An alternative estimator for the number of factors for high-dimensional time series. A robust approach.

Valdério A. Reisen
DEST-CCE/PPGEA/PPGECON - Federal University of Espírito Santo, Brazil
valderioanselmoreisen@gmail.com

CentraleSupélec, Gif-Sur-Yvette
Summary

1. Introduction
2. Introduction
3. Time Domain: A robust estimator of the ACF
4. Factor Analysis-Methodology
5. Factor Analysis - Simulation cases
6. Application
7. Conclusions
Abstract

This paper considers the factor modeling for high-dimensional time series with short and long-memory properties and in the presence of additive outliers. The factor model studied by Lam and Yao (2012) is extended to the presence of additive outliers. The estimators of the number of factors are obtained by the robust covariance matrix. The methodology is analyzed in terms of the convergence rate of the number factors and by means of Monte Carlo simulations. Application with the aiming to reduce the dimensionality of the data set: The pollutant PM$_{10}$ in the Greater Vitória region (ES, Brazil).
Introduction: Main topics of this talk

- Factor Analysis (FA) for multivariate time series with long-memory and short-memory properties and outliers;
- A robust dimension reduction estimator for the number of components in FA is proposed (an extension Lam and Yao (2012));
- A simulation study to show the performance of the method for TS under additive outliers
- The application of the suggested estimator to a real data set. The pollutant PM$_{10}$ in the Greater Vitória region (ES, Brazil).
Figure 1: Geographical location of RAMQAr’s stations.
Figure 2: PM$_{10}$ concentrations of RAMQAr’s stations
There are some high levels of concentrations and long dependence among the air pollution data from AAQMN of Greater Vitória Region.

- High level of concentrations → **health impact**.
- High level of concentrations → **may be outliers**.
Lemma 1

(Fajardo et al. (2009)) Suppose that \( y_1, y_2, \ldots, y_n \) is a set of time series observations and let \( \hat{\rho}_y(h) = \hat{\gamma}_y(h)/\hat{\gamma}_y(0) \), then

i. For \( m = 1 \) (one outlier),

\[
\lim_{n \to \infty} \lim_{\omega \to \infty} \hat{\rho}_y(h) = 0.
\]

ii. For \( m = 2 \) and \( T_2 = T_1 + l \), such that \( h < T_1 < T_2 < n - h \), we have

\[
\lim_{n \to \infty} \left\{ \lim_{\omega_1 \to \infty} \lim_{\omega_2 \to \pm \infty} \hat{\rho}_y(h) \right\} = \begin{cases} 
0, & \text{if } h \neq l, \\
\pm 0.5, & \text{if } h = l.
\end{cases}
\]
Proposition 1

(Cotta et al. (2017)) Suppose that $Z_{1,t}, Z_{2,t}, \ldots, Z_{n,t}$ is a set of $k$-dimensional time series observations of Model 15 and $m$ is the expected number of additive outliers as stated in (15). Let

$$\hat{\rho}_{ij}^{Z}(h) = \frac{\gamma_{ij}^{Z}(h)}{\sqrt{\gamma_{ii}^{Z}(0)}\sqrt{\gamma_{jj}^{Z}(0)}}, \quad \text{for } \forall i, j = 1, \ldots, k,$$

Then

a. For $m = 1$ (one outlier occurring only at $Z_{i,t}$),

$$\lim_{n \to \infty} \text{plim} \omega_{i} \to \infty \hat{\rho}_{ij}^{Z}(h) = 0.$$
Proposition 1 (cont.)

\[ b. \quad \text{For } m = 2 \text{ (two outliers occurring at } Z_{i,t} \text{ or/and at } Z_{j,t} \text{) and assuming that } \hat{\gamma}^Z_{ij}(h) \neq 0, \text{ for } Z_{i,t} \text{ and } Z_{j,t}, \text{ it follows} \]

\[
\lim_{n \to \infty} \text{plim} \quad \hat{\rho}^Z_{ij}(h) = 0.
\]

