UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL

HIGOR HENRIQUE ARANDA COTTA

ROBUST METHODS IN MULTIVARIATE TIME SERIES

VITÓRIA
2019

HIGOR HENRIQUE ARANDA COTTA

ROBUST METHODS IN MULTIVARIATE TIME SERIES

Thesis submitted to Programa de Pós-graduação em Engenharia Ambiental of Centro Tecnológico da Universidade Federal do Espírito Santo, for the degree of Doctor of Science in Environmental Engineering.
Supervisors: Prof. Valdério Anselmo Reisen, PhD. Prof. Pascal Bondon, PhD.

VITÓRIA
2019

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Aranda Cotta, Higor Henrique, 1987-
A662r Robust Methods in Multivariate Time Series / Higor Henrique Aranda Cotta. - 2019.
154 f. : il.

Orientador: Valdério Anselmo Reisen.
Coorientador: Pascal Bondon.
Tese (Doutorado em Engenharia Ambiental) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Séries temporais multivariadas. 2. Poluição do ar. 3. Função de autocovariância e autocorrelação. 4. Outliers. 5. Robustez. 6. Análise fatorial. I. Reisen, Valdério Anselmo. II. Bondon, Pascal. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 628

NNT : 2019SACLC064

Thèse de doctorat

Robust Methods in Multivariate Time Series

Thèse de doctorat de Universidade Federal do Espírito Santo et de l'Université Paris-Saclay, préparée à UFES et CentraleSupélec

Ecole doctorale n°9 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Traitement du signal et des images

Thèse présentée et soutenue à Vitória, Espírito Santo, Brésil, le 22 août 2019, par

M. HIGOR HENRIQUE ARANDA COTTA

Composition du Jury :

Mme Taciana Toledo de Almeida Albuquerque Professeure, Universidade Federal de Minas Gerais	Présidente
Mme Glaura da Conceição Franco Professeure, Universidade Federal de Minas Gerais	Rapporteuse
M. Wilfredo Omar Palma Manríquez Professeur, Pontificia Universidad Catolica de Chile	Rapporteur
M. Alexandre Renaux Maître de Conférences, Paris-Saclay	Examinateur
M. Neyval Costa Reis Junior Professeur, Universidade Federal do Espírito Santo	Examinateur
M. Márton Ispány Professeur, University of Debrecen	Examinateur
M. Valdério Anselmo Reisen Professeur, Universidade Federal do Espírito Santo	Directeur de thèse
M. Pascal Bondon Directeur de Recherche, CNRS	Directeur de thèse

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Méthodes robustes dans le séries chronologiques multivariées

Mots clés : Séries chronologiques multivariées, robustesse, valeurs aberrantes, domaine temporel, domaine fréquentiel.

Résumé :
Ce manuscrit propose de nouvelles méthodes d'estimation robustes pour les fonctions matricielles d'autocovariance et d'autocorrélation de séries chronologiques multivariées stationnaires pouvant présenter des valeurs aberrantes aléatoires additives. Ces fonctions jouent un rôle important dans l'identification et l'estimation des paramètres de modèles de séries chronologiques multivariées stationnaires. Les données aberrantes aléatoires peuvent impacter le niveau d'une ou plusieurs composantes du vecteur multivarié. Ceci augmente la variabilité globale de la série, ce qui a un impact sur le périodogramme matriciel et conduit à une diminution des valeurs de la fonction matricielle d'autocorrélation. Nous proposons tout d'abord de nouveaux estimateurs des fonctions matricielles d'autocovariance et d'autocorrélation construits en utilisant une approche spectrale à l'aide du périodogramme matriciel qui est l'estimateur naturel de la densité spectrale matricielle. Comme dans le cas des estimateurs classiques des fonctions d'autocovariance et d'autocorrélation matricielles, ces estimateurs sont affectés par des observations aberrantes. Ainsi, toute procédure d'identification ou d'estimation les utilisant est directement affectée, ce qui entraîne des conclusions erronées. Pour atténuer ce problème, nous proposons l'utilisation de techniques statistiques robustes pour créer des estimateurs résistants aux observations aléatoires aberrantes.
Dans un premier temps, nous proposons de nouveaux estimateurs des fonctions d'autocorvariance et d'autocorrélation de séries chronologiques univariées. Les domaines temporel et fréquentiel sont liés par la relation existant entre la fonction d'autocovariance et la densité spectrale. Le périodogramme étant sensible aux données aberrantes, nous obtenons un estimateur robuste en le remplaçant par le M-périodogramme. Le M-périodogramme est obtenu en remplaçant dans le calcul des estimations des coefficients de Fourier liés au périodogramme, la régression standard des moindres carrés par la M-régression robuste. Les propriétés asymptotiques des estimateurs sont établies. Leurs performances sont étudiées au moyen de simulations numériques pour différentes tailles d'échantillons et différents scénarios de contamination. Les résultats empiriques indiquent que les méthodes proposées fournissent des valeurs proches de celles obtenues par la fonction d'autocorrélation classique quand les données ne sont pas contaminées et resistent à différent scénarios de contamination. Ainsi, les estimateurs proposés dans cette thèse sont des méthodes alternatives utilisables pour des séries chronologiques présentant ou non des valeurs aberrantes.
Les estimateurs obtenus pour des séries chronologiques univariées sont ensuite étendus au cas de séries multivariées. Cette extension est simplifiée par le fait que le calcul du périodogramme croisé ne fait intervenir que les coefficients de Fourier de chaque composante de la série. De nouveau, la relation de dualité entre les domaines temporel et fréquentiel est utilisée via le lien entre la fonction matricielle d'autocovariance et la densité spectrale matricielle d'une série chronologique multivariée stationnaire. Le M-périodogramme matriciel apparaît alors comme une alternative robuste au périodogramme matriciel pour construire des estimateurs robustes des fonctions matricielles d'autocovariance et d'autocorrélation. Les propriétés asymptotiques sont étudiées et des expériences numériques sont réalisées. Comme exemple d'application avec des données réelles, nous utilisons les fonctions proposées pour ajuster un modèle autoregressif par la méthode de Yule-Walker à des données de pollution collectées dans la région de Vitória au Brésil (particules de diamètre inférieur à 10 micromètres, PM_{10}).
Enfin, l'estimation robuste du nombre de facteurs dans les modèles factoriels de grande dimension est considérée afin de réduire la dimensionnalité. Il est bien connu que les valeurs aberrantes aléatoires affectent les matrices de covariance et de corrélation et que les techniques qui dépendent du calcul de leurs valeurs propres et vecteurs propres, telles que l'analyse en composantes principales et l'analyse factorielle, sont affectées. Ainsi, en présence de valeurs aberrantes, les critères d'information proposés par Bai & Ng (2002) tendent à surestimer le nombre de facteurs. Pour atténuer ce problème, nous proposons de remplacer la matrice de covariance standard par la matrice de covariance robuste proposée dans ce manuscrit. Nos simulations de Monte Carlo montrent qu'en l'absence de contamination, les méthodes standards et robustes sont équivalentes. En présence d'observations aberrantes, le nombre de facteurs estimés augmente avec les méthodes non robustes alors qu'il reste le même en utilisant les méthodes robustes. À titre d'application avec des données réelles, nous étudions des concentrations de polluant PM_{10} mesurées dans la région de l'Île-de-France en France.

Title : Robust methods in multivariate time series

Keywords : Multivariate time series, robustness, outliers, time domain, frequency domain.

Abstract :
This manuscript proposes new robust estimation methods for the autocovariance and autocorrelation matrices functions of stationary multivariates time series that may have random additives outliers. These functions play an important role in the identification and estimation of time series model parameters. Random additive outliers can impact the level of one or more components of the multivariate vector. This increases the overall variability of the series, which has an impact on the periodogram matrix and leads to a decrease in the values of the autocorrelation matrix function. We first propose new estimators of the autocovariance and of autocorrelation matrices functions constructed using a spectral approach considering the periodogram matrix periodogram which is the natural estimator of the spectral density matrix. As in the case of the classic autocovariance and autocorrelation matrices functions estimators, these estimators are affected by aberrant observations. Thus, any identification or estimation procedure using them is directly affected, which leads to erroneous conclusions. To mitigate this problem, we propose the use of robust statistical techniques to create estimators resistant to aberrant random observations.
As a first step, we propose new estimators of autocovariance and autocorrelation functions of univariate time series. The time and frequency domains are linked by the relationship between the autocovariance function and the spectral density. As the periodogram is sensitive to aberrant data, we get a robust estimator by replacing it with the M-periodogram. The M-periodogram is obtained by replacing the Fourier coefficients related to periodogram calculated by the standard least squares regression with the ones calculated by the M-robust regression. The asymptotic properties of estimators are established. Their performances are studied by means of numerical simulations for different sample sizes and different scenarios of contamination. The empirical results indicate that the proposed methods provide close values of those obtained by the classical autocorrelation function when the data is not contaminated and it is resistant to different contamination scenarios. Thus, the estimators proposed in this thesis are alternative methods that can be used for time series with or without outliers.
The estimators obtained for univariate time series are then extended to the case of multivariate series. This extension is simplified by the fact that the calculation of the cross-periodogram only involves the Fourier coefficients of each component from the univariate series. Again, the duality relationship between time and frequency domains is considered via the link between the autocovariance matrix function and the spectral density matrix stationary multivariate time series. The M-periodogram matrix is a robust periodogram matrix alternative to build robust estimators of the autocovariance and autocorrelation matrices functions. The asymptotic properties are studied and numerical experiments are performed. As an example of an application with real data, we use the proposed functions to adjust an autoregressive model by the Yule-Walker method to Pollution data collected in the Vitória region Brazil (particles smaller than 10 micrometers in diameter, PM_{10}).
Finally, the robust estimation of the number of factors in large factorial models is considered in order to reduce the dimensionality. It is well known that the values random additive outliers affect the covariance and correlation matrices and the techniques that depend on the calculation of their eigenvalues and eigenvectors, such as the analysis principal components and the factor analysis, are affected. Thus, in the presence of outliers, the information criteria proposed by Bai & Ng (2002) tend to overestimate the number of factors. To alleviate this problem, we proposeto replace the standard covariance matrix with the robust covariance matrix proposed in this manuscript. Our Monte Carlo simulations show that, in the absence of contamination, the standard and robust methods are equivalent. In the presence of outliers, the number of estimated factors increases with the non-robust methods while it remains the same using robust methods. As an application with real data, we study pollutant concentrations PM_{10} measured in the Île-de-France region of France.

Título : Métodos robustos em séries temporais multivariadas

Palavras-Chave : Séries temporais multivariadas, robustez, observações discrepantes, domínio do tempo, domínio da frequência.

Resumo :
Este manuscrito é centrado em propor novos métodos de estimação das funções de autocovariância e autocorrelação matriciais de séries temporais multivariadas com e sem presença de observações discrepantes aleatórias. As funções de autocovariância e autocorrelação matriciais desempenham um papel importante na análise e na estimação dos parâmetros de modelos de série temporal multivariadas. Primeiramente, nós propomos novos estimadores dessas funções matriciais construídas, considerando a abordagem do domínio da frequência por meio do periodograma matricial, um estimador natural da matriz de densidade espectral. Como no caso dos estimadores tradicionais das funções de autocovariância e autocorrelação matriciais, os nossos estimadores também são afetados pelas observações discrepantes. Assim, qualquer análise subsequente que os utilize é diretamente afetada causando conclusões equivocadas. Para mitigar esse problema, nós propomos a utilização de técnicas de estatística robusta para a criação de estimadores resistentes às observações discrepantes aleatórias.
Inicialmente, nós propomos novos estimadores das funções de autocovariância e autocorrelação de séries temporais univariadas considerando a conexão entre o domínio do tempo e da frequência por meio da relação entre a função de autocovariância e a densidade espectral, do qual o periodograma tradicional é o estimador natural. Esse estimador é sensível às observações discrepantes. Assim, a robustez é atingida considerando a utilização do M-periodograma. O M-periodograma é obtido substituindo a regressão por mínimos quadrados com a M-regressão no cálculo das estimativas dos coeficientes de Fourier relacionados ao periodograma. As propriedades assintóticas dos estimadores são estabelecidas. Para diferentes tamanhos de amostras e cenários de contaminação, a performance dos estimadores é investigada. Os resultados empíricos indicam que os métodos propostos provem resultados acurados. Isto é, os métodos propostos obtêm valores próximos aos da função de autocorrelação tradicional no contexto de não contaminação dos dados. Quando há contaminação, os M-estimadores permanecem inalterados. Deste modo, as funções de M-autocovariância e de M-autocorrelação propostas nesta tese são alternativas viáveis para séries temporais com e sem observações discrepantes.
A boa performance dos estimadores para o cenário de séries temporais univariadas motivou a extensão para o contexto de séries temporais multivariadas. Essa extensão é direta, haja vista que somente os coeficientes de Fourier relativos à cada uma das séries univariadas são necessários para o cálculo do periodograma cruzado. Novamente, a relação de dualidade entre o domínio da frequência e do tempo é explorada por meio da conexão entre a função matricial de autocovariância e a matriz de densidade espectral de séries temporais multivariadas. É neste sentido que, o presente artigo propõe a matriz M-periodograma como um substituto robusto à matriz periodograma tradicional na criação de estimadores das funções matriciais de autocovariância e autocorrelação. As propriedades assintóticas são estudas e experimentos numéricos são realizados. Como exemplo de aplicação à dados reais, nós aplicamos as funções propostas no artigo na estimação dos parâmetros do modelo de série temporal multivariada pelo método de Yule-Walker para a modelagem dos dados MP_{10} da região de Vitória/Brasil.
Finalmente, a estimação robusta dos números de fatores em modelos fatoriais aproximados de alta dimensão é considerada com o objetivo de reduzir a dimensionalidade. É sabido que dados discrepantes afetam as matrizes de covariância e correlação. Em adição, técnicas que dependem do cálculo dos autovalores e autovetores dessas matrizes, como a análise de componentes principais e a análise fatorial, são completamente afetadas. Assim, na presença de observações discrepantes, o critério de informação proposto por Bai & Ng (2002) tende a superestimar o número de fatores. De forma a resolver esse problema, nós propomos substituir a matriz de covariância amostral usual pela matriz M-covariância robusta proposta no segundo artigo. Nossas simulações de Monte Carlo mostram, como esperado, que dentro do cenário de não contaminação, os métodos usuais e robustos são equivalentes. Já na presença de observações discrepantes o número estimado de fatores obtidos considerando os autovalores e autovetores da matriz de covariância usual aumenta, enquanto ao se utilizar os autovalores e autovetores da matriz M-covariância estima-se corretamente o verdadeiro número de fatores. Em um problema real, são considerados os dados de MP_{10} da região de Ilê-de-France/França.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

PhD report:

ROBUST METHODS IN MULTIVARIATE TIME SERIES

PhD student:
Higor Henrique Aranda Cotta

Advisors:
Pascal Bondon
CentraleSupélec, L2S - France

Valdério Anselmo Reisen
Universidade Federal do Espírito Santo PPGEA-UFES- Brazil

Vitória
2019

Acknowledgments

I would like to thank my supervisor, Prof. Valdério Reisen, for ALL the time spent together, patient guidance, encouragement, and counseling since the beginning, 10 years ago. What I became as a researcher is thanks to him and he has also been a role model in many aspects. I could not have hoped for a better supervisor. Without his incredible patience and timely wisdom, my thesis work would not have gone so smoothly. His enthusiasm and passion for science were contagious for me and anyone working with him. The dedication that he puts into work is impressive. I am forever grateful for having a mentor like him.

I would also like to thank my other supervisor, Prof. Pascal Bondon. Although we started working only during my Ph.D., Prof. Bondon is now an important part of my academic career. Thanks to him, I could experience living my part of life in a different country where I had amazing academic and life experiences. Through him, I saw how academia works in a different country and what Brazil could do to improve. I have been lucky to have another supervisor who cared so much about my work, and who responded to my questions and queries so promptly even when I was in Brazil, and he was in France. The time spent commuting home by train was pleasant in his company.

I would like to offer my special thanks to Prof. Céline Lévy-Leduc for her contributions to many of my works through good ideas and insights. Her work with Prof. Reisen always motivated and pushed me to do my best.

My deepest thanks to Prof. Márton Ispány. Marton was the external member of the jury of my master's dissertation, and now he is a member of the jury of my Ph.D. thesis. We spent some time together during his research meetings with Prof. Reisen and the discussions were fruitful.

