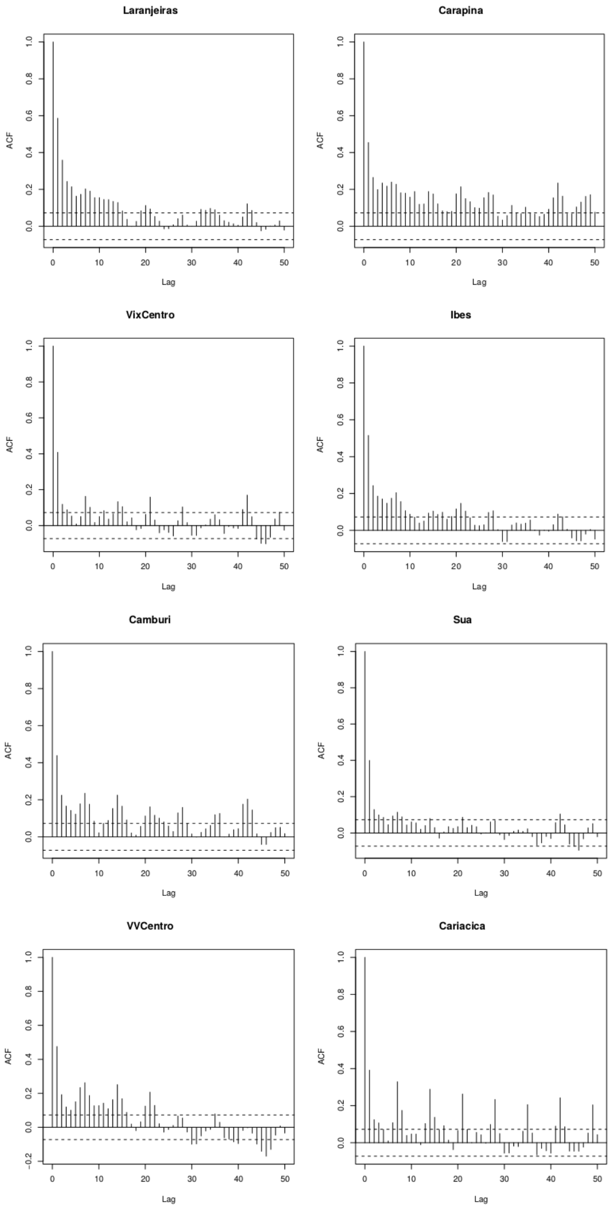


Figure 2: Classical ACF estimates of the PM_{10} pollutant concentrations.

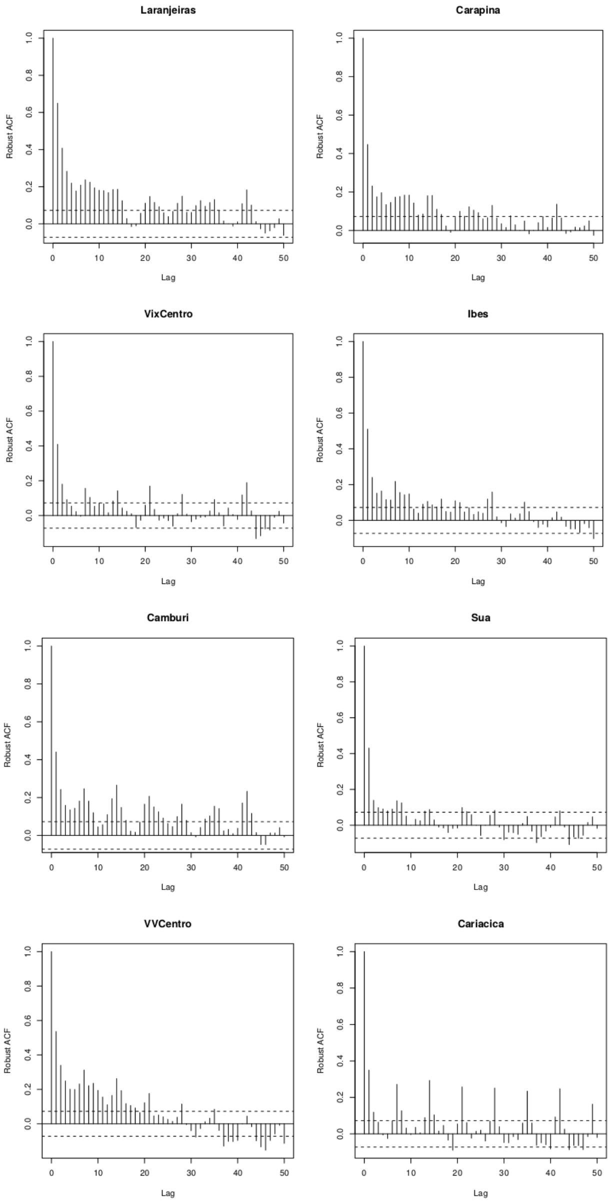
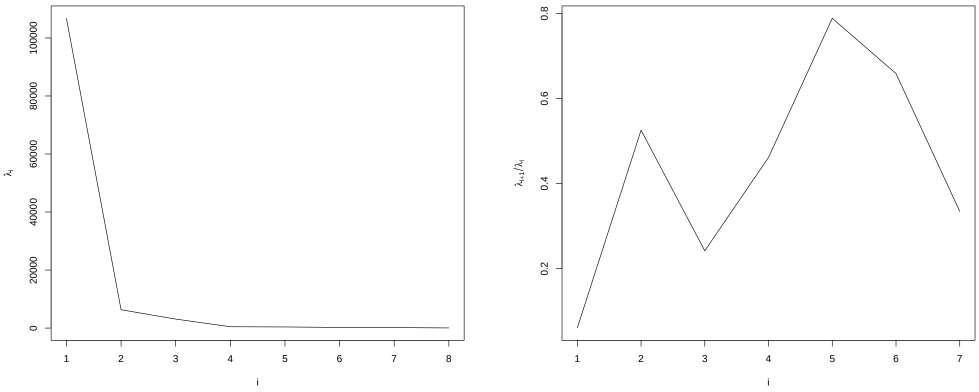
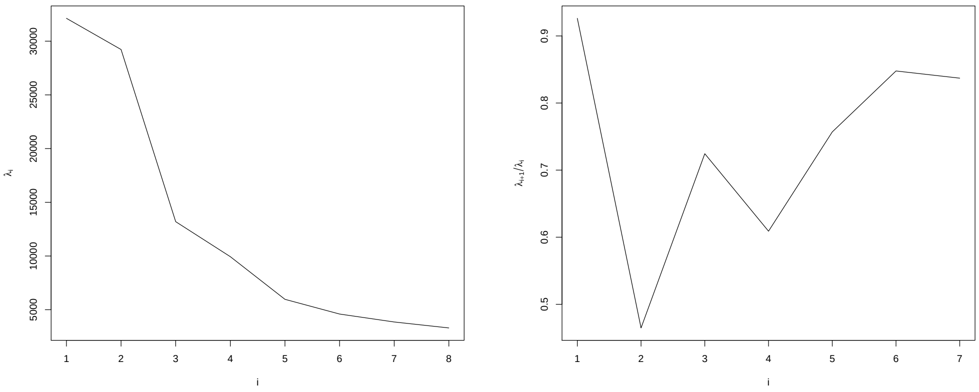