In (a.) and (b.), \( w_i \) and \( w_j \) are the magnitudes of the additive outliers occurring at position \( i \) and \( j \).
Rousseeuw & Croux (1993) proposed a robust scale estimator function which is based on the $k$th order statistic of $\binom{n}{2}$ distances \{$|y_i - y_j|, i < j\}$, and can be written as

$$Q_n(y) = c \times \{|y_i - y_j|; i < j\}(k),$$

where $y = (y_1, y_2, \ldots, y_n)'$, $c$ is a constant used to guarantee consistency ($c = 2.2191$ for the normal distribution), and $k = \left\lfloor \frac{\binom{n}{2} + 2}{4} \right\rfloor + 1$. Rousseeuw & Croux (1993) showed that the asymptotic breakdown point of $Q_n(\cdot)$ is 50%, which means that the time series can be contaminated by up to half of the observations with outliers and $Q_n(\cdot)$ will still yield sensible estimates.
Robust Sample autocovariance

Ma & Genton (2000) suggested the following robust sample autocovariance function

\[
\hat{\gamma}_Q(h) = \frac{1}{4} \left[ Q_{n-h}^2(u + v) - Q_{n-h}^2(u - v) \right],
\]  

(2)

where \( u \) and \( v \) are vectors containing the initial \( n - h \) and the final \( n - h \) observations, respectively. Note that, the above ACF is not necessarily positive semi-definite.
The references below are background works which present theoretical and applied results related to the robust ACF $\hat{\gamma}_Q(h)$


Factor Analysis - Model

Let $Z_t, t \in \mathbb{Z}$, be a $k$-dimensional zero-mean vector of an observed time series. Also, let $x_t$ be an unobserved $r$-dimensional vector of common factors. It is assumed that these series are generated by $r$ ($r \leq k$) factors, $x_t$, plus a measurement error $\epsilon_t$ as

$$z_t = Px_t + \epsilon_t,$$

where $P$ is a $k \times r$ matrix of parameters of rank $r$ (factor loading matrix), and $\epsilon_t$ is a $k$-dimensional white-noise sequence with full-rank covariance matrix $\Sigma_{\epsilon}$. Thus, all the common dynamic structure comes through the common factors, $x_t$. Assumes that $P'P = I$. When $r$ is much smaller than $k$ an effective dimension-reduction is achieved. $Z_t$ is driven by a much lower-dimension process $x_t$. 
The key to the inference for the model in Equation (3) is to determine the number of factors $r$ and to estimate the $k \times r$ factor loading matrix $P$, or more precisely the factor loading space $\Omega(P)$. Once an estimator is obtained, say, $\hat{P}$, a natural estimator for the factor process is

$$\hat{x}_t = \hat{P}'z_t,$$

(4)

and the resulting residuals are

$$\hat{\epsilon}_t = (I_d - \hat{P}\hat{P}')z_t.$$

(5)

The estimation of $P$ is suggested by Lam & Yao (2012).
Suppose that the vector of common factors $x_t = (x_1,t, \ldots, x_r,t)'$, $t \in \mathbb{Z}$, follows a zero-mean $r$-dimensional stationary vector Fractional Autoregressive Moving Average process (VARFIMA($p_x, d_x, q_x$)) given by

$$
\phi_x(B)D_x[(1 - B)^d]x_t = \theta_x(B)a_t,
$$

(6)

where

$$
\phi_x(B) = I - \phi_1 B - \cdots - \phi_p B^p
$$

$$
\theta_x(B) = I + \theta_1 B + \cdots + \theta_q B^q
$$

are matrix polynomials in the backshift operator $B$, the $\phi'$s and the $\theta'$s are $r \times r$ matrices, the roots of the determinant polynomial $|\phi_x(B)|$ are all outside the unit circle, and those of $|\theta_x(B)|$ are all outside the unit circle. In addition, $a_t$ is a sequence of $r \times r$ vector Gaussian White noise with zero mean and covariance matrix $\Sigma_a$. 
Remark 1

The VARFIMA process $x_t$, $t \in \mathbb{Z}$, can be written as an infinite stationary second-order moving average representation as follows:

$$x_t = \sum_{j=0}^{\infty} \Psi_j a_{t-j},$$

where the innovations $a_t = [a_{1,t}, \ldots, a_{r,t}]'$ are $r$-dimensional martingale differences with respect to an increasing sequence of $\sigma$-fields $\{F_t\}$ such that for some $\lambda > 0$, $\sup_t E(|a_{i,t}|^{2+\lambda} |F_{t-1}) < \infty$, a.s., for all $i = 1, \ldots, r$. Let $E(a_t a_t'|F_{t-1}) = \Sigma_a$, a.s. The $r \times r$ matrix coefficients $\Psi_j$ are often referred to as impulse responses. The main characterization of the process $x_t$ considered in this paper is that impulses responses $\Psi_j$ converge at slow hyperbolic rates as $j \to \infty$. More precisely, there are $r$ memory parameters $d_1, d_2, \ldots, d_r$, whose values lie in $(0,0.5)$ such that the impulse responses $\Psi_j$ can be approximated by...
Remark 1 (cont.)