I must thank Prof. Jane Meri and Prof. Neyval Reis Jr, great researchers from UFES. Also, Prof. Neyval has a special place for being the first person, besides my advisor at the time, to read and correct my first paper. I also thank Prof. Josu Arteche for being part of my midterm examination committee. I also deeply thank Prof. Raul de Lacerda for being the head of the ERASMUS smart^2 program, which allowed me to live abroad for a while.

I am very grateful to the professors of UFES, in special, Eliana Zandonade, Adelmo Bertolde and Antônio Pego e Silva for teaching me the little I know about the interesting field of statistics. I must also thank the secretaries Rose Leão and Shanna Pavan. I am also grateful to the professors of IFES, Edilson Luiz do Nascimento (my first advisor), Sergio Nery, Ernani Ribeiro, Chico Rapchan, Elton Moura, Matheus Costa, Vanessa Battestin and Isaura Nobre.

I must express my gratitude to my friends from NuMEs, PPGEA and UFES, Adriano Sgrâncio, Alessandro Sarnaglia, Bartolomeu Zamprogno, Carlo Solci, Dennys Mourão, Edson Zambon, Fátima Leite, Faradiba Serpa, Filipe Carneiro, Juliana Bottoni, Milena Machado, Mônica Castro, Paulo Prezotti, Pedro Berger, Raí Machado, Wanderson Pinto, Wesley Campos, Wilson Benaquio and Wharley Borges, for all direct and indirect help and also for the happy moments.

I passed some time abroad living in France. I was away from my family and during that time, the new friends that I made became my family. I especially thank Amine Agouran, Gabriel Matos, Hidayet Zaimaga, Kuba Orlowski and Ricardo Bertoglio. Thanks to all of you, my staying in France was the best possible ever! You helped me relax from work and you were my company during sad and happy times. You are now part of my family. I would like to express my deep gratitude to all my friends from L2S, Amine Othmane, Aymeric Thibault, Guacira Costa, Filipe Perez(and his wife Rafaela), Kike Diez, Zicheng Liu, Djibrilla Incha and Violeta Rozman. I also want to thank my neighbors from Maison du Brésil/Cité Universitaire, with whom I shared

an amazing time and also to the working staff of the house, Ablo(in memoriam), Denise, Fred, Mohammed, and Simita.

I thank the CNPq, CAPES, ERAMUS and FAPES for the direct and indirect financial support that made this thesis possible.

I also thank my family for all the support and encouragement during my entire life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1 General Introduction

Environmental and climate change impact our lives and the air pollution is one of the biggest health-threatening problems faced by many countries across the globe. The effects of air pollution on human beings are widely investigated and statistical models have been heavily employed in order to establish the association between air pollution levels, meteorological variables, and acute health effects. At this point, air pollutants are usually monitored and recorded over time at multiple sites of interest. Therefore, they may be viewed as a multivariate time series. Due to the nature of air pollution phenomenon, air pollution time series are complex to analyze. Thus, models which adequately describe the physical behavior of the pollution variables are essential for accurately modeling and forecasting the data.

One common issue present in air pollution time series data is the occurrence of high levels of pollutant concentrations influenced by external events that can cause changes in the dynamic of the series. These high-level observations may be due to exogenous forces, such as unfavorable meteorological conditions, failure in the control process, high pollution episodes, etc. In this context, apart from the negative impacts on the environment and on human health, they may cause undesirable statistical properties such as estimates with large bias, spurious model choices and statistical decisions and, in time series, the reduction of the power correlation among other problems. In the statistical methods, these phenomena are usually caused by outliers or aberrant or atypical observations. Therefore, high levels of pollutants cause a similar effect on the statistical functions such as outliers does. From this empirical connection property between outliers and high-levels of concentrations arises the unstable and noneffective problem in any statistical analysis in the air pollution area. Therefore, the use of a robust method becomes a crucial step in any statistical methodology when dealing with pollutant concentrations and correlated variables.

As well discussed in the literature of time series, among all types of outliers, additive outliers are the most harmful type of aberrant observations and can significantly destroy the correlation structure of a time series see, for example, Chan (1992, 1995), Molinares et al. (2009), Cotta, Reisen, Bondon & Stummer (2017) and the references therein. In the study of time series, the sample mean, autocorrelation (ACF), and partial autocorrelation (PACF) functions play a fundamental role on some steps of the analysis and model estimation, and are known to be highly affected by atypical observations. Consequently, if not dealt properly this issue may lead to spurious results and wrong conclusions.

Given this background, one possible way to mitigate the effects of additive outliers in the statistical analysis is to use methods that are robust against outliers. The robust estimation theory

has been extensively studied by the statistical community since the 1970s following the seminal works of Huber and Maronna Huber (1964), Maronna (1976). Several efforts have been done by the time series analysis community in order to weaken the impact of aberrant observations. A concise review of the fundamentals can be found in Zoubir et al. (2012) and a review of some robust methods for the estimation of the autocovariance and autocorrelation function can be found in Dürre et al. (2015).

In this context, as an approach for solving this problem, Ma & Genton (2000) proposed a highly robust estimator of the autocovariance function (ACOVF) and autocorrelation function, denoted by $\hat{\gamma}^{Q_h}$ and $\hat{\rho}^{Q_n}$, respectively. These estimators are based on the Q_n scale estimator proposed by Rousseeuw & Croux (1993), whose asymptotic properties were studied by Lévy-Leduc et al. (2011*b*) for univariate time series. The highly robust performance of $\hat{\gamma}_Q$ motivated its adoption by Molinares et al. (2009) to obtain an estimator of the spectral density function which is robust against additive outliers for the univariate scenario. For the multivariate time series context, Cotta, Reisen, Bondon & Stummer (2017) extended $\hat{\gamma}^{Q_h}$ and $\hat{\rho}^{Q_n}$ to accommodate multiple time series. The estimators of Ma & Genton (2000) and Cotta, Reisen, Bondon & Stummer (2017) does not yield a non-negative definite sample covariance matrix. Thus, the development of robust ACF and ACOVF multivariate estimators that yield non-negative definite matrix is still an open problem and this is one contribution of this thesis.

In addition, time series analysis in the frequency domain is based on the study of the spectral density function from which the periodogram is an estimator. As demonstrated by Molinares et al. (2009), the periodogram lacks robustness properties against outliers. Thus, robust methods to minimize the effect of outliers on the estimation of periodogram have to be considered. In this direction, different robust periodogram methods have been proposed by the literature, see, for instance, Molinares et al. (2009), Zhang & Chan (2005), Li (2008, 2010), Sarnaglia et al. (2016), Reisen, Lévy-Leduc & Taqqu (2017), among others.

Furthermore, it is known that the periodogram can be also computed from the least squares estimates of the Fourier coefficients and is hence sensitive to outliers in data. Thus, to mitigate this problem, one may consider the use of robust regression methods, e.g., a robust regression method, instead of the standard least square method. The M-periodogram is obtained by replacing the standard least squares regression with the robust M-regression method. Recently, this approach has been considered by Sarnaglia et al. (2016) and Reisen, Lévy-Leduc & Taqqu (2017) providing good results in the estimation of the coefficients of Periodic autoregressive and moving average model (PARMA) models and the fractional parameter of autoregressive fractionally integrated moving average (ARFIMA) models, respectively.

Moreover, a very important time series result establishes that, under some assumptions, the time and frequency domains are connected by the Fourier transforms of the autocovariance function and the spectral density. In this thesis, this relationship is considered to propose estimation methods for the autocovariance and autocorrelation functions starting from an estimator of the spectral density. Firstly, the standard periodogram case is considered. However, this approach is not robust and the resulting estimators are sensitive to outliers. The robustness property is achieved when considering the M-periodogram. Thus, the resulting robust autocovariance and autocorrelation functions are positive semi-definite by construction since the periodogram or the M-periodogram is always positive.

Nowadays, thanks to the development of air quality monitoring technology of data sampling units, air pollution data may be collected by air quality monitors scattered across different areas of the region in short time intervals. In this scenario, the collected air pollution time series data may be of order millions and it is coined as Big Data. This high dimension presents new challenges from theoretical and applied points of view due to the enormous number of parameters and coefficients to be estimated by many standard statistical techniques. Another issue arrives

when the number of variables is much bigger than the quantity of sampling units. In addition, the large quantity of variables also poses a problem to the standard descriptive and graphing statistical tools. Many researchers have developed Big Data visualization techniques, analytic models and machine learning based models to conduct data analysis to achieve better accuracy in evaluation and prediction.

In this direction, one possible approach to deal with time series of high dimension is to consider dimension reduction techniques such as principal component analysis (PCA) and factor analysis (FA) before analyzing the data using standard statistical techniques, e.g, time series or multivariate analysis. This approach has been considered in Reisen, Sgrancio, Lévy-Leduc, Bondon, Monte, Cotta & Ziegelmann (2019), Stock & Watson (2002), for air pollution and economic data, respectively. However, since PCA and FA techniques are built using the sample covariance or correlation matrices, they are sensitive to outliers occasioning a spurious dimension reduction or a wrong selection of the number of factors. Therefore, in order to mitigate this problem, a possible solution is to use robust estimators of these matrices. This approach is considered in the recent works of Cotta (2014) and Reisen, Sgrancio, Lévy-Leduc, Bondon, Monte, Cotta & Ziegelmann (2019) where robust estimation methods were applied to PCA and FA, respectively, in an air pollution data set. In this direction, this thesis proposes a robust variant of the information criteria given in Bai & Ng (2002) for selecting the number of factors in an approximate factor model.

Based on the issues discussed above, that is, robust methods in the time and frequency domain to compute a robust autocovariance matrix and time series with high dimension becomes the main core of this thesis. As the first and second contributions, it is proposed a robust estimation method of the autocorrelation and autocovariance matrix functions of stationary univariate and multivariate time series, respectively, starting from the spectral domain. The third contribution is to introduce a robust method for estimating the number of factors in an approximate factor model. As additional contributions, the methodologies are used in real data related to pollutant variables collected at at the stations of the Automatic Air Quality Monitoring Network (AAQMN) of Greater Vitória (GV) and of Île-de-France (IDF) regions with the aim to verify the effect, if any, of high pollution levels in the estimation model and in the dimension reduction. These applications become very important in the context to show how these series can be analyzed with different aims when using robust ACF functions. These contributions are presented in three papers, which are shown in the third, fourth and fifth chapters of this thesis. Not that, since all theoretical results are new and due to some analytical complexities, the asymptotic properties of the estimation methods are not totally established here and they are left for future work. However, the finite sample size investigations clearly demonstrate that the methods perform quite well and support their use in real problems.

In the first paper, we present a robust estimation method for the autocorrelation and autocovariance functions of stationary univariate time series. Some theoretical properties of the estimators are proved and their finite sample size performances are investigated through a numerical simulation study.

The second paper extends the results of the first paper in order to accommodate multivariate time series. We provide some theoretical results regarding the proposed estimators and their performance are also investigated by means of numerical experiments. An application to real data is conducted in order to demonstrate the usefulness of the method.

In the third paper, we propose a robust method for estimating the number of factors in an approximate factor model. We study the effect of additive outliers on the standard estimator of the number of factors and we employ the robust estimator proposed in the second paper to robustify the estimation of the number of factors. We analyze through simulations our proposed method and an application to real data is also considered.

This thesis is structured as follows: This introduction, the research objectives, the study region and the real data used in the applications of the proposed models are presented in Chapter 1. Chapter 2 presents a brief review of the bibliography regarding some robust statistical models and concepts that are essential for the development and understanding of the thesis. As mentioned before, Chapters 3, 4 and 5 are the original contributions and results of this thesis, described in the form of three articles. The general conclusions, final comments, and further research lines are presented in Chapter 6, followed by the references used in this manuscript. To end this, Appendix A displays papers that I have made contributions during my Ph.D. period. Note that these papers are directly connected, from the theoretical and empirical point of views, with this thesis.

2 Objectives

2 .1 General

To analyse multivariate data using any statistical technique or model it is crucial, as one of the first steps, to compute the sample covariance matrix which, in more complex data, such as the case of time series with aberrant observations and/or high dimension, this sample function may give unprecise estimates, as well discussed in the literature. In this context, this thesis proposes to use matrices encountered in time series analysis and Fourier vector basis to obtain the matrix that will approximately diagonalize the sample covariance matrix of univariate and multivariate time series with absolutely summable autocovariance function and series with high dimension. In addition, a robust diagonal matrix is suggested to obtain a robust ACF matrix which displays the robustness property against abrupt observations and heavy-tailed distributions. These proposed methodologies can be very useful in practical situation when dealing with a high dimension reduction, cluster and discriminant analysis, PCA, FA, regression methods among others which are statistical tools widely used in the Air Pollution area. Therefore, the analysis and interpretation of the dynamic of PM_{10} pollutants measured at the stations of the Automatic Air Quality Monitoring Network (AAQMN) of Greater Vitória (GV) and of Île-de-France (IDF) regions are also part of the contribution of this thesis in the sense to verify the effect, if any, of high pollution levels in the estimation model and in the dimension reduction:

2 .2 Specific

The specific objectives are:

- To propose methods for the estimation of autocovariance and autocorrelation functions of univariate and multivariate time series from the connection between time and frequency domains;
- To propose robust alternatives for the estimation of autocovariance and autocorrelation functions of univariate and multivariate time series considering the M-periodogram;
- To propose a robust method based on the proposed robust M-covariance estimator to estimate the factors and to select the number of factors when additive outliers are present in an approximate factor model;
- To study and interpret the dynamic behavior of PM_{10} pollutant data measured at the stations of the Automatic Air Quality Monitoring Network (AAQMN) of Greater Vitória and Île-de-France regions;
- To make available to the scientific community all computer codes and programs created in this thesis to make this research reproducible.

3 Regions of study

3 .1 Brazil

The Greater Vitória Region (GVR) is located on the southeast coast of Brazil (latitude 20º 19S, longitude 40º20W) with a population of approximately 1.900.000 inhabitants. The climate is tropical humid with average temperatures ranging from 24ºC to 30ºC. The region has many ports being an important cargo transport hub in Brazil. Also, there are many industries presented in the region, such as steel plants, iron ore pellet mill, stone quarrying, cement and food industry and asphalt plant. Figures 1.1 and 1.2 present the main pollution sources of the region.

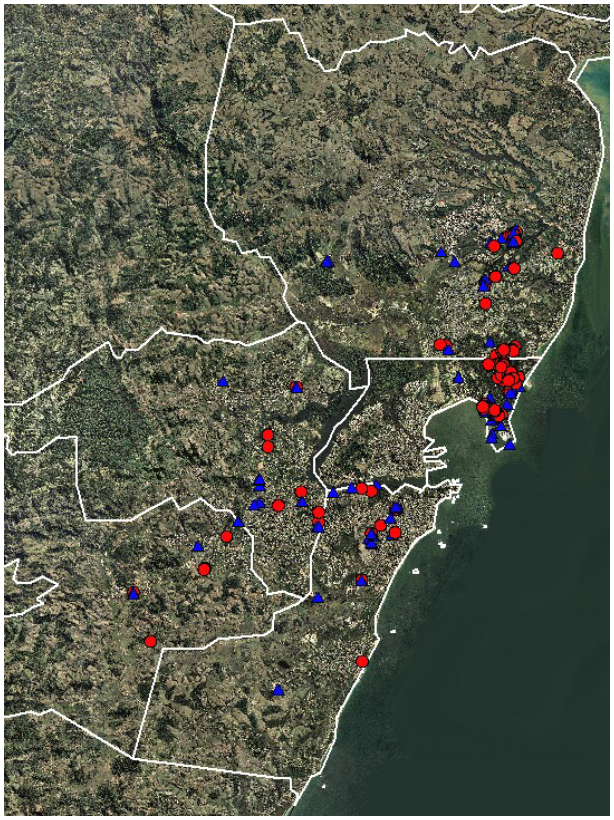


Figure 1.1: Industries present in the region.

The region is characterized by mountainous regions: in the Northwest (Mestre Álvaro) and in the West (Região Serrana), some plains (Airport and mangrove) and plateaus (Planalto Serrano) in the North. In the southern, a plain region (Barra do Jucu). All portions are interspersed by rockies of small and medium size. These conditions, in general, are favorable to wind circulation and the dispersion of pollutants.

This region is the main economic pole of the state which represents approximately 63.13 % of the state's Gross Domestic Product (GDP), where 65.55 % of this sector comes from the tertiary sector, 34.03 % of the secondary sector and 0.42 % of the sector the economy. In this region, steel, pelletizing, mining (quarries), cement, food industry, asphalt plant, etc are found.