Figure 3: Robust ACF estimates of the PM_{10} pollutant concentrations.



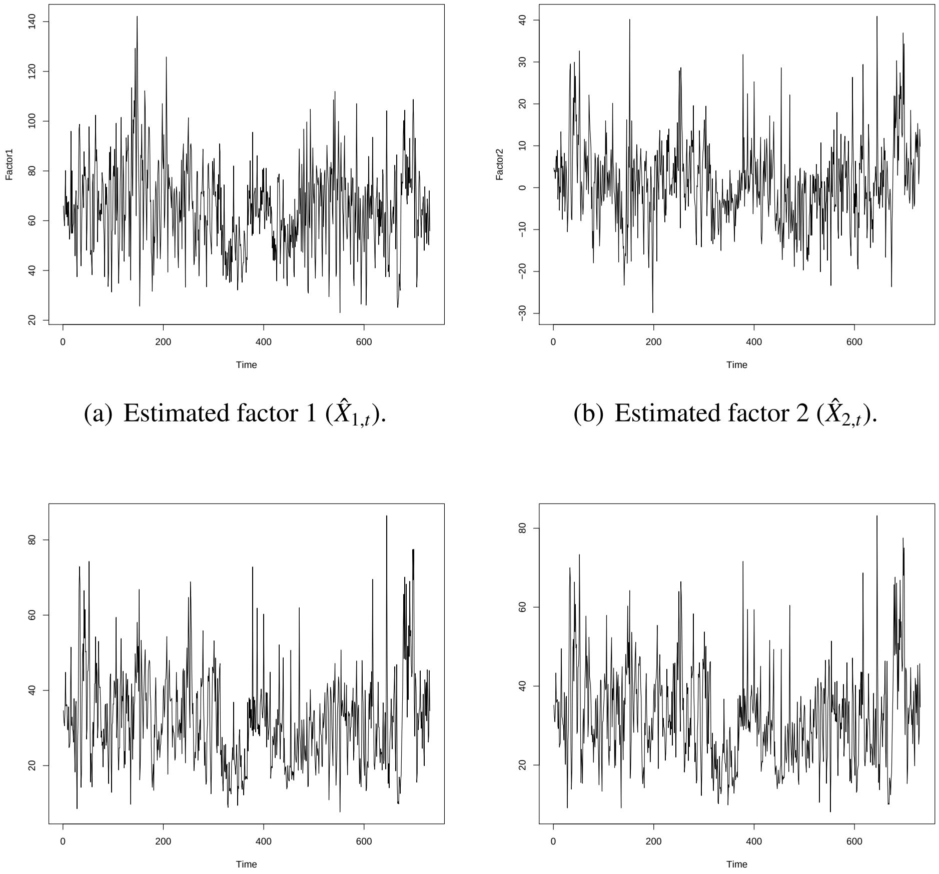
(a) (b)

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



(a) (b)

Figure 5: A scree plot (a) and the plot of the ratios (b) of the estimated eigenvalues of $\hat{\boldsymbol{M}}_Q$.



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

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

(c) Observed Laranjeiras series.

(d) Estimated Laranjeiras series.

Figure 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.

List of Tables

Table 1: $\mathbf{\Phi}_1$ coefficients matrix of the VAR(1) process.

$\mathbf{\Phi}_1$		
0.60	0.00	0.00
0.00	-0.50	0.00
0.00	0.00	0.30

Table 2: Relative frequency estimates for $P(\hat{r} = 3)$.

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	0.740	0.960	1	1	1

Table 3: Relative frequency estimates for $P(\hat{r_Q} = 3)$.

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 4: Relative frequency estimates for dimensional reduction, $n = 100$.

	$p = 0$			$p = 0.05$ 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	0.330	0.585	0.250	0.230	0.290	0.140	0.410	0.450	0.180	0.380	0.440
$k = 0.5n$	0.100	0.280	0.710	0.240	0.240	0.260	0.100	0.220	0.680	0.160	0.310	0.530
$k = 0.8n$	0.040	0.200	0.785	0.130	0.120	0.210	0.040	0.270	0.690	0.060	0.290	0.650