\[ \Psi_j \sim D \left[ \frac{1}{\Gamma(d)} j^{d-1} \right] \Pi, \quad \text{as } j \to \infty, \tag{8} \]

where \( \Gamma(\cdot) \) is a gamma function and \( \Pi \) is a nonsingular \( r \times r \) matrix of constants that are independent of \( j \) and may be functions of a smaller set of unknown parameters. The notation \( D[j^{d-1}/\Gamma(j)] \) represents a \( r \times r \) diagonal matrix with \( j^{d_1-1}/\Gamma(d_1), \ldots, j^{d_r-1}/\Gamma(d_r) \) on the diagonal. In fact, for any univariate function \( f \) of a single variable, the notation \( D[f(d)] \) represents \( r \times r \) diagonal matrix with \( f(d_1), \ldots, f(d_r) \) on the diagonal. Also, the notation \( \sim \) is defined as follows: given two sequences of matrices \( U_j \) and \( V_j \), as \( j \to \infty \), for \( i \) and \( r \), where \( u_{i,r,j} \) and \( v_{i,r,j} \) are the \((i,r)\)th elements of \( U_j \) and of \( V_j \), respectively. Let \( \psi'_{i,j} \) and \( \pi_{i} \) be the \( i \)th rows of \( \Psi_j \) and \( \Pi \), respectively; then Equation (8) implies that \( \psi'_{i,j} \sim j^{d_{i}-1}\Gamma(d_{i})^{-1}\pi_{i}', \) as \( j \to \infty \), for all \( i = 1, \ldots, r \). Note that the conditions on the memory parameters \( d_{i} \in (0,0.5) \), for \( i = 1, \ldots, r \), ensure that the impulse responses are square-summable and the infinite sum in Equation (7) exists.
Remark 2

The autocovariances $\Gamma(j) \equiv \text{Cov}(x_t, x_{t+j})$ of the process $x_t$ must also converge at hyperbolic rates as follows (Chung (2002)).

\begin{equation}
\Gamma(j) \sim D(j^{d-0.5}) \cdot A \cdot D(j^{d-0.5}), \quad \text{as} \quad j \to \infty,
\end{equation}

where the $(i, r)$th element of the $r \times r$ matrix $A$ is $\text{Gamma}(1 - d_i - d_r)/[\Gamma(d_r) \times \Gamma(1 - d_r)]. \pi_i^t \Sigma_a \pi_r.$
Remark 2 (Cont.)

Hence, as $j \to \infty$, not only do the autocovariances $\gamma_{i,i}(j)$ of each $x_{i,t}$ die out slowly at a hyperbolic rate, i.e.,

$$\gamma_{i,i}(j) \sim j^{2d_i-1} \frac{\Gamma(1 - 2d_i)}{[\Gamma(d_i) \Gamma(1 - d_i)]} \pi_i' \Sigma a \pi_i,$$

the covariances $\gamma_{i,r}(j)$ between the current $x_{i,t}$ and the future $x_{r,t+j}$, for $i \neq r$, also vanish at hyperbolic rates, i.e.,

$$\gamma_{i,r}(j) \sim j^{d_i+d_r-1} \frac{\Gamma(1 - d_i - d_r)}{[\Gamma(d_r) \times \Gamma(1 - d_r)]} \pi_i' \Sigma a \pi_r.$$

Hosking (1996) presents the result for the univariate case.
Lemma 2

Let $z_t = Px_t + \epsilon_t$ as defined in Equations (3) and (6) and assume that $\Gamma_z(h) = E[z_{t-h}z_t']$ are the covariance matrices of the process $z_t$ and $\Gamma_x(h) = E[x_{t-h}x_t']$ are the covariance matrices for the generating vector $x_t$. Then

$$\Gamma_z(0) = P \Gamma_x(0) P' + \Sigma_\epsilon,$$

(10)

$$\Gamma_z(h) = P \Gamma_x(h) P', \quad h \geq 1,$$

(11)

where $\text{rank}(\Gamma_z(h)) = \text{rank}(\Gamma_x(h))$, as $h \geq 1$. 
Lemma 3