The automatic air pollution monitoring network of GVR is consisted by nine monitoring stations distributed in the cities of this region as follows: three stations in Serra (Cidade Continental,Laranjeiras and Carapina), three stations in Vitória (Jardim Camburi, Enseada do Suá and Vitória Centro), two stations in Vila Velha (Vila Velha Centro and Ibes) and one station in Cariacica (at the regional food distribution center, CEASA). Figure 1.3 presents the geographical

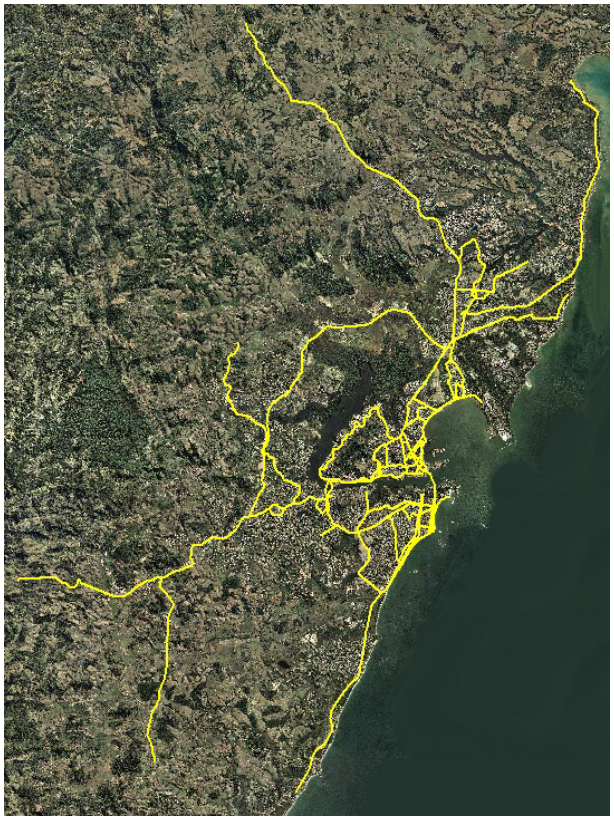


Figure 1.2: Main roads of the region.

location of each station.

The network monitors the following pollutants: Respirable Particles ($PM_{2.5}$), Inhalable Particles (PM_{10}) Total Suspended Particles (TSP), Sulfur Dioxide (SO_2), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Nitrogen Oxides (NO_x), Carbon Monoxide (CO), Ozone (O_3), Methane (CH), Non-methane Hydrocarbons (HCnM) and Total Hydrocarbons (HC). In addition to these pollutants, some meteorological parameters are monitored: Wind Direction (WD), Wind Speed (WS), Standard Deviation of Wind Direction (STDWD), Rainfall (R), Relative Humidity (RH), Temperature (T), Atmospheric pressure (P) and Solar radiation (SR). Not all pollutants and meteorological parameters are monitored by all stations, the list of pollutants and parameters monitored by each station is presented in Table 1.1.

Table 1.1: Pollutants and meteorological parameters measured at the AAQMN from RGV.

Station	$PM_{2.5}$	PM_{10}	PTS	SO_2	NO	NO_2	NO_x	CO	O_3	CH	HCnM	HCT	Meteo. Parameters
Cidade Continental		x	x	x		x							WD,WS
Laranjeiras		x	x	x	x	x	x	x	x				
Carapina		x	x										WD,P,R,SR,STDWD,T,RH,WS
Jardim Camburi		x	x	x	x	x	x						
Enseada do Suá	x	x	x	x	x	x	x	x	x	x	x	x	WD, STDWD,WS
Vitória Centro		x	x	x	x	x	x	x		x	x	x	
Ibes	x	x	x	x	x	x	x	x	x	x	x	x	WD, STDWD,WS,T,RH
Vila Velha Centro		x		x									
Cariacica		x	x	x	x	x	x	x	x				WD, STDWD,WS,T,RH

This thesis considers the data obtained at the stations of the AAQMN of GVR. The data analyzed are hourly concentrations of atmospheric pollutants or meteorological parameters of each station measured in the period from January 2005 to December 2011. The pollutants studied are the PM_{10} and the SO_2. The data set was made available by the Instituto Estadual

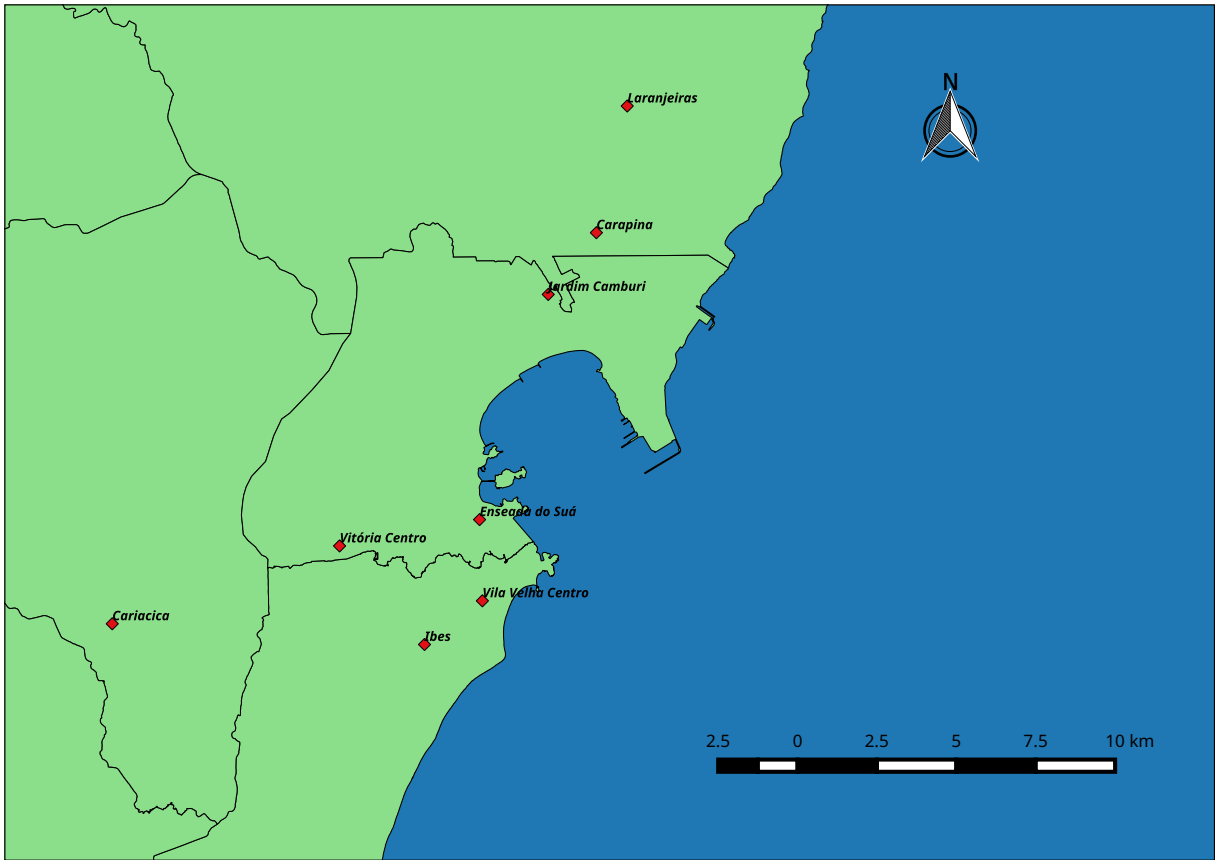


Figure 1.3: Geographical location of the stations of GVR.

do Meio Ambiente (IEMA).

3 .2 France

The Île-de-France (IDF) is a region of France which encompasses Paris and its neighboring cities and departments, namely the departments of Val-d'Oise, Seine-et-Marne, Seine-Saint-Denis, Ville-de-Paris, Hauts-de-Seine, Val-de-Marne, Essonne, and Yvelines. This region is the most populated the metropolitan area containing Paris and the surrounding area has around 12 million inhabitants, 18% of the population of France IAU (2018). The population is concentrated in the highly urbanized area of Paris and immediately surrounding cities. The outer parts of the Ile-de-France remain largely rural, where agriculture land, forest and natural spaces occupy 78 percent of the region. This most important economic region of France concentrates nearly 30 percent of the French GDP and accounts for 23.2% of France's workforce. IAU (2018).

The IDF has a more mature and broader network of stations than RGV. However, as in the case of the AAQMN of RGV, not all stations monitor all pollutants. Thus, a list of IDF stations with their respectively monitored pollutants are presented in Table 1.2. Figure 1.4 presents the geographical location of each station.

The data analyzed are hourly concentrations of atmospheric pollutants of each station measured in the period from March 17th to June 11th of 2019 (91) days ($T = 91$) The pollutant studied is the PM_{10}. The data set was made available by the European Environmental Agency.

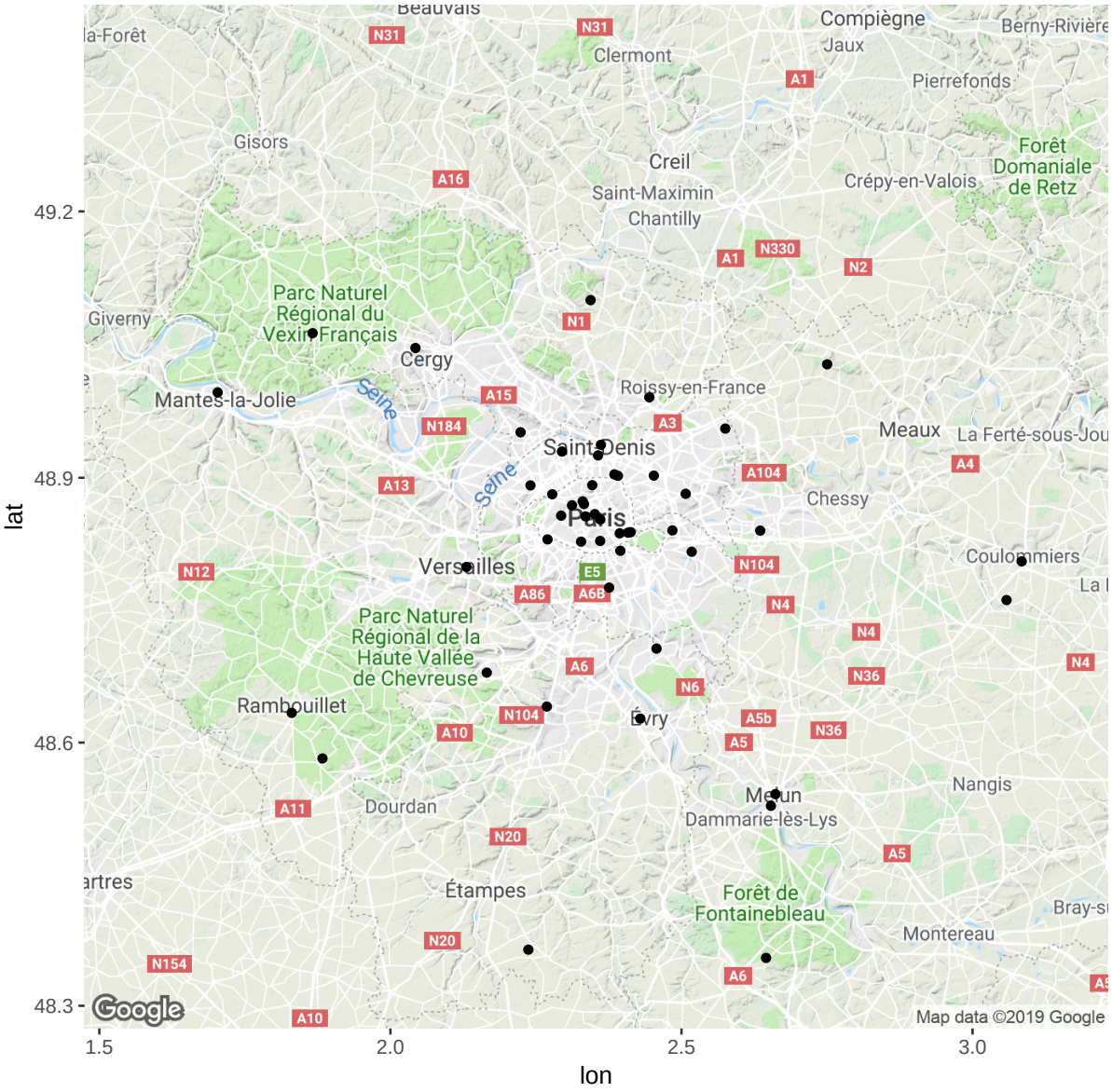


Figure 1.4: Geographical location of the stations of IDF.

Table 1.2: Pollutants measured at the AAQMN from IDF.

City	Station(code)	$PM_{2.5}$	PM_{10}	NO_2	SO_2	O_3	CO
Paris	FR04004		x	x		x	
Paris	FR04012		x	x			x
Paris	FR04014			x			
Paris	FR04031		x	x			
Paris	FR04037			x		x	
Paris	FR04060			x			
Paris	FR04071			x			
Paris	FR04118		x	x			
Paris	FR04131		x	x			
Paris	FR04135			x			
Paris	FR04141			x			
Paris	FR04143	x	x	x		x	x
Paris	FR04179		x	x			
Paris	FR04329	x	x	x			
Hauts-de-Seine	FR04002	x	x	x			
Hauts-de-Seine	FR04017			x	x	x	
Hauts-de-Seine	FR04150		x	x			
Val-de-Marne	FR04034	x	x	x	x	x	
Val-de-Marne	FR04099		x				
Val-de-Marne	FR04101			x		x	
Val-de-Marne	FR04146			x			
Seine-Saint-Denis	FR04018			x	x		x
Seine-Saint-Denis	FR04058	x	x	x			x
Seine-Saint-Denis	FR04059			x			
Seine-Saint-Denis	FR04100			x		x	
Seine-Saint-Denis	FR04123		x	x			
Seine-Saint-Denis	FR04156	x	x	x			
Seine-Saint-Denis	FR04319		x	x		x	
Seine-et-Marne	FR04069			x		x	
Seine-et-Marne	FR04098		x	x		x	
Seine-et-Marne	FR04122	x	x	x			
Seine-et-Marne	FR04142					x	
Seine-et-Marne	FR04173		x				
Seine-et-Marne	FR04324					x	
Seine-et-Marne	FR04328	x		x	x	x	
Essonne	FR04049					x	
Essonne	FR04066	x	x			x	
Essonne	FR04149			x		x	
Essonne	FR04180			x			
Essonne	FR04323			x			
Yvelines	FR04029			x		x	
Yvelines	FR04038			x		x	
Yvelines	FR04063			x		x	
Yvelines	FR04181	x	x			x	
Val-d'Oise	FR04023		x			x	
Val-d'Oise	FR04024	x		x			
Val-d'Oise	FR04048	x				x	
Val-d'Oise	FR04051			x			
Val-d'Oise	FR04158		x			x	

Chapter 2

Brief review of the literature

As mentioned in Chapter 1, the field of robust statistics is not new and several robust techniques have been proposed to deal with diversified types of outliers. Since this thesis is about robust statistical methods, we shall present some a brief description of the definitions and what is already done by the literature.

Data from the most diverse areas of knowledge may present the presence of discrepant observations, outliers, and the failure to adopt adequate techniques capable of behaving this data can lead to inconsistent results. Fox (1972) defines as outliers observations that are not in accord with the others from the same data set. Rousseeuw & Van Zomeren (1990) defined as outliers those observations that deviate from the estimates provided by a statistical model suggested by the bulk of the data set. Huber & Ronchetti (2009) define robust statistics as those that are insensitive to small deviations from the original assumptions, usually, the data have a Gaussian or assumed to be known distribution. Note that despite being related robust statistics are not necessarily non-parametric statistics.

Maronna et al. (2006) define robust statistics as a tool to increase the reliability and accuracy of the model and the statistical analysis. Tukey (1975) affirms that it is possible to use classical, non-robust statistics and robust statistics concomitantly, but one must proceed with caution when the two methods provide different values. Thus, one can consider robust statistics as a statistic that is not sensitive to outliers.

These outliers observations may be due to several factors:

- Measurement error;
- Typesetting error;
- External interventions;
- Extreme observations
- Unexpected alteration of physical conditions.

There are several statistical methods for determining whether or not an observation is an outlier. However, even if an observation is identified as a possible outlier by some statistical method, in certain areas of knowledge, it is a real and observed observation and may have some impact, for example, to the environment and human health. In this context, one can not simply exclude it.