If the factors are independent for all lags and the matrix $\Sigma_\epsilon$ is diagonal, then

a) All of the covariance matrices $\Gamma_x(h)$ are diagonal;

b) The matrices $\Gamma_z(h)$ are symmetric for $h \geq 1$;

c) By Spectral Decomposition, the columns of $P$ will be eigenvectors of $\Gamma_z(h)$ with eigenvalues $\gamma_i(h)$, where $\gamma_i(h)$ are the diagonal elements of $\Gamma_x(h)$. 
Proposition 2

Suppose \( z_t = P x_t + \epsilon_t \), where \( x_t \) is a \( r \)-dimensional VARFIMA\((p_x, d_x, q_x)\) process, \( P \) is a \( k \times r \) matrix \((k \geq r)\) of rank \( r \), and \( \epsilon_t \) is a \( k \)-dimensional white noise sequence with covariance \( \Sigma_\epsilon \). Then \( z_t \) follows a \( k \)-dimensional VARFIMA\((p_z, d_z, q_z)\) with \( p_z = p_x \), \( d_z = d_x \) and \( q_z = \max(p_x, q_x) \).
Remark 3

In Equation (6), consider \( y_t = D_x [(1 - B)^d] x_t \). Then, if the parameters of long memory (\( d_i \)) are all equal to zero, the model presented in Equation 6 becomes the short memory model described by Peña & Box (1987), i.e., a VARMA(\( p, q \)). Thus, \( \phi_x(B) D_x [(1 - B)^d] x_t = \theta_x(B) a_t \) becomes \( \phi_y(B) y_t = \theta_y(B) a_t \).
Following the same line as Lam & Yao (2012), the estimation of $P$ if performed by an eigenanalysis on

$$\hat{M} = \sum_{h=1}^{h_0} \hat{\Gamma}_z(h)\hat{\Gamma}_z(h)' ,$$  \hspace{1cm} (12)

where $h_0$ is a prescribed integer and $\hat{\Gamma}_z(h)$ denotes the sample covariance matrix of $z_t$ at lag $h$.

**ratio-based estimator for $r$.** The estimator for the number of factors $r$ is given by:

$$\hat{r} = \arg\min_{1 \leq i \leq R} \frac{\hat{\lambda}_{i+1}}{\hat{\lambda}_i} ,$$  \hspace{1cm} (13)

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_k$ are the eigenvalues of $\hat{M}$, and $r < R < k$ is a constant. As suggested in Lam and Yao (2012), in practice, $R = p/2$ can be the starting point.
Based on Equation 12 and on the robust ACF estimator, the robust M estimator is here suggested as

$$\hat{M}_{Q_n} = \sum_{h=1}^{h_0} \hat{\Gamma}_{z,Q_n}(h)\hat{\Gamma}_{z,Q_n}(h)'.$$  \hspace{1cm} (14)

where $\hat{\Gamma}_{z,Q_n}(h)$ denotes the sample robust covariance matrix of $z_t$ at lag $h$.

Therefore, the estimator $\hat{r}_{Q_n}$ for the number of factors is similarly obtained from Equation (13). Note that $\hat{M}_{Q_n}$ is positive semi-definite.
Proposition 3

Let $h$ be a fixed positive integer and $(\hat{\Gamma}_Q(h))_{1 \leq i, j \leq p} = (\hat{\gamma}_{i,j}^Q(h))_{1 \leq i, j \leq p}$, where $\hat{\gamma}_{i,j}^Q(h)$ is Robust ACF function defined previously. Assume that the process is short-memory, then

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

where $(\hat{\lambda}_j^Q)_{1 \leq j \leq p}$ and $(\lambda_j)_{1 \leq j \leq p}$ denote the eigenvalues of $(\sum_{h=1}^{h_0} \hat{\Gamma}_Q(h)\hat{\Gamma}_Q(h)')$ and $(\sum_{h=1}^{h_0} \Gamma(h)\Gamma(h)')$, respectively, where $(\Gamma(h))_{1 \leq i, j \leq p} = (\gamma_{i,j}(h))_{1 \leq i, j \leq p}$ and $h_0$ is a fixed integer larger than 1.
Theoretical results

**Proposition 4**

Let $h$ be a fixed positive integer and 
$$ \left( \hat{\Gamma}_Q(h) \right)_{1 \leq i, j \leq p} = \left( \hat{\gamma}_{i,j}^Q(h) \right)_{1 \leq i, j \leq p}, $$

where $\hat{\gamma}_{i,j}^Q(h)$ is defined previously. Assume that the vector process has long-memory property, then

(i) If, for all $i$ in $\{1, \ldots, k\}$, $D_i > 1/2(d_i < 1/4)$,

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

(ii) If, there exists $i_0$ in $\{1, \ldots, k\}$ such that $D_{i_0} < 1/2(d_i > 1/4)$,

$$ n^{D_{i_0}} \sup_{1 \leq j \leq p} \left| \hat{\lambda}_j^Q - \lambda_j \right| = O_p(1), \text{ as } n \to \infty, $$
(background):

- COTTA, H., REISEN, V.A., BONDON, P., STUMMER, W. Robust principal component analysis with air pollution data: . 2017. EUSIPCO.
- Reisen, V. A, Lévy-Leduc, C., Zambon, E. Robust Factor Model for long-memory multivariate time series with application to stock market. IN REVISION.
Monte Carlo experiments were conducted to analyze the effect of high-dimensional time series with additive outliers on the factor modeling and time series with long and short-memory dependency.

The empirical study is divided into two cases of $x_t$ (Equation 6), which follows a VARFIMA model with $r = 3$: (1) short-memory process that is, $x_t$ is a VARMA process; (2) long-memory process; $d = (d_1, d_2, d_3)'$, for at least one $d_1, d_2, d_3 \in (0, 0.5)$.

The VARFIMA model was generated with independent $a_t$ from $N(0, I)$ and $\Phi$ coefficients, which are displayed in Table 1. The sample size is $n = 50, 100, 200, 400, 800$ and $1600$, and $k = 0.2n, 0.5n, 0.8n$. Here, an only one case is presented.

All $k \times r$ elements of matrix $P$ were generated as independent observations from the uniform distribution on the interval $[-1,1]$ (simulation method was according to Lam and Yao (2012)).
The empirical study is divided into two cases of $x_t$ (Equation 6), which follows a VARFIMA model with $r = 3$: (1) short-memory process where $d = (0, 0, 0)'$; that is, $x_t$ is a VARMA process; (2) long-memory process where $d = (d_1, d_2, d_3)'$, for at least one $d_1, d_2, d_3 \in (0, 0.5)$. The VARFIMA model was generated with independent $a_t$ from $N(0, I)$ and $\Phi$ coefficients, which are displayed in Table 1. The sample size is $n = 50, 100, 200, 400, 800$ and $1600$, and $k = 0.2n, 0.5n, 0.8n$ and $1.2n$.

<table>
<thead>
<tr>
<th>$\Phi_1$ (Model 1)</th>
<th>$\Phi_1$ (Model 2)</th>
<th>$\Phi_1$ (Model 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Let \( \{z_t\}, \ t = 1, ..., \ t \in \mathbb{Z} \), be a vector process contaminated by additive outliers defined as follows:

\[
z_t = P x_t + \omega \circ \delta_t,
\]

(15)

where "\circ" is the Hadamard product. \( \omega = [\omega_1, ..., \omega_k]' \) is a magnitude vector of additive outliers. \( \delta_t = [\delta_{1t}, ..., \delta_{kt}]' \) is a random vector indicating the occurrence of an outlier at time \( t \), in variable \( k \), such as \( \mathbb{P}(\delta_{k,t} = -1) = \mathbb{P}(\delta_{k,t} = 1) = p/2 \) and \( \mathbb{P}(\delta_{k,t} = 0) = 1 - p \), where \( \mathbb{E}[\delta_{k,t}] = 0 \) and \( \mathbb{E}[\delta_{k,t}^2] = \text{Var}(\delta_{k,t}) = p \). The model described above assumes that \( \{Z_t\} \) and \( \{\delta_t\} \) are independent processes. Also, it is assumed that the elements of \( \delta_t \) are not correlated and temporally uncorrelated, i.e, \( \mathbb{E}(\delta_t \delta_t') = \Sigma_\delta = \text{diag}(p, ..., p) \) and \( \mathbb{E}(\delta_t \delta_{t+h}') = 0 \) for \( h \neq 0 \).
### Table 2: Relative frequency estimates for $f_{rel}(r = 3)$ in the simulation with 200 replications - Model 1. (Classical ACF): similar results of Lam and Yao (2012)