Besides, some methods may identify as outliers observations that are not. For example, the 3-sigma rule identifies as outliers observations that are distant from the sample mean by two or three times the standard deviation. When considering the Gaussian distribution, where the

interval of three standard deviations around the mean contains approximately 99.7 % of the observations, it is expected to identify some observations as outliers.

The common choice made by a wide range of scientists and practitioners to mitigate this problem is to exclude the suspected observations' values from the data set. However, as pointed out by (Maronna et al. 2006, chap. 1), the removal of an atypical observation may lead to other complications since the exclusion is based on subjective decisions.

One possible way of dealing with outliers is to replace the outlying observation with a more plausible value in the approach called filtering. However, since this approach completely modifies the original time series data, it is not considered in this thesis. Therefore, one viable option to mitigate this adversity without modifying the original data set is to use robust statistical methods.

In the context of time series, one may classify outliers as four types of outliers: additive outliers, innovational outliers, level shifts, and temporary changes. Each one of them affects affect the observed time series in its particular way. As pointed by Tsay et al. (2000), the effect of an outlier depends not only on its size and the underlying model but also on the interaction between the size and the dynamic structure of the model, especially with multivariate time series. This thesis is focused on additive outliers since they are the most harmful type as they negatively affect the correlation structure of the time series.

Related to additive outliers in univariate time series, Chan (1992, 1995) studied the impact of additive outliers on the autocovariance and autocorrelation functions. As pointed out by the authors, the increment in the variance of the time series due to the presence of additive outliers will decrease the autocorrelation values. The work of Molinares et al. (2009) presents the impact on the periodogram, an estimator of the spectral density.

For multivariate time series, the impact on the autocorrelation and autocovariance matrix can be found in Cotta, Reisen, Bondon & Stummer (2017) and similar conclusions to the univariate case are obtained when outliers are present in one or more elements of the time series vector.

In the past, some robust autocorrelation estimators have been proposed but the lack of computational power has limited the adoption of these techniques. Nowadays, because of an increment in the processing power of modern computers, the usage of robust methodologies is a workable task. Nonetheless, most of these estimators rely on the computation of robust pair-wise covariances and correlations, but the resulting estimator is not positive semi-definite.

Considering this approach, Ma & Genton (2000) proposed a highly robust estimator of the autocovariance function (ACOVF) and autocorrelation function (ACF), denoted by $\hat{\gamma}^{Q_h}$ and $\hat{\rho}^{Q_n}$, respectively. These estimators are based on the Q_n scale estimator proposed by Rousseeuw & Croux (1993), whose asymptotic properties were studied by Lévy-Leduc et al. (2011*b*) for univariate time series. For the multivariate context, Cotta, Reisen, Bondon & Stummer (2017) extended $\hat{\gamma}^{Q_h}$ and $\hat{\rho}^{Q_n}$ to accommodate multiple time series. Although the estimators of Cotta, Reisen, Bondon & Stummer (2017) do not yield a positive semi-definite matrices, they were used by Reisen, Sgrancio, Lévy-Leduc, Bondon, Monte, Cotta & Ziegelmann (2019) to robustly estimate the number of factors with air pollution data.

Robust ACF and ACOVF may also be obtained by methods based on Signs and Ranks, popular nonparametric statistics. However, they are not very robust against outliers and some transformations required for them to be unbiased may destroy the positive semidefiniteness of the estimators Dürre et al. (2015).

Another approach is to interpret the autocorrelation function as a linear regression and then to apply some robust regression method for its calculation. This method was suggested by

Chang & Politis (2016) and also does not yield a positive semi-definite estimator without being modified.

Some positive semi-definite estimators for the ACF are available in the literature are built from the relationship between the ACF and PACF. However, to obtain an estimator of the ACOVF, the variance of the series must be robustly estimated beforehand.

Alternatively, one might construct an estimator for ACF and ACOVF by robustly estimating, considering, for example, the minimum covariance determinant estimator, the whole autocorrelation matrix at a given lag starting from the data organized in a suitable matrix. The resulting estimator is positive semi-definite, but it does not have the Toeplitz structure with constant off-diagonals.

It is possible to find a review and comparison of some robust methods in Dürre et al. (2015). In this direction, a common problem shared by many of the robust estimation methods is that they do not yield a positive semi-definite covariance or correlation matrix and/or a Toeplitz matrix. In order to solve this problem, many estimators rely on some modification to be positive semi-definite with a Toeplitz structure. In this direction, our approach proposed in this thesis has the advantage that yields a positive semi-definite correlation matrix with Toeplitz structure with no approximation nor modification of the original data set, such as trimming or outlier removal.

Chapter 3

Paper 1: Robust autocovariance estimation from the frequency domain for univariate stationary time series

In this Chapter, we present the first original contribution of this thesis: the robust estimation of the autocovariance and autocorrelation functions of univariate stationary time series. The key idea is to consider the connection between time and frequency domains where the autocovariance function and spectral density are linked through the Fourier transform to construct estimators for the autocovariance function starting from an estimator of the spectral density. We present some theoretical results for our estimators. A simulation study evaluates the performance of the estimators for small sample size with and without occurrence of additive outliers.

Abstract

It is well-know that, under some assumptions, $\mathrm{P}_N \, \Gamma_N \, \mathrm{P}'_N - 2\pi \, \mathrm{D}_N$ converges to zero uniformly as $N \to \infty$, where Γ_N is the covariance matrix of the first N observations from a stationary $\{Y_t\}_{t\in\mathbb{Z}}$ process with absolutely summable covariances and spectral density $f(.)$, D_N is a diagonal $N \times N$ matrix with elements of $f(.)$ and P_N is a vector of eigenvalues of Γ_N. In this context, this paper proposes two estimates of Γ_N in the frequency domain by using the standard periodogram and a $M-$periodogram function. The asymptotic properties of the proposed estimators are established. The empirical investigation shows that the methods display estimates fairly close to the ACF function in the context of non-contaminated data. On the other hand, in the presence of additive outliers, the M-estimator remains unaffected to the presence of additive outliers while, as expected, when using the the classical periodogram the estimates are totally corrupted. Therefore, the ACF M-estimator proposed here becomes an alternative method to estimate Γ_N in time series with and without outliers. These methodologies can be very useful in the context of estimating models with a high-dimension time series data set.

1 Introduction

Atypical observations (outliers) are present in time series of diversified origins. It is well-known that outliers significantly affect the correlation structure of a time series even when only one atypical observation is present, see, for example, Chan (1992, 1995), Molinares et al. (2009) and the references therein. As a possible approach for solving this problem, Ma & Genton (2000) proposed a highly robust estimator of the autocovariance function (ACOVF) and autocorrelation function (ACF), denoted by $\hat{\gamma}_{Q_N}$ and $\hat{\rho}_{Q_N}$, respectively. These estimators are based on the Q_n

scale estimator proposed by Rousseeuw & Croux (1993), whose asymptotic properties were studied by Lévy-Leduc et al. (2011*b*) for univariate time series.

As noticed by Ma & Genton (2000), their robust ACOVF estimator does not provide a non-negative definite sample covariance matrix. Although this is an undesirable property for an autocovariance function estimator, the highly robust performance of $\hat{\gamma}_{Q_N}$ motivated its adoption by Molinares et al. (2009) to obtain an estimator of the spectral density function which is robust against additive outliers.

Time series analysis in the frequency domain is based on the study of the spectral density function from which the periodogram is an estimator. As demonstrated by Molinares et al. (2009), the periodogram lacks robustness properties against outliers. Therefore, robust methods to minimize the effect of outliers on the estimation of periodogram have to be considered. In this direction, different robust periodogram methods have been proposed by the literature, see, for instance, Molinares et al. (2009), Zhang & Chan (2005), Li (2008, 2010), Sarnaglia et al. (2016), Reisen, Lévy-Leduc & Taqqu (2017), among others.

It is known that the periodogram can be obtained directly from a least squares estimates of the Fourier coefficients and is hence sensitive to outliers in data. Thus, to mitigate this problem, one may consider the use of robust regression methods, e.g., a robust M-regression, instead of the standard approach.

Recently, this approach has been considered by Sarnaglia et al. (2016) and Reisen, Lévy-Leduc & Taqqu (2017) providing good results in the estimation of the coefficients of PARMA models and the fractional parameter of ARFIMA models, respectively.

In addition, under some assumptions, $\mathrm{P}_N \Gamma_N \mathrm{P}'_N - 2\pi \mathrm{D}_N$ converge to zero uniformly as $N \to \infty$, where Γ_N is the covariance matrix of the first N observations from a stationary $\{Y_t\}_{t\in\mathbb{Z}}$ process with absolutely summable covariances and spectral density $f(.)$, D_N is a diagonal $N \times N$ matrix with elements of $f(.)$ and P_N is a vector of eigenvalues of Γ_N. This fact establishes the connection between time and frequency domains by means of the Fourier transform.

In this paper, two contributions are established. Firstly, it is demonstrated that the elements of $\mathrm{P}_N \hat{\Gamma}_N \mathrm{P}'_N - 2\pi \mathrm{D}_N$ also converge to zero uniformly as $N \to \infty$, where $\hat{\Gamma}_N$ is built from the standard periodogram with suitable window. The second contribution suggests replacing the standard periodogram with the robust M-periodogram in order to obtain $\hat{\Gamma}^M_N$.

The outline of this paper is as follows: Besides the introduction, Section 2 discusses the estimation of the ACOVF and ACF from the robust M-periodogram. Section 3 summarizes the simulation experiments and the robust performance of the estimators comparing them to Ma & Genton (2000)'s. Concluding remarks are given in Section 6 .

2 The model and the estimation of the autocovariance function

Let $\{Y_t\}$, $t = 1, 2, ...$, be a stationary process with autocovariance function $\gamma_Y(h) = \mathbb{C}ov[Y_t, Y_{t+h}]$, $h = 0, 1, ...$, which satisfies

(**A1**) $\sum_{h=-\infty}^{\infty} |\gamma_Y(h)| < \infty$.

Under Assumption (A1), the spectral density of $\{Y_t\}$, $t = 1, 2, ...$, is defined as

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_Y(h) e^{-ih\lambda}, \qquad for \quad all \quad \lambda \in [-\pi, \pi]. \tag{3.1}$$

Let $\{Y_1, Y_2, ..., Y_N\}$ be the first N observations of $\{Y_t\}_{t\in\mathbb{Z}}$. The covariance matrix of $\{Y_1, Y_2, ..., Y_N\}$ is defined by

$$\Gamma_N = [\gamma(i-j)]_{i,j=1}^{N}. \tag{3.2}$$

The following properties establish, from the time to the frequency domain, a very useful analytical connection between functions, that is, they allow to interpret the spectral density as a multiple of the matrix of covariance of the process multiplied by orthogonal vectors. Actually, this has similar interpretation of the classical result of spectral diagonalization of an any positive semi-definite square matrix.

Proposition 1. *Let $\boldsymbol{A}$ be an $N \times N$ regular circulant matrix with first row $[a_0, \ldots, a_{N-1}]$.*

1. *$\boldsymbol{A}$ has eigenvectors g_j with (not necessarily distinct) eigenvalues*

$$\delta_j = \sum_{h=0}^{N-1} a_h \omega^{jh} = p(\omega_j) = \sum_{h=0}^{N-1} a_h \omega_j^h, \quad j = 0, 1, \ldots, N-1, \tag{3.3}$$

where $\omega_j = \omega^j$, $p(\omega_j) = a_0 + a_1\omega_j + \ldots + a_{N-1}\omega_j^{N-1}$, and

$$g_j = N^{-1/2}[1, \omega^j, \omega^{2j}, \ldots, \omega^{(N-1)}j]'. \tag{3.4}$$

2. *Setting $j = 0$ in (1), $\delta_0 = a_0 + a_1 + ... + a_{n-1}$ is always an eigenvalue.*
3. *The eigenvectors are mutually orthogonal, that is, $g_j^* g_k = 0$ if $k \neq j$ and $g_j^* g_k = 1$ for $k = j$, where g_j^* is the conjugate transpose of g_j.*
4. *If $\boldsymbol{F}$ is an $N \times N$ Fourier matrix then, it is unitary and $\boldsymbol{F}\boldsymbol{A}\boldsymbol{F}^* = \boldsymbol{\Delta}$, that is $\boldsymbol{A}\boldsymbol{F}^* = \boldsymbol{F}^*\boldsymbol{A}$, where $\Delta = diag[\delta_0, \delta_1, \ldots, \delta_{N-1}]$. Also $\boldsymbol{A} = \boldsymbol{F}^*\boldsymbol{\Delta}F$.*

Proof. This proposition is directly derived from 8.18 to 8.28 in Seber (2008). □

The following Corollaries are straightforward derived from Proposition 1.

Corollary 1. *If $\boldsymbol{A}$ is symmetric regular circulant then:*

1. *$a_h = a_{N-h}, h = 1, \ldots, m$, where*

$$m = \begin{cases} N/2 & N \text{ even} \\ (N-1)/2 & N \text{ odd.} \end{cases} \tag{3.5}$$

2. *The eigenvectors are g_j given by Proposition 1.*
3. *The eigenvalues of $\boldsymbol{A}$ are*

$$\delta_j = \sum_{h=0}^{N-1} a_h \cos(2\pi jh/n), \qquad j = 0, \ldots, N-1. \tag{3.6}$$

4. *A spectral decomposition of $\boldsymbol{A}$ is*

$$\boldsymbol{A} = \sum_{j=0}^{N-1} \delta_j g_j g_j^*. \tag{3.7}$$

Corollary 2. *Let δ_j be an eigenvalue of $\boldsymbol{A}$ as in Corollary 1. Then, for N odd, the pair of real orthogonal eigenvectors corresponding to δ_j are*

$$\mathrm{c}_j = (\delta_j + \delta_{N-j})/\sqrt{2} = \sqrt{2/N}[1, \cos\lambda_j, \cos 2\lambda_j, \ldots, \cos(N-1)\lambda_j],$$

and

$$\mathrm{s}_j = \mathrm{i}(\delta_{N-j} + \delta_j)/\sqrt{2} = \sqrt{2/N}[1, sin\lambda_j, \sin 2\lambda_j, \ldots, \sin(N-1)\lambda_j],$$

for $j = 1, \ldots, [N/2]$ and setting $\mathrm{c}_0 = \sqrt{1/N}[1, 1, 1, \ldots, 1]$. Additionally, $\mathrm{P}_N\, A\, \mathrm{P}'_N = \Delta^s$ where $\Delta^s = diag[\delta_0, \delta_1, \delta_1, \ldots, \delta_{[N/2]}, \delta_{[N/2]}]$ and $\mathrm{P}_N = [\mathrm{c}_0, \mathrm{c}_1, \mathrm{s}_1, \ldots, \mathrm{c}_{[N/2]}, \mathrm{s}_{[N/2]}]'$. For the case when N is even, both δ_0 and $\delta_{n/2}$ have multiplicity 1 and $\mathrm{P}_N = [\mathrm{c}_0, \mathrm{c}_1, \mathrm{s}_1, \ldots, 2^{-1/2}\,\mathrm{c}_{N/2}]'$. Again, $\mathrm{P}_N\, A\, \mathrm{P}'_N = \Delta^s$.

Proposition 2. *Let Γ_N be the covariance matrix of the first N observations from $\{Y_t\}_{t\in\mathbb{Z}}$ which satisfies (A1), and let $f(.)$ be its spectral density as given by (4.6). Let $\lambda_j = 2\pi j/N, j = 0, \ldots, [N/2]$, where $[.]$ denotes the integer part of $N/2$. Let D_N be an $N \times N$ matrix,*

$$\mathrm{D}_N = \begin{cases} \mathrm{diag}\{f(0), \ldots, f(\lambda_{[N/2]}), f(\lambda_{[N/2]})\} & \text{if } N \text{ is odd,} \\ \mathrm{diag}\{f(0), \ldots, f(\lambda_{(N-2)/2}), f(\lambda_{(N-2)/2}), f(\lambda_{N/2})\} & \text{if } N \text{ is even.} \end{cases} \tag{3.8}$$

and let P_N be the eigenvectors which will lead to the diagonalization of Γ_N defined in Corollary 2. Then, the components x_{ij}^N of the matrix

$$\mathrm{P}_N\, \Gamma_N\, \mathrm{P}'_N - 2\pi\, \mathrm{D}_N, \tag{3.9}$$

converge to zero uniformly as $N \to \infty$,i.e. $\sup_{1\le i,j\le N} |x_{ij}^{(N)}| \to 0$.

Proof. Consider the results given in Corollaries 1 and 2 and let $\boldsymbol{A} = circ[\gamma(0), \gamma(1), \gamma(2), \ldots, \gamma(2), \gamma(1)]$. Since the elements of the matrix $\Delta^s - 2\pi\, \mathrm{D}_N$ are bounded in absolute value by $\sum_{|h|>[N/2]|\gamma(h)|}$ which converges to zero as $N \to \infty$, it suffices to show that

$$|\, \mathrm{P}_{N,i}\, \boldsymbol{A}\, \mathrm{P}'_{N,j} - \mathrm{P}_{N,i}\, \Gamma_N\, \mathrm{P}'_{N,j}\, | \to 0 \quad \text{uniformly in } i \text{ and } j. \tag{3.10}$$

We have

$$|\, \mathrm{P}_{N,i}(\boldsymbol{A} - \Gamma_N)\, \mathrm{P}'_{N,j}\, | = \left| \sum_{m=1}^{c} (\gamma(m) - \gamma(N-m)) \sum_{k=1}^{m} (p_{ik}p_{j,N-m+k} - p_{i,N-m+k}p_{jk}) \right|, \tag{3.11}$$

where $c = [(N-1)/2]$. This expression is bounded by

$$4n^{-1}\left(2\sum_{m=1}^{c} m|\gamma(m)| + 2\sum_{m=1}^{c} m|\gamma(N-m)|\right) \le 8\sum_{m=1}^{c}\frac{m}{N}|\gamma(m)| + 8\sum_{m=N-c}^{N-1}\frac{c}{N}|\gamma(m)|.$$

The first term converges to zero as $n \to \infty$ by the dominated convergence theorem since the summand is dominated by $|\gamma(m)|$ and $\sum_{m=1}^{\infty}|\gamma(m)| < \infty$. The second term also goes to zero since it is bounded by $\sum_{m=[N/2]}^{\infty}|\gamma(m)|$. Since both terms are independent of i and j, the proof is complete. Similar proof of (3) is given in, for example, Brockwell & Davis (2013), Proposition 4.5.2.

□

Related to Proposition 3, the following remark gives the upper and lower bounds of the covariance matrix of $\{Y_1, Y_2, ..., Y_N\}$, among many other interesting properties which can be derived from.

Remark 1. *Let $\{Y_t\}$ be a stationary process with spectral density $f(.)$ such that*

$$w = \inf_{\lambda} f(\lambda) > 0 \quad \text{and} \quad W = \sup_{\lambda} f(\lambda) < \infty,$$

and denote by $\lambda_1, \ldots, \lambda_N (\lambda_1 \leq \ldots \leq \lambda_N)$ the eigenvalues of the covariance matrix of $\{Y_1, Y_2, ..., Y_N\}$. Then,

$$2\pi w \leq \lambda_1 \leq \lambda_N \leq 2\pi W. \tag{3.12}$$

The proof of Remark 1 is given in Brockwell & Davis (2013).

Apart from its mathematical interpretation, Proposition 3 can also lead to a very useful and elegant result, from statistical and applied point of view, when dealing with a sample from the process from which time series models are built, estimated and tested. This is one of the main core of this paper and it is discussed as follows.

Let now $\{Y_1, Y_2, ..., Y_N\}$ be a sample of $\{Y_t\}_{t\in\mathbb{Z}}$ and $\hat{f}_N(.)$ be an estimator of the spectral density in (4.6). Given $\hat{\mathrm{D}}_N$ as an estimator of D_N in (4.9) by replacing $f(.)$ with $\hat{f}_N(.)$, an alternative estimator of Γ_N can be obtained by

$$\hat{\Gamma}_N = 2\pi \, \mathrm{P}'_N \, \hat{\mathrm{D}}_N \, \mathrm{P}_N \, . \tag{3.13}$$

where P_N was defined previously.

The focus is now on the properties of $\hat{f}_N(.)$ related to $f(.)$ that allow the convergence of $\hat{\gamma}_N(.)$ towards $\gamma(.)$. Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a second order stationary process. Given a sample $\{Y_1, Y_2, ..., Y_N\}$, the classical periodogram function, at the Fourier frequency $\lambda_j = 2\pi j/N, j = 1, \ldots, [N/2]$, is defined as

$$I_N(\lambda_j) = \frac{1}{2\pi N} \left| \sum_{k=1}^{N} Y_k \exp(\mathrm{i}k\lambda_j) \right|^2 . \tag{3.14}$$

Although the periodogram is a natural estimator of $f(.)$, it is well-known that $I_N(.)$ is not a consistent estimator of $f(.)$ in the sense that the variance of $I_N(.)$ does not go to zero as N goes to infinite. In addition, $I_N(.)$ has an erratic and wildly fluctuating form. These features make the periodogram to be a poor estimator of the spectral density $f(.)$ see, for example, Priestley (1981). Sinse $\{Y_t\}_{t\in\mathbb{Z}}$ is a stationary process with autocovariance function that satisfies Assumption 1, one way to obtain an estimator of the spectral density with reduced variance is simply to omit some terms of $I_N(.)$ which correspond to the tail of the sample autocovariance function. In general, omitting the terms in $I_N(.)$ will increase the bias, however if these correspond to the tails of the sample ACF satisfying Assumption 1, this will not seriously affect the bias of this "new" periodogram. In this context, the new periodogram is given as follows and it is usually called truncated window periodogram.

Before introducing a class of consistent estimators of the spectral density of $\{Y_t\}_{t\in\mathbb{Z}}$, the following assumptions are introduced:

(**A2**) $\{M_N := M\}$ is a sequence of positive integers with $M \to \infty$ and $\frac{M}{N} \to 0$ as $N \to \infty$.

(**A3**) $\{W_N(.)\}$ is a sequence of weight functions with $W_N(k) = W_N(-k)$ and $W_N(k) \geq 0$, for all k.

(**A4**) $\sum_{|k|\leq M} W_N(k) = 1$ and $\sum_{|k|\leq M} W_N^2(k) \to 0$ as $N \to \infty$.

In view of (A1)-(A4), a consistent class of estimators has the form

$$\hat{f}_{T,N}(\lambda_j) = (2\pi)^{-1} \sum_{k=-m}^{m} W_N(k) I_N(\lambda_{j+k}). \tag{3.15}$$

For more details of the properties of $\hat{f}_{T,N}$ see, for example, Brockwell & Davis (2013), Priestley (1981) and Fuller (1996).

Theorem 1. *Let $\{Y_1, Y_2, ..., Y_N\}$ be a sample observation of a second order stationary time series $\{Y_t\}_{t\in\mathbb{Z}}$ satisfying (A1). Let $\hat{f}_{T,N}(.)$ be an estimator of the spectral density of $\{Y_t\}_{t\in\mathbb{Z}}$ satisfying (A2) to (A4). Define $\hat{\mathrm{D}}_N$ as in (4.9) but replacing $f(.)$ with $\hat{f}_{T,N}(.)$. Let $\hat{\Gamma}_N$ as in (3.13) where $\hat{\gamma}_N(h)$ is obtained. Then,*

$$\hat{\Gamma}_N - 2\pi\, \mathrm{P}_N'\, \mathrm{D}_N\, \mathrm{P}_N = o_p\left(\frac{1}{\sqrt{N}}\right) \quad as \quad N \to \infty. \tag{3.16}$$

Proof of Theorem 3. Let $\mathrm{e}_i = [\delta_{ij}^{\diamond}]_{i,j=1}^{N}$ where $\delta_{ij}^{\diamond} = 1$ if $i = j$ and $\delta_{ij}^{\diamond} = 0$ if $i \neq j$.

$$\begin{aligned}
\hat{\gamma}_N(h) - \gamma(h) &= \mathrm{e}_{h+1}'(\hat{\Gamma}_N - \Gamma_N)\,\mathrm{e}_1 \\
&= 2\pi(\mathrm{P}_N\,\mathrm{e}_{h+1})'(\hat{\mathrm{D}}_N - \mathrm{D}_N)\,\mathrm{P}_N\,\mathrm{e}_1 \\
&= \frac{1}{N}\{\hat{f}_{T,N}(0) - f(0) + \sqrt{2}\sum_{j=1}^{[N/2]}(\hat{f}_{T,N}(\lambda_j) - f(\lambda_j))\cos(\lambda_j h)\}
\end{aligned} \tag{3.17}$$

By Cauchy–Schwarz inequality

$$\begin{aligned}
((\mathrm{P}_N\,\mathrm{e}_{h+1})'(\hat{\mathrm{D}}_N - \mathrm{D}_N)\,\mathrm{P}_N\,\mathrm{e}_1)^2 &= \frac{1}{N^2}\{1 + \sum_{j=1}^{[N/2]}\cos(\lambda_j h)^2\}\{(\hat{f}_N(0) - f(0))^2 \\
&\quad + \sum_{j=1}^{[N/2]}(\hat{f}_{T,N}(\lambda_j) - f(\lambda_j))^2\} \\
&\leq \{1 + 2[N/2]\}\{(\hat{f}_{T,N}(0) - f(0))^2 \\
&\quad + \sum_{j=1}^{[N/2]}(\hat{f}_{T,N}(\lambda_j) - f(\lambda_j))^2\}.
\end{aligned} \tag{3.18}$$

Based on $(A2)$ to $(A4)$,

$$\mathbb{E}[(\mathrm{P}_N\,\mathrm{e}_{h+1})'(\hat{\mathrm{D}}_N - \mathrm{D}_N)\,\mathrm{P}_N\,\mathrm{e}_1]^2 \leq \frac{1}{N}\sup_{-\pi\lambda\pi}\mathbb{E}[|\hat{f}_{T,N}(\lambda) - f(\lambda)|^2] \tag{3.19}$$

where

$$\sup_{-\pi\lambda\pi}\mathbb{E}[|\hat{f}_{T,N}(\lambda) - f(\lambda)|^2] \to 0,$$

by Remark 1 on page 353 in Brockwell & Davis (2013). Hence, $\hat{f}_{T,N}(.)$ converges in mean square to $f(.)$ uniformly on $[-\pi, \pi]$. □

As discussed in Reisen, Lévy-Leduc & Taqqu (2017) among others, one alternative way to derive the periodogram function $I_N(\lambda_j)$ is based on the Least Square (LS) estimates of a two-

dimensional vector $\boldsymbol{\beta}' = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model

$$Y_i = c'_{Ni}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)}\cos(i\lambda_j) + \beta^{(2)}\sin(i\lambda_j) + \varepsilon_i \ , \ 1 \leq i \leq N, \ \boldsymbol{\beta} \in \mathbb{R}^2 \ , \tag{3.20}$$

where ε_i denotes the deviation of Y_i from $c'_{Ni}\boldsymbol{\beta}$, $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] < \infty$. In the sequel (ε_i) is assumed to be a function of a stationary Gaussian process, see (4.23) for a precise definition. Then,

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\mathrm{Arg\,min}} \sum_{i=1}^{N} (Y_i - c'_{Ni}(\lambda_j)\boldsymbol{\beta})^2 \ , \tag{3.21}$$

where

$$c'_{Ni}(\lambda_j) = (\cos(i\lambda_j) \ \sin(i\lambda_j)) \ . \tag{3.22}$$

The solution of (3.21) is

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (C'C)^{-1}C'\mathbf{Y} \ , \tag{3.23}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)'$, C and $C'C$ are defined by

$$C = \begin{pmatrix} \cos(\lambda_j) & \sin(\lambda_j) \\ \cos(2\lambda_j) & \sin(2\lambda_j) \\ \vdots & \vdots \\ \cos(N\lambda_j) & \sin(N\lambda_j) \end{pmatrix} \tag{3.24}$$

and

$$C'C = \begin{pmatrix} \sum_{k=1}^{N}\cos(k\lambda_j)^2 & \sum_{k=1}^{N}\cos(k\lambda_j)\sin(k\lambda_j) \\ \sum_{k=1}^{N}\cos(k\lambda_j)\sin(k\lambda_j) & \sum_{k=1}^{N}\sin(k\lambda_j)^2 \end{pmatrix} = \frac{N}{2}\,\mathrm{Id}_2 \tag{3.25}$$

where Id_2 is the identity matrix 2 by 2. Hence,

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \frac{2}{N}C'\mathbf{Y} = \frac{2}{N}\left(\sum_{k=1}^{N} Y_k\cos(k\lambda_j) \quad \sum_{k=1}^{N} Y_k\sin(k\lambda_j)\right)' = (\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j), \hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))'. \tag{3.26}$$

In view of (3.14),

$$I_N(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j)\|^2 = \frac{N}{8\pi}\left((\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j))^2 + (\hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))^2\right) =: I_N^{\mathrm{LS}}(\lambda_j) \ , \tag{3.27}$$

where $\|\cdot\|$ denotes the classical Euclidean norm and $\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j), \hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))'$ is the least square estimates of $\boldsymbol{\beta}' = (\beta^{(1)}, \beta^{(2)})$ see, for example, Fajardo et al. (2018) and Reisen, Lévy-Leduc & Taqqu (2017) and references therein. Note that $I_N(\lambda_j)$ in (3.27) can be derived for different choices of ε_i, $i = 1, \ldots, N$.

It is now discussed an alternative spectral estimator which is robust against outliers and heavy-tailed distribution.

For the results discussed here, it is supposed that, in Equation 3.20,

$$\varepsilon_i = G(\eta_i) \ , \tag{3.28}$$

where, G is a non null real-valued and skew symmetric measurable function (*i.e.* $G(-x) = -G(x)$, for all x) and $(\eta_i)_{i\geq 1}$ is a stationary Gaussian process with zero mean and unit variance. Additional assumptions of $(\eta_i)_{i\geq 1}$ will be given in the sequel of the paper.

Let now $\psi(.)$ be a function satisfying the following assumptions.

(**A5**) $0 < \mathbb{E}[\psi^2(\varepsilon_1)] < \infty$.

(**A6**) The function ψ is absolutely continuous with its almost everywhere derivative ψ' satisfying $\mathbb{E}[|\psi'(\varepsilon_1)|] < \infty$ and such that the function $z \mapsto \mathbb{E}[|\psi'(\varepsilon_1 - z) - \psi'(\varepsilon_1)|]$ is continuous at zero.

(**A7**) ψ is nondecreasing, $\mathbb{E}[\psi'(\varepsilon_1)] > 0$ and $\mathbb{E}[\psi'(\varepsilon_1)^2] < \infty$.

(**A8**) ψ is skew symmetric, *i.e.* $\psi(-x) = -\psi(x)$, for all x.

It is now introduced the M-periodogram which presents similar performance (from theoretical and empirical meaning) to $I_N(\lambda)$, $\lambda \in [-\pi, \pi]$, but with robustness property against additive outliers and asymmetric and heavy-tail distributions. The M-periodogram is based on the M-estimator $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}$ of the parameter $\boldsymbol{\beta}$ defined in Equation (3.20). The M-estimator $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}} = (\hat{\beta}_N^{(1)}, \hat{\beta}_N^{(2)})'$ is defined as the solution (t_1, t_2) of

$$\sum_{i=1}^{N} \cos(i\lambda_j)\, \psi(Y_i - \cos(i\lambda_j)t_1) = 0 \text{ and } \sum_{i=1}^{N} \sin(i\lambda_j)\, \psi(Y_i - \sin(i\lambda_j)t_2) = 0. \tag{3.29}$$

$\hat{\beta}_N^{(1)}$ and $\hat{\beta}_N^{(2)}$ can be also seen as the minimizers with respect to t_1 and t_2, respectively, of

$$\left| \sum_{i=1}^{N} \cos(i\lambda_j)\, \psi(Y_i - \cos(i\lambda_j)t_1) \right| \text{ and } \left| \sum_{i=1}^{N} \sin(i\lambda_j)\, \psi(Y_i - \sin(i\lambda_j)t_2) \right|, \tag{3.30}$$

where ψ satisfies the same assumptions as in Koul & Surgailis (2000). By analogy to (3.27), the robust periodogram $I_N^M(\lambda_j)$ at $\lambda_j = 2\pi j/N, j = 1, \ldots, [N/2]$, is defined by

$$I_N^M(\lambda_j) = \frac{N}{8\pi} \|\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}(\lambda_j)\|^2 = \frac{N}{8\pi} \left((\hat{\beta}_N^{(1)}(\lambda_j))^2 + (\hat{\beta}_N^{(2)}(\lambda_j))^2 \right) . \tag{3.31}$$

The asymptotic properties of $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}$ are established in the short and long-range dependence frameworks in Reisen, Lévy-Leduc, Cotta, Bondon & Ispany (2019), Reisen, Lévy-Leduc & Taqqu (2017) and Reisen, Lévy-Leduc & Taqqu (2017), respectively. In the case of this paper, that is, in short-range dependence process, the following assumptions are introduced.

(**A9**) Let η_t, $t \in \mathbb{Z}$, be i.i.d. standard Gaussian random variables and let a_j be real numbers such that $\sum_{j \geq 0} |a_j| < \infty$ and $a_0 = 1$. Then,

$$\varepsilon_i = \sum_{j \geq 0} a_j \eta_{i-j}.$$

(**A10**) ψ is the Huber function that is $\psi(x) = \max[\min(x, c), -c]$, for all x in $\mathbb{R}$, where c is a positive constant.

Theorem 2. *Assume that (A9) and (A10) hold and that $\boldsymbol{\beta} = \mathbf{0}$ in (3.20) so that $Y_i = \varepsilon_i$. Then, for any fixed j, $\hat{\boldsymbol{\beta}}_N^M$ defined by (3.30) satisfies*

$$\sqrt{\frac{N}{2}} (F(c) - F(-c)) \hat{\boldsymbol{\beta}}_N^M(\lambda_j) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Delta}^{(j)}\right), \; N \to \infty ,$$

where F is the c.d.f. of ε_1 and

$$\boldsymbol{\Delta}^{(j)} = \sum_{k \in \mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\} \begin{pmatrix} \cos(k\lambda_j) & \sin(k\lambda_j) \\ -\sin(k\lambda_j) & \cos(k\lambda_j) \end{pmatrix}.$$

Theorem 2 is proved in Section 5 of Reisen, Lévy-Leduc, Cotta, Bondon & Ispany (2019).

Corollary 3. *Under the assumptions of Theorem 2, $I_N^M(\lambda_j)$ defined in (3.31) satisfies for any fixed j,*

$$I_N^M(\lambda_j) \xrightarrow{d} \frac{X^2+Y^2}{4\pi(F(c)-F(-c))^2}, \quad as\ N\to\infty\ ,$$

where

$$X \sim \mathcal{N}\left(0, \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\cos(k\lambda_j)\right),\ Y \sim \mathcal{N}\left(0, \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\cos(k\lambda_j)\right)$$

and

$$\mathbb{C}ov(X,Y) = \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\sin(k\lambda_j).$$

The proof of Corollary 3 is a straightforward consequence of Theorem 2 and (3.31).

As previously mentioned, the main objective of this paper is to obtain a robust ACF function which satisfies all assumptions of the definition of ACF presented in Proposition 1.5.1 and Definition 1.5.1 (non-negative definiteness) (Theorem 1.5.1) in Brockwell & Davis (2013). In order to obtain such estimator, in (3), the diagonal elements of matrix D_N is replaced by the robust spectral estimator $I_N^M(.)$. This leads to the following equation

$$\hat{\Gamma}_N^M = 2\pi\, \mathrm{P}_N'\, \hat{\mathrm{D}}_n^M\, \mathrm{P}_N, \tag{3.32}$$

where $\hat{\mathrm{D}}_N^M$ is defined similarly as (4.9) but replacing $f(.)$ by $I_N^M(.)$.

Based on the above discussion and the simulation results presented in Section 3 , the follow statement is proposed:

Statement 1. *Let $\{Y_1, Y_2, ..., Y_N\}$ be a sample observation of a second order stationary time series $\{Y_t\}_{t\in\mathbb{Z}}$ satisfying (A1). Let $I_N^M(.)$ be an estimator of the spectral density of $\{Y_t\}_{t\in\mathbb{Z}}$ and define $\hat{\mathrm{D}}_N$ as in (4.9) but replacing $f(.)$ with $I_N^M(.)$. Let $\hat{\Gamma}_N$ as in (3.32) where $\hat{\gamma}_N^M(h)$ is obtained. Suppose that (A1) to (A10) hold. Then, $\hat{\gamma}_N^M(h) - \gamma(h) = o_p\left(\frac{1}{\sqrt{N}}\right)$ as $N\to\infty$ for $h = 0, \ldots, N-1$.*

Due to some analytical complexities, the proof of this result will be left for the version of the paper that will be submitted.

Now, given a sample $\{Y_1, Y_2, ..., Y_N\}$ of $\{Y_t\}_{t\in\mathbb{Z}}$, the $N\times N$ autocorrelation matrix $\hat{\boldsymbol{\rho}}_N$ and its robust version $\hat{\boldsymbol{\rho}}_N^M$ are, respectively,

$$\hat{\boldsymbol{\rho}}_N = \frac{\hat{\Gamma}_N}{\hat{\gamma}_{11}}, \tag{3.33}$$

and

$$\hat{\boldsymbol{\rho}}_N^M = \frac{\hat{\Gamma}_N^M}{\hat{\gamma}_{11}^M}, \tag{3.34}$$

where $\hat{\gamma}_{11} = \hat{\gamma}_N(0)$ and $\hat{\gamma}_{11}^M = \hat{\gamma}_N^M(0)$. Finally, the ACOVF and ACF, $\hat{\gamma}_N(.)$ and $\hat{\rho}_N(.)$, and their robust counterparts, namely $\hat{\gamma}_N^M$, $\hat{\rho}_N^M$, are extracted from the rows of the circulant matrices $\hat{\Gamma}_N$, $\hat{\boldsymbol{\rho}}_N$ $\hat{\Gamma}_N^M$ and $\hat{\boldsymbol{\rho}}_N^M$, respectively.

3 Monte Carlo experiments

This section reports a Monte Carlo simulation study to investigate the performance of the robust sample ACOVF and ACF estimators discussed previously. For the numerical experiments, the data generating process of $\{Y_t\}_{t\in\mathbb{N}}$ is an autoregressive process of order 1 ($AR(1)$) as follows:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \tag{3.35}$$

where $|\phi| < 1$ and ε_t is a zero mean Gaussian white noise process with variance σ^2.

Let $\{Y_1, Y_2, ..., Y_N\}$ be a realization of $\{Y_t\}_{t\in\mathbb{N}}$, the standard biased ACF estimator is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -N < h < N \tag{3.36}$$

where

$$\hat{\gamma}(h) = N^{-1} \sum_{t=1}^{N-|h|} (Y_{t+|h|} - \bar{Y})(Y_t - \bar{Y}), \quad -N < h < N \tag{3.37}$$

where $\bar{Y} = N^{-1}\sum_{t=1}^{N} Y_t$.

The contaminated process $\{Z_t\}_{t\in\mathbb{N}}$ is

$$Z_t = Y_t + \omega\delta_t, \tag{3.38}$$

where ω is the magnitude of the outliers affecting $\{Y_t\}$. $\{\delta\}_{t\in\mathbb{N}}$ are independent random variables with $\mathsf{Pr}(\delta_t = -1) = \mathsf{Pr}(\delta_t = 1) = p/2$, and $\mathsf{Pr}(\delta_t = 0) = 1 - p$ with $0 < p < 1$. Notice that δ_t is the product of Bernoulli(p) and Rademacher independent random variables. It follows that $\mathbb{E}[\delta_t] = 0$ and $\mathbb{V}ar[\delta_t] = p$.

For the purpose of comparison between the proposed sample ACF and the Robust ACF of Ma & Genton (2000), the latter is summarized in the sequel. Given a sample $\{Y_1, Y_2, ..., Y_N\}$ of $\{Y_t\}_{t\in\mathbb{Z}}$, based on the Q_N scale estimator proposed by Rousseeuw & Croux (1993), Ma & Genton (2000) suggested the following highly robust estimator of the ACOVF

$$\hat{\gamma}_{Q_N}(h) = \frac{1}{4}\left[Q^2_{N-h}(U+V) - Q^2_{N-h}(U-V)\right], \tag{3.39}$$

where U and V are vectors containing the initial $N-h$ and the final $N-h$ observations of $\{Y_1, Y_2, ..., Y_N\}$, respectively. Then, the autocorrelation function can be obtained from

$$\hat{\rho}_{Q_N}(h) = \frac{Q^2_{N-h}(U+V) - Q^2_{N-h}(U-V)}{Q^2_{N-h}(U+V) + Q^2_{N-h}(U-V)}, \tag{3.40}$$

where U and V are also vectors containing the initial $N-h$ and the final $N-h$ of $\{Y_1, Y_2, ..., Y_N\}$. It can be shown that $|\hat{\rho}_{Q_N}(h)| \leq 1$.

The asymptotic results of the above robust autocovariance in time series with short and long memory properties were the motivation for the papers of Lévy-Leduc et al. (2011*b*) and Lévy-Leduc et al. (2011*a*). Theorem 4 in Lévy-Leduc et al. (2011*b*) presents the central limit theorem for the autocorrelation given by (3.39). The non-positive definiteness property of (3.39) was one of the motivations to propose a new robust autocovariance and autocorrelation function estimators.

In the simulations, $\phi = 0.7$, $\sigma^2 = 1$, $\omega = 15$ and $p = 0.05, 0.10$ and 0.15 are set. The sample sizes are $N = 200, 500$ and 1000, and each experiment is replicated 1000 times. Two scenarios are

considered: (i) the samples are uncontaminated ($p = 0$), and (ii) the samples are contaminated ($p \neq 0$). Under both scenarios, the comparison between the estimators is done by contrasting the plots and the averages of empirical root mean square error (RMSE) and bias of the theoretical $\rho(h) = \phi^h$ with $\hat{\rho}(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, for $h = 0, 1, \ldots, 8$.

In this direction, Figure 4.1 displays the plots of $\rho(h)$ and the means of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$ for the uncontaminated scenario. The case of $p = 0.05$ is shown in Figure 4.2. For the uncontaminated case, we observe similar behavior in all the plots indicating that all the estimators capture the correlation structure of the series. The effects of additive outliers appear by comparing the true correlation to the sample estimates under a contaminated scenario, see Figure 4.2. Not surprisingly, $\hat{\rho}(h)$ and $\hat{\rho}^N(h)$ were completely affect while $\hat{\rho}^M(h)$ and $\hat{\rho}^{Q_N}(h)$ provided values much closer to the theoretical ones. Related to Figures 4.1 and 4.2, Figure 3.3 and 3.4 show the boxplot of the simulated ACF values for both scenarios, respectively.

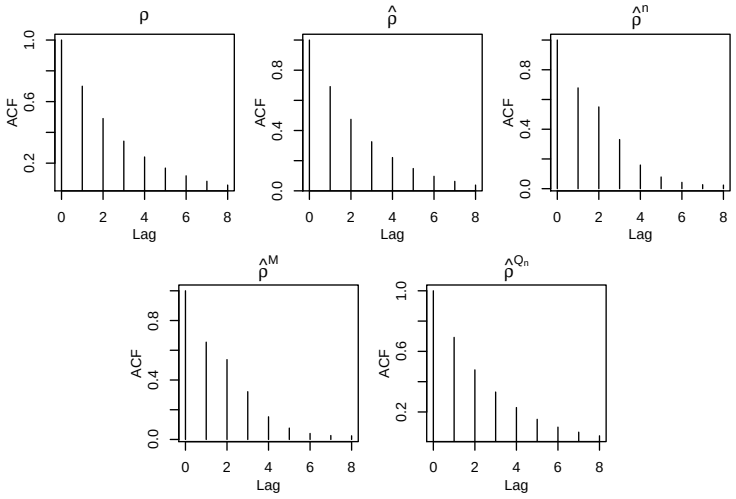


Figure 3.1: Autocorrelation function of z_t. From left to right and top to bottom, plots are $\rho(h)$, $\hat{\rho}(h)$, $\hat{\rho}^N(h)$, and $\hat{\rho}^M(h)$, $\hat{\rho}^{Q_N}(h)$ when $p = 0$.

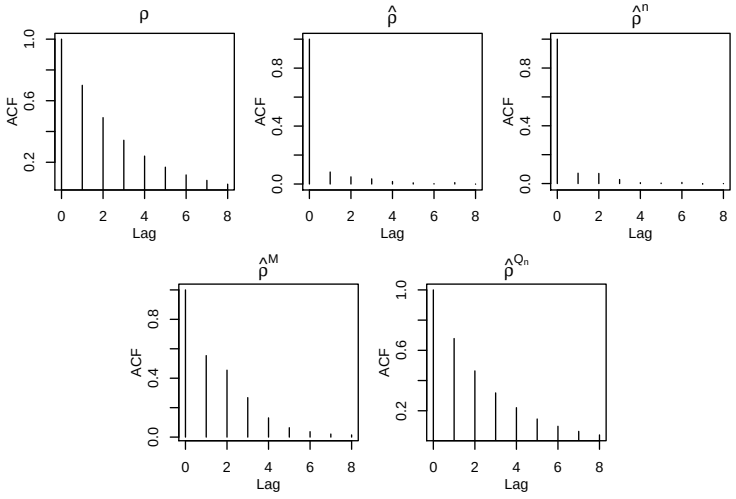


Figure 3.2: Autocorrelation function of z_t. From left to right and top to bottom, plots are $\rho(h)$, $\hat{\rho}(h)$, $\hat{\rho}^N(h)$, and $\hat{\rho}^M(h)$, $\hat{\rho}^{Q_N}(h)$ when $p = 0.05$.

In Table 4.1, we present the root mean squared errors (RMSE) of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$ as N increases, for $h = 0, 1, \ldots, 8$ and $p = 0$. The results for the contaminated scenario $p \neq 0$

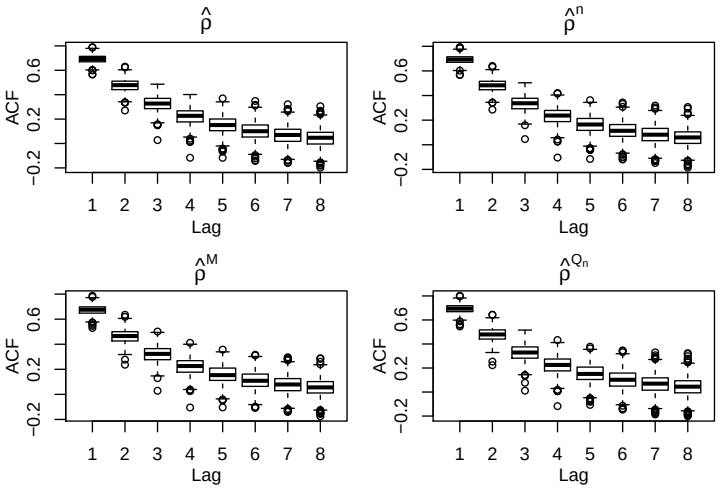


Figure 3.3: Boxplots of estimated ACF of z_t when $p = 0$.

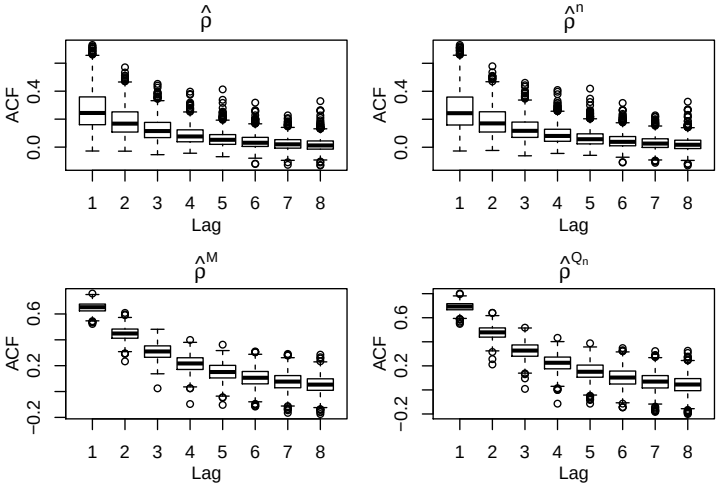


Figure 3.4: Boxplots of estimated ACF of z_t when $p = 0.05$.

are displayed in Tables 4.3, 3.4 and 3.5, for $p = 0.05$, $p = 0.10$ and $p = 0.15$, respectively. From Table 4.1, we observe that the samples and the robust estimators have RMSE close to the each other in the absence of contamination. As expected, $\hat{\rho}(h)$ and $\hat{\rho}_N(h)$ performed best. Comparing $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, we find that $\hat{\rho}_{Q_N}(h)$ has a slight better performance than $\hat{\rho}_N^M(h)$ for $h = 1$. Therefore, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$ are useful even in the context which the presence of additive outliers is uncertain.

Table 3.1: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0$.

	N	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	200	0.0514	0.0794	0.0962	0.1060	0.1118	0.1137	0.1141	0.1134
	500	0.0347	0.0540	0.0645	0.0696	0.0726	0.0729	0.0742	0.0752
	1000	0.0231	0.0352	0.0425	0.0475	0.0506	0.0521	0.0528	0.0532
$\hat{\rho}_N(h)$	200	0.0510	0.0753	0.0894	0.0983	0.1049	0.1079	0.1085	0.1092
	500	0.0344	0.0526	0.0626	0.0676	0.0706	0.0709	0.0724	0.0738
	1000	0.0230	0.0347	0.0415	0.0463	0.0494	0.0510	0.0517	0.0524
$\hat{\rho}_N^M(h)$	200	0.0635	0.0832	0.0926	0.0987	0.1030	0.1050	0.1050	0.1060
	500	0.0461	0.0610	0.0670	0.0699	0.0707	0.0701	0.0707	0.0719
	1000	0.0352	0.0437	0.0464	0.0485	0.0502	0.0509	0.0511	0.0518
$\hat{\rho}_{Q_N}(h)$	200	0.0533	0.0835	0.1018	0.1131	0.1218	0.1235	0.1252	0.1270
	500	0.0373	0.0589	0.0702	0.0761	0.0785	0.0777	0.0794	0.0810
	1000	0.0251	0.0383	0.0459	0.0503	0.0538	0.0555	0.0564	0.0576

Table 3.2: Bias of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0$.

	N	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	200	-0.0193	-0.0279	-0.0322	-0.0340	-0.0316	-0.0305	-0.0298	-0.0292
	500	-0.0078	-0.0126	-0.0136	-0.0150	-0.0161	-0.0167	-0.0167	-0.0164
	1000	-0.0054	-0.0081	-0.0091	-0.0104	-0.0113	-0.0112	-0.0110	-0.0092
$\hat{\rho}_N(h)$	200	-0.0186	-0.0175	-0.0134	-0.0089	-0.0024	0.0013	0.0036	0.0050
	500	-0.0073	-0.0081	-0.0055	-0.0041	-0.0035	-0.0031	-0.0026	-0.0020
	1000	-0.0051	-0.0058	-0.0051	-0.0051	-0.0050	-0.0043	-0.0038	-0.0018
$\hat{\rho}_N^M(h)$	200	-0.0384	-0.0367	-0.0287	-0.0201	-0.0104	-0.0041	-0.0008	0.0015
	500	-0.0273	-0.0279	-0.0216	-0.0160	-0.0117	-0.0086	-0.0065	-0.0046
	1000	-0.0250	-0.0254	-0.0206	-0.0162	-0.0130	-0.0096	-0.0070	-0.0038
$\hat{\rho}_{Q_N}(h)$	200	-0.0170	-0.0252	-0.0294	-0.0302	-0.0293	-0.0269	-0.0284	-0.0302
	500	-0.0078	-0.0131	-0.0146	-0.0151	-0.0161	-0.0162	-0.0157	-0.0158
	1000	-0.0053	-0.0080	-0.0089	-0.0097	-0.0106	-0.0100	-0.0097	-0.0082

Now, considering the occurrence of outliers, in Table 4.3, we see that the RMSE of the sample estimate is much larger than the RMSE of the robust estimators when the percentage of contamination is 5%, and, thus, confirming that even a small fraction of contamination can make the sample ACF useless. Moreover, the RMSE of the robust estimators are almost the same in the uncontaminated and the contaminated cases, but, again, $\hat{\rho}_{Q_N}(h)$ performs better than $\hat{\rho}_N^M(h)$ for $h = 1, 2, 3$.

Table 3.3: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0.05$.

	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	0.6377	0.4510	0.3186	0.2252	0.1617	0.1186	0.0887	0.0693
$\hat{\rho}_N(h)$	0.6365	0.4493	0.3165	0.2228	0.1591	0.1160	0.0866	0.0673
$\hat{\rho}_N^M(h)$	0.1418	0.1151	0.0940	0.0798	0.0722	0.0677	0.0659	0.0659
$\hat{\rho}_{Q_N}(h)$	0.0432	0.0642	0.0741	0.0768	0.0786	0.0783	0.0797	0.0809

To empirically investigate the breakdown point of the proposed estimator, the RMSE of $\hat{\rho}_N^M(h)$ as the percentage of outliers in $\{z_t\}$ increases are presented in Tables 4.3 and 3.4, respectively. Comparing both tables, note that increasing the percentage of outliers reduces the performance

of both estimators. However, not surprisingly, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$ are less affect by the outliers, although $\hat{\rho}_{Q_N}(h)$ still performs slightly better than $\hat{\rho}_N^M(h)$.

Table 3.4: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0.10$.

	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	0.6574	0.4645	0.3280	0.2339	0.1668	0.1220	0.0916	0.0713
$\hat{\rho}_N(h)$	0.6558	0.4627	0.3256	0.2317	0.1644	0.1197	0.0894	0.0694
$\hat{\rho}_N^M(h)$	0.2317	0.1717	0.1268	0.0984	0.0803	0.0711	0.0668	0.0625
$\hat{\rho}_{Q_N}(h)$	0.0549	0.0722	0.0779	0.0787	0.0786	0.0783	0.0799	0.0800

Table 3.5: RMSE of $\hat{\rho}(h)$, $\hat{\rho}_N(h)$, $\hat{\rho}_N^M(h)$ and $\hat{\rho}_{Q_N}(h)$, for $h = 0, 1, \ldots, 8$ and $p = 0.15$.

	0	1	2	3	4	5	6	7	8
$\hat{\rho}(h)$	0.6587	0.4688	0.3325	0.2363	0.1700	0.1228	0.0922	0.0725	
$\hat{\rho}_N(h)$	0.6572	0.4669	0.3304	0.2339	0.1678	0.1211	0.0904	0.0709	
$\hat{\rho}_N^M(h)$	0.3091	0.2232	0.1630	0.1195	0.0943	0.0766	0.0677	0.0612	
$\hat{\rho}_{Q_N}(h)$	0.0738	0.0860	0.0861	0.0834	0.0835	0.0804	0.0802	0.0795	

4 Conclusion

This paper presented a new estimation method for the autocovariance and autocorrelation functions of stationary univariate processes with absolutely summable autocovariance function. The procedure consists in replacing the traditional periodogram by the robust M-periodogram in the inverse diagonalization procedure of the matrix containing the estimated spectral density. The proposed method is also robust to additive outliers. Therefore, the authors suggest the use of the proposed method in a time series in which there are occurrences of additive outliers and/or heavy-tail distribution.

Acknowledgments

The authors gratefully acknowledges partial financial support from FAPES/ES, CAPES/Brazil and CNPq/Brazil and, ERASMUS/France and CentraleSupélec/Frace. The project is cofinanced by the European Union and the European Social Fund.

Chapter 4

Paper 2: A robust alternative method for the estimation of the covariance and the correlation matrices for multivariate time series

A natural extension of the estimators proposed in the first paper constructed in order to consider multivariate stationary time series. The approach is similar of the one from the first paper. That is, first we obtain the Fourier or robust Fourier coefficients from each series individually. Then, from the Fourier coefficients we construct the periodogram matrix which is used to construct the estimator of the autocovariance matrix function of the time series vector. A simulation study is presented to investigate the performance of the estimators for some finite sample sizes. An application to an air pollution data set is also considered.

This Chapter presents the second original contribution:

Abstract

This paper extends the results from Cotta, A. Reisen & Bondon (2017) for the case of multivariate time series with absolutely summable autocovariance function. A very important time series result establishes that, under some assumptions, the time and frequency domains are connected by the Fourier transform of the autocovariance and the spectral density. This connection allows each element of the matrix $\boldsymbol{H}_N \boldsymbol{\Gamma}_N^{\boldsymbol{Y}} \boldsymbol{H}_N^* - 2\pi\, \mathrm{D}_N$ converges to zero uniformly as $N \to \infty$, where, $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$ is the covariance matrix of the first N observations from a r-dimensional stationary process $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ with absolutely summable covariances and spectral density matrix $\boldsymbol{f}^{\boldsymbol{Y}}(.)$, D_N is a $2N \times 2N$ diagonal matrix with elements of $\boldsymbol{f}_{ij}^{\boldsymbol{Y}}(.)$ and $\boldsymbol{H}_N$ is a matrix of eigenvectors of $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$. That is, the first method considers the duality connection between the autocovariance matrix function and the spectral density matrix and implements an inverse diagonalization procedure considering the periodogram matrix and the Fourier transform vector basis in order to obtain a new estimator for $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$. However, as in the standard sample autocovariance matrix function case, this new estimator is also sensitive to additive outliers. The robustness property is achieved by replacing the standard periodogram matrix with the M-periodogram matrix. The asymptotic properties of the two proposed estimators are established. The finite sample size investigation shows that both methods perform close to the standard sample autocovariance matrix function in the case of non-contaminated data. Under the contaminated data scenario, the estimator built using the M-periodogram matrix remains unaffected while, as expected, the one employing the classical periodogram matrix is totally corrupted. Hence, the autocovariance and autocor-

relation matrices M-estimators proposed are a viable alternative estimators of $\mathbf{\Gamma}_N^{\boldsymbol{Y}}$ and $\boldsymbol{\rho}_N^{\boldsymbol{Y}}$, respectively, for multivariate time series with and without outliers.

1 Introduction

It is well-known that outliers may lead to a complete destruction of the correlation structure and, thus, leading to model misspecification and wrong conclusions. See, for example, Chan (1995), Molinares et al. (2009), Cotta, Reisen, Bondon & Stummer (2017) and the references therein. This issue is found for univariate and multivariate time series. However, as pointed by Tsay et al. (2000), most of the outliers studies are devoted only to the univariate time series or to the multivariate time uncorrelated processes.

In order to address this issue in a multivariate time series context, Cotta, Reisen, Bondon & Stummer (2017) studied the impact of additive outliers on the autocovariance (ACOVF) and autocorrelation (ACF) matrices functions and proposed robust estimation methods for these functions as a way to mitigate the impact of outlying observations. The proposed estimators are based on the univariate robust estimators proposed by Ma & Genton (2000) which make use of the Q_n scale estimator proposed by Rousseeuw & Croux (1993). Although highly robust, the autocovariance functions of Ma & Genton (2000) and Cotta, Reisen, Bondon & Stummer (2017) do not yield positive definite matrices.

The duality between time and frequency domains is given by the autocovariance function and the spectral density connected though the Fourier transform. In this direction, Cotta, A. Reisen & Bondon (2017) proposed two estimators for the autocovariance and autocorrelation functions of univariate stationary time series. The ACOVF and ACF are obtained from the inverse diagonalization procedure of the matrix containing the periodogram. The robust version of the estimators is obtained by fitting a robust harmonic regression to obtain a robust version of the discrete Fourier transform, and, thus the so-called M-periodogram that was studied by Reisen, Lévy-Leduc & Taqqu (2017), Fajardo et al. (2018) and Reisen, Lévy-Leduc, Cotta, Bondon & Ispany (2019) for short and long-memory processes. These estimators yield positive definite matrices by construction since the periodogram and the M-periodogram are strictly positive.

Therefore, in this paper, we shall extend to the multivariate stationary time series context the robust estimator of the autocovariance and the autocorrelation functions proposed by Cotta, A. Reisen & Bondon (2017) from the frequency domain. The robust estimators are obtained from the robust M-periodogram matrix which is achieved by calculating the Fourier coefficients from the robust M-regression. Then, the multivariate ACOVF and ACF are obtained.

The outline of this paper is as follows: besides the introduction, Section 2 presents the multivariate time series model and discusses the estimation of the ACOVF and ACF from the cross-periodogram. Section 3 presents the robust ACOVF and ACF obtained from the M-cross-periodogram. Section 4 summarizes the simulation experiments and the robust performance of the estimators comparing them to the standard sample estimators. Concluding remarks are given in Section 6 .

2 Time series model with additive outliers

2 .1 Linear Time Series

Let $\boldsymbol{Y}_t = [Y_{1t}, Y_{2t}, \ldots, Y_{rt}]'$, $t \in \mathbb{Z}$ be a r-dimensional, $r \in \mathbb{N}$, linear vector process defined by

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{t-j}, \tag{4.1}$$

where $\mathbb{E}(\boldsymbol{Y}_t) = \boldsymbol{\mu}$ and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_r]'$, $\boldsymbol{\Psi}_0$ is the identity $r \times r$ matrix, $\boldsymbol{\Psi}_j$, $j = 1, \ldots, \infty$ are $r \times r$ matrices of coefficients satisfying $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}\|^2 < \infty$, where $\|\boldsymbol{\Psi}\|$ is the norm of matrix $\boldsymbol{\Psi}$ defined by $\|\boldsymbol{\Psi}\|^2 = \mathrm{Tr}(\boldsymbol{\Psi}'\boldsymbol{\Psi})$. In 4.1, the $\boldsymbol{\varepsilon}_t$ form a vector white noise processes with $\boldsymbol{\Psi} = [\varepsilon_{1t}, \ldots, \varepsilon_{rt}]'$ such that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}$ and $\mathbb{C}ov(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t+h}) = \Sigma_{\boldsymbol{\varepsilon}} \mathbb{1}_{\{h=0\}}$, where $\mathbb{1}_{\{h=0\}} = \boldsymbol{I}$, if $h = 0$ and $\mathrm{I}_{\{h\neq 0\}} = \mathbf{0}$ otherwise. Although the elements of $\{\boldsymbol{\varepsilon}_t\}$ at different times are uncorrelated, they may be contemporaneously correlated.

It results from (4.1) that

$$\boldsymbol{\Gamma}^{\boldsymbol{Y}}(h) = \mathbb{C}ov(\boldsymbol{Y}_t, \boldsymbol{Y}_{t+h}) = \textstyle\sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \Sigma_{\boldsymbol{\varepsilon}} \boldsymbol{\Psi}'_{j+h}, \quad h \geq 0, \tag{4.2}$$

is a $r \times r$ matrix where the (i,j)th element, $i, j = 1, \ldots, r$, of $\boldsymbol{\Gamma}^{\boldsymbol{Y}}(h)$ is denoted by $\gamma_{ij}^{\boldsymbol{Y}}(h)$.

The lag-h, $h \in \mathbb{Z}$ autocorrelation matrix function of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ is defined by

$$\boldsymbol{\rho}^{\boldsymbol{Y}}(h) = \boldsymbol{C}^{-1/2} \boldsymbol{\Gamma}^{\boldsymbol{Y}}(h) \boldsymbol{C}^{-1/2}, \tag{4.3}$$

where $\boldsymbol{C}^{-1/2} = \mathrm{diag}\left[\sqrt{\gamma_{11}^{\boldsymbol{Y}}(0)}, \ldots, \sqrt{\gamma_{rr}^{\boldsymbol{Y}}(0)}\right]$. The (i,j)th element of $\boldsymbol{\rho}^{\boldsymbol{Y}}(h)$ is

$$\rho_{ij}^{\boldsymbol{Y}}(h) = \frac{\mathbb{C}ov(Y_{it}, Y_{j,(t+h)})}{\sqrt{\mathbb{V}ar(Y_{it})\mathbb{V}ar(Y_{jt})}} = \frac{\gamma_{ij}^{\boldsymbol{Y}}(h)}{\sqrt{\gamma_{ii}^{\boldsymbol{Y}}(0)\gamma_{jj}^{\boldsymbol{Y}}(0)}}. \tag{4.4}$$

We here denote by $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}(h)$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}(h)$ the standard sample estimates of $\boldsymbol{\Gamma}^{\boldsymbol{Y}}(h)$ and $\boldsymbol{\rho}^{\boldsymbol{Y}}(h)$, respectively, i.e., the estimate that we obtain by replacing the unknown covariances in (4.3) by their sample estimates.

In the sequel, for simplicity of presentation and without loss of generality, let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a bi-dimensional multivariate time series with $r = 2$, where $\boldsymbol{Y}_t = [Y_{1t}, Y_{2t}]'$, which satisfies the assumption

(A11) $\sum_{h=-\infty}^{\infty} |\gamma_{ij}^{\boldsymbol{Y}}(h)| < \infty$, for $i, j = 1, 2$.

Under Assumption (A11), the spectral density matrix of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ is defined as

$$\boldsymbol{f}^{\boldsymbol{Y}}(\lambda) = \left[f_{ij}^{\boldsymbol{Y}}(\lambda)\right]_{i,j=1}^{2} = \begin{bmatrix} \boldsymbol{f}_{11}^{\boldsymbol{Y}} & \boldsymbol{f}_{12}^{\boldsymbol{Y}} \\ \boldsymbol{f}_{21}^{\boldsymbol{Y}} & \boldsymbol{f}_{22}^{\boldsymbol{Y}} \end{bmatrix}, \quad \lambda \in [-\pi, \pi], \tag{4.5}$$

where

$$f_{ij}^{\boldsymbol{Y}}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_{ij}^{\boldsymbol{Y}}(h) e^{-\mathrm{i}h\lambda}. \tag{4.6}$$

Let $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ be the first N, $N \in \mathbb{N}$, observations of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$. The $2N \times 2N$ covariance matrix of $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ is defined by

$$\boldsymbol{\Gamma}_N^{\boldsymbol{Y}} = [\boldsymbol{\Gamma}_{N,ij}^{\boldsymbol{Y}}]_{i,j=1}^{2} = \begin{bmatrix} \boldsymbol{\Gamma}_{N,11}^{\boldsymbol{Y}} & \boldsymbol{\Gamma}_{N,12}^{\boldsymbol{Y}} \\ \boldsymbol{\Gamma}_{N,21}^{\boldsymbol{Y}} & \boldsymbol{\Gamma}_{N,22}^{\boldsymbol{Y}} \end{bmatrix}, \tag{4.7}$$

where $\boldsymbol{\Gamma}_{N,ij}^{\boldsymbol{Y}} = [\gamma_{N,ij}^{\boldsymbol{Y}}(l-m)]_{l,m=1}^{N}$ and $\gamma_{N,ij}^{\boldsymbol{Y}}(.) = \gamma_{ij}^{\boldsymbol{Y}}(.)$.

Proposition 3. *Let $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$ be the covariance matrix of the first N, $N \in \mathbb{N}$, observations from $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ which satisfies (A11), and let $\boldsymbol{f}^{\boldsymbol{Y}}(.)$ be its spectral density matrix as given by* (4.5)*. Let*

$\lambda_k = 2\pi k/N, k = 0, \ldots, N-1$. *Let* D_N *be an* $2N \times 2N$ *matrix,*

$$\boldsymbol{D}_N = [\boldsymbol{D}_{N,ij}]_{i,j=1}^2 = \begin{bmatrix} \boldsymbol{D}_{N,11} & \boldsymbol{D}_{N,12} \\ \boldsymbol{D}_{N,21} & \boldsymbol{D}_{N,22} \end{bmatrix}, \tag{4.8}$$

where

$$\mathrm{D}_{N,ij} = \mathrm{diag}[f_{ij}^{\boldsymbol{Y}}(0), f_{ij}^{\boldsymbol{Y}}(\lambda_1), \ldots, f_{ij}^{\boldsymbol{Y}}(\lambda_{(N-1)})]. \tag{4.9}$$

Define a transformation matrix $\boldsymbol{H}_N$ *by*

$$\boldsymbol{H}_N = \begin{bmatrix} \boldsymbol{G}_N & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_N \end{bmatrix}, \tag{4.10}$$

where $\boldsymbol{G}_N$ *is an* $N \times N$ *matrix where the rows are the eigenvectors which shall lead to the diagonalization of* $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$*, given by*

$$g_{N,k} = N^{-1/2}[1, e^{-\mathrm{i}\pi k/N}, e^{-\mathrm{i}\pi 2k/N}, \ldots, e^{-\mathrm{i}\pi(N-1)k/N}], k = 0 \ldots, N-1. \tag{4.11}$$

Let $\boldsymbol{H}_N^*$ *be the conjugate transpose of* $\boldsymbol{H}_N$*.*

Then, every element of the matrix

$$\boldsymbol{H}_N \boldsymbol{\Gamma}_N^{\boldsymbol{Y}} \boldsymbol{H}_N^* - 2\pi\, \mathrm{D}_N, \tag{4.12}$$

converge to zero uniformly as $N \to \infty$*.*

See the proof in Fuller (1996), Theorem 7.4.1.

Proposition 3 is very interesting from statistical point of view as it establishes the connection between time and frequency domains for multivariate stationary processes through the relation between the autocovariance matrix and the spectral density matrix. We shall now focus on considering this dual connection to construct an estimator for the autocovariance matrix function starting from an estimator of the spectral density matrix.

Let $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ be a sample of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ and $\hat{\boldsymbol{f}}_N^{\boldsymbol{Y}}(.)$ be an estimator of the spectral density matrix in (4.5). Given $\hat{\mathrm{D}}_N$ as an estimator of D_N in (4.8) by replacing $\boldsymbol{f}^{\boldsymbol{Y}}(.)$ with $\hat{\boldsymbol{f}}_N^{\boldsymbol{Y}}(.)$, an estimator for $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$ is given by

$$\hat{\boldsymbol{\Gamma}}_N^{\star\boldsymbol{Y}} = 2\pi \boldsymbol{H}_N^* \hat{\mathrm{D}}_N \boldsymbol{H}_N. \tag{4.13}$$

Let $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ be a sample of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$, a natural estimator of the spectral density matrix of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ is obtained by replacing $\gamma_{ij}^{\boldsymbol{Y}}(.)$ with $\hat{\gamma}_{ij}^{\boldsymbol{Y}}(.)$ in (4.6) for $i, j = 1, 2$. Thus, the periodogram matrix is defined by

$$\boldsymbol{I}_N^{\boldsymbol{Y}}(\lambda_k) = [I_{N,ij}^{\boldsymbol{Y}}(\lambda_k)]_{i,j=1}^2 = \begin{bmatrix} I_{N,11}^{\boldsymbol{Y}}(\lambda_k) & I_{N,12}^{\boldsymbol{Y}}(\lambda_k) \\ I_{N,21}^{\boldsymbol{Y}}(\lambda_k) & I_{N,22}^{\boldsymbol{Y}}(\lambda_k) \end{bmatrix}, \tag{4.14}$$

where

$$I_{N,ij}^{\boldsymbol{Y}}(\lambda_k) = \frac{1}{2\pi} \sum_{h=-(N-1)}^{N-1} \hat{\gamma}_{N,ij}^{\boldsymbol{Y}}(h) \exp(-\mathrm{i}h\lambda_k), \tag{4.15}$$

and $\lambda_k = 2\pi k/N$, with $k = 0, \ldots, N-1$.

The focus is on the properties of $\hat{\boldsymbol{f}}_N^{\boldsymbol{Y}}(.)$ related to $\boldsymbol{f}^{\boldsymbol{Y}}(.)$ that allows the convergence of $\hat{\boldsymbol{\Gamma}}_N^{\star\boldsymbol{Y}}$ towards $\boldsymbol{\Gamma}_N^{\boldsymbol{Y}}$. As discussed in Priestley (1981), without a suitable lag window, the periodogram matrix is not a consistent estimator of the spectral density since the variance of $\boldsymbol{I}_N^{\boldsymbol{Y}}(.)$ does not go to zero as $N \to \infty$. In order to obtain a consistent class of estimators of the spectral density matrix of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$, the following assumptions are introduced:

(**A12**) $\{M_N\}$ is a sequence of positive integers, $M_N \to \infty$ and $\frac{M_N}{N} \to 0$ as $N \to \infty$.

(**A13**) $\{W_N(.)\}$ is a sequence of weight functions with $W_N(k) = W_N(-k)$ and $W_N(k) \geq 0$, for all k.

(**A14**) $\sum_{|k|\leq M_N} W_N(k) = 1$ and $\sum_{|k|\leq M_N} W_N^2(k) \to 0$ as $N \to \infty$.

Under Assumptions (A11)-(A14), a consistent class of estimators has the form

$$\boldsymbol{I}_N^{\star\boldsymbol{Y}}(\lambda_k) = [I_{N,ij}^{\star\boldsymbol{Y}}(\lambda_k)]_{i,j=1}^2 = \begin{bmatrix} I_{N,11}^{\star\boldsymbol{Y}}(\lambda_k) & I_{N,12}^{\star\boldsymbol{Y}}(\lambda_k) \\ I_{N,21}^{\star\boldsymbol{Y}}(\lambda_k) & I_{N,22}^{\star\boldsymbol{Y}}(\lambda_k) \end{bmatrix}, \tag{4.16}$$

where

$$I_{N,ij}^{\star\boldsymbol{Y}}(\lambda_k) = \frac{1}{2\pi} \sum_{h=-M_{N,ij}}^{M_{N,ij}} W_{N,ij}(k) I_{N,ij}^{\boldsymbol{Y}}(\lambda_{k+h}), \tag{4.17}$$

and $\lambda_k = 2\pi k/N$, with $k = 0, \ldots, N-1$. The quantity $M_{N,ij}$ is called truncation point and the function $W_{N,ij}(.)$ is called lag window. It should be noted that as advocated in Priestley (1981), there is no reason for the lag window nor the truncation point to be the same for all (ij)-elements of $\boldsymbol{I}_N^{\star\boldsymbol{Y}}(.)$. However, they should be chosen to match the rating of decay of each $\hat{\gamma}_{N,ij}^{\boldsymbol{Y}}(.)$ function. Henceforth, it is assumed that all $\hat{\gamma}_{N,ij}^{\boldsymbol{Y}}(.)$ for $i, j = 1, 2$, have the same rate of decay, truncation point and lag window.

Theorem 3. *Let $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ be an N observed sample of a second order stationary time series $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ satisfying (A11). Let $\boldsymbol{I}_N^{\star\boldsymbol{Y}}(.)$ be an estimator of the spectral density matrix of $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$. Define $\hat{\mathrm{D}}_N$ as in (4.9), but replacing $\boldsymbol{f}^{\boldsymbol{Y}}(.)$ with $\boldsymbol{I}_N^{\star\boldsymbol{Y}}(.)$. Let $\hat{\boldsymbol{\Gamma}}_N^{\star\boldsymbol{Y}}$ as in (4.13). Then,*

$$\hat{\boldsymbol{\Gamma}}_N^{\star\boldsymbol{Y}} - 2\pi \boldsymbol{H}_N^* \mathrm{D}_N \boldsymbol{H}_N = o_p\left(\frac{1}{\sqrt{N}}\right) \quad as \quad N \to \infty. \tag{4.18}$$

Proof. Taken each $\mathrm{D}_{N,ij}$ and $\hat{\mathrm{D}}_{N,ij}$, $\hat{\boldsymbol{\Gamma}}_{N,ij}^{\star\boldsymbol{Y}}$ and $\boldsymbol{\Gamma}_{N,ij}^{\boldsymbol{Y}}$ as in (4.9), (4.13) and (4.7), respectively , for $i, j = 1, 2$. When, $i = j$, the proof is presented in Theorem 1 of Cotta, A. Reisen & Bondon (2017). The case when $i \neq j$ is obtained by arguments completely analogous to those of Theorem 1 in Cotta, A. Reisen & Bondon (2017).

□

3 Robust estimation from the robust M-cross-periodogram

3 .1 Additive outlier model

A parametric class of linear time series satisfying (4.1) is the vector autoregressive moving average (VARMA) model of orders (p, q) defined by the difference equation

$$\boldsymbol{\Phi}(B)(\boldsymbol{Y}_t - \boldsymbol{\mu}) = \boldsymbol{\Theta}(B)\boldsymbol{\varepsilon}_t, \tag{4.19}$$

where B is the backward shift operator ($B\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1}$), $\boldsymbol{\Phi}(B) = I - \sum_{i=1}^{p} \boldsymbol{\Phi}_i B^i$ and $\boldsymbol{\Theta}(B) = I + \sum_{i=1}^{q} \boldsymbol{\Theta}_i B^i$ where $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Theta}_i$ are $r \times r$ matrices, and $\{\boldsymbol{\varepsilon}_t\}$ is a vector white noise process. When the polynomials $\boldsymbol{\Phi}(z)$ and $\boldsymbol{\Theta}(z)$ satisfy $\det(\boldsymbol{\Phi}(z)) \neq 0$ and $\det(\boldsymbol{\Theta}(z)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$, (4.19) has a unique stationary causal and invertible solution and the matrices $\boldsymbol{\Psi}_j$ are determined uniquely by $\boldsymbol{\Psi}(z) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j z^j = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$ for $|z| \leq 1$. We suppose that the observed time series $\{\boldsymbol{Z}_t\}$ results from the contamination of $\{\boldsymbol{Y}_t\}$ by additive random outliers, i.e.,

$$\boldsymbol{Z}_t = \boldsymbol{Y}_t + \boldsymbol{\Omega}\boldsymbol{\delta}_t, \tag{4.20}$$

where $\boldsymbol{\Omega} = \text{diag}[\omega_1, \ldots, \omega_r]$ and $\omega_i, i = 1, \ldots, r$, is the magnitude of the outliers which affects $\{Y_{it}\}$, $\boldsymbol{\delta}_t = [\delta_{1t}, \ldots, \delta_{rt}]'$ is a random vector indicating the occurrence of an outlier at time t. We assume that $\{\boldsymbol{Y}_t\}$ and $\{\boldsymbol{\delta}_t\}$ are uncorrelated processes and that $\mathbb{P}(\delta_{it} = -1) = \mathbb{P}(\delta_{it} = 1) = p_i/2$, $\mathbb{P}(\delta_{it} = 0) = 1 - p_i$ for $i = 1, \ldots, r$ where $0 \leq p_i < 1$. Then $\mathbb{E}(\delta_{it}) = 0$ and $\mathbb{V}ar(\delta_{it}) = p_i$. We assume also that $\mathbb{C}ov(\boldsymbol{\delta}_t, \boldsymbol{\delta}_t) = \Sigma_{\boldsymbol{\delta}} = \text{diag}[p_1, \ldots, p_r]$ and that $\mathbb{C}ov(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t+h}) = 0$ when $h \neq 0$.

It follows from (4.20) that $\mathbb{E}(\boldsymbol{Z}_t) = \mathbb{E}(\boldsymbol{Y}_t)$, $\boldsymbol{\Gamma}^{\boldsymbol{Z}}(0) = \boldsymbol{\Gamma}^{\boldsymbol{Y}}(0) + \boldsymbol{\Omega}\Sigma_{\boldsymbol{\delta}}\boldsymbol{\Omega}'$ and $\boldsymbol{\Gamma}^{\boldsymbol{Z}}(h) = \boldsymbol{\Gamma}^{\boldsymbol{Y}}(h)$ when $h \neq 0$. Therefore

$$\rho_{ij}^{\boldsymbol{Z}}(h) = \begin{cases} \dfrac{\gamma_{ij}^{\boldsymbol{Y}}(h)}{\sqrt{(\gamma_{ii}^{\boldsymbol{Y}}(0)+p_i\omega_i^2)(\gamma_{jj}^{\boldsymbol{Y}}(0)+p_j\omega_j^2)}}, & h \neq 0, \\ \dfrac{\gamma_{ij}^{\boldsymbol{Y}}(0)+p_i\omega_i^2\,\mathbb{1}_{\{i=j\}}}{\sqrt{(\gamma_{ii}^{\boldsymbol{Y}}(0)+p_i\omega_i^2)(\gamma_{jj}^{\boldsymbol{Y}}(0)+p_j\omega_j^2)}}, & h = 0. \end{cases} \tag{4.21}$$

We observe that $\rho_{ij}^{\boldsymbol{Z}}(h) \to 0$ as $|\omega_i| \to \infty$ or $|\omega_j| \to \infty$ when $h \neq 0$, these conclusions are deeper analyzed in Remark 2. The recent works of Lévy-Leduc et al. (2011*b*,*a*) and Chan (1995) discuss this problem in univariate time series with short and long memory properties.

Remark 2. *Suppose that $\{\boldsymbol{Z}_{1t}, \boldsymbol{Z}_{2t}, \ldots, \boldsymbol{Z}_{nt}\}$ are r-dimensional time series observations following model* (4.20) *and m is the observed number of additive outliers. Let $\hat{\rho}_{ij}^{\boldsymbol{Z}}(h) = \hat{\gamma}_{ij}^{\boldsymbol{Z}}(h)/(\sqrt{\hat{\gamma}_{ii}^{\boldsymbol{Z}}(0)\hat{\gamma}_{jj}^{\boldsymbol{Z}}(0)})$, for all $i, j = 1, \ldots, r$. Then*

a. For $m = 1$ (one outlier occurring only at Z_i),

$$\lim_{n\to\infty} \operatorname*{plim}_{\omega_i\to\infty} \hat{\rho}_{ij}^{\boldsymbol{Z}}(h) = 0.$$

b. For $m = 2$ (two outliers occurring at Z_{it} or/and at $Z_{j,t}$) and assuming that $\hat{\gamma}_{ij}^{\boldsymbol{Z}}(h) \neq 0$, for Z_{it} and Z_{jt}, it follows

$$\lim_{n\to\infty} \operatorname*{plim}_{\substack{\omega_i\to\infty \\ and/or \\ \omega_j\to\infty}} \hat{\rho}_{ij}^{\boldsymbol{