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>1600</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 0.2n$</td>
<td></td>
<td>0.170</td>
<td>0.585</td>
<td>0.870</td>
<td>0.995</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 0.5n$</td>
<td></td>
<td>0.395</td>
<td>0.710</td>
<td>0.975</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 0.8n$</td>
<td></td>
<td>0.435</td>
<td>0.740</td>
<td>0.960</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 3: Relative frequency estimates for dimensional reduction - $d = [0.1, 0.2, 0.4]'$ and Model 1’s $\Phi$ coefficients, r=3.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\Gamma}_z$</th>
<th>$\hat{\Gamma}_z$</th>
<th>$\hat{\Gamma}_{z,Q_n}$</th>
<th>$\hat{\Gamma}_{z,Q_n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 0$</td>
<td>$p = 0.05$</td>
<td>$p = 0$</td>
<td>$p = 0.05$</td>
</tr>
<tr>
<td></td>
<td>$n = 100$</td>
<td>$n = 100$</td>
<td>$n = 100$</td>
<td>$n = 100$</td>
</tr>
<tr>
<td>$r = 1$</td>
<td>0.110</td>
<td>0.310</td>
<td>0.150</td>
<td>0.160</td>
</tr>
<tr>
<td>$r = 2$</td>
<td>0.260</td>
<td>0.270</td>
<td>0.310</td>
<td>0.320</td>
</tr>
<tr>
<td>$r = 3$</td>
<td>0.630</td>
<td>0.260</td>
<td>0.540</td>
<td>0.710</td>
</tr>
</tbody>
</table>

- The third column gives the simulation results using $\hat{\Gamma}_{z,Q_n}$ when $p = 0$. As one can see, the $\hat{\Gamma}$ estimates using $\hat{\Gamma}_{z,Q_n}$ present similar results of $\hat{\Gamma}_z$ when $p = 0$. Both methods have the same asymptotical properties.

- The presence of atypical observations in the data leads to a reduction of the estimated frequencies when $\hat{\Gamma} = 3$ for all values of $k$ in the classical method. This does not occur when the robust estimator is utilized, the results are quite close to the ones from the first column.
The application of the proposed methodology consists of applying the PCA to cluster stations with the same behavior of measured pollutants. This section presents an application of the methodology for $PM_{10}$ concentrations measured at the Air Quality Automatic Monitoring Network (AQAMN) of the Greater Vitória Region (GVR). The application was divided into two parts: 1) reduction of the dimensions, and 2) forecasting. The data are:

- $PM_{10}$ measured in $\mu g/m^3$;
- Daily average;
- The period: January 2005 to December 2009;
- All eight RAMQAr’s stations.
The RAMQAr - Time series

Figure 3: $\text{PM}_{10}$’s concentrations of RAMQAr’s stations
The RAMQAr - Descriptive statistics

Figure 4: Boxplot of PM\(_{10}\)’s of RAMQAr’s stations.
The RAMQAr - ACFs

Figure 5: $\widehat{PM}_{10} \hat{\rho}_{Z_n}(h)$ function for RAMQAr’s Stations
The RAMQAr - The Robust ACF X ACF- IBIS
The RAMQAr - FA’s Classical ACF

Figure 6: Plots of estimated eigenvalues (a) and ratios of estimated eigenvalues of $\hat{M}$ (b). The test indicated 1 factor.
The RAMQAr - FA’s ROBUST

Figure 7: Plots of estimated eigenvalues (a) and the ratios of estimated eigenvalues of $\hat{M}_Q$ (b). The Robust indicated 2 factors.
The AF analysis suggests the following model for $Z_t$ (the daily PM$_{10}$ concentrations of the stations):

$$Z_t = p_1 w_t + p_2 v_t + \varepsilon_t,$$

where $w_t$ denotes the first factor, $v_t$ is the second factor of solution $\hat{Z} = \hat{P}\hat{X}$, and $\varepsilon_t$ is a vector white-noise process.
Main Conclusions

- A robust factor model for high-dimensional time series with short and long-memory and additive outliers is proposed.
- This study considered the effects of different correlation structures and additive outliers on a vector linear process and its implication in the analysis and interpretation of Factor Analysis calculated from the correlation matrix of this process;
- It was shown that the existence of outliers destroys the correlation and cross-correlation of a vector time series;
- This article applied the proposed methodology to identify pollution behavior for the pollutant $PM_{10}$ in the Greater Region of Vitória to enable better management of the local monitoring network.
- The results in this paper will hopefully stimulate further research on using robust estimation methods and long-memory models to represent and forecast environmental time series.
Thanks!!!
valderioanselmoreisen@gmail.com


**URL**: http://dx.doi.org/10.1214/12-AOS970


**URL**: http://www.jstor.org/stable/2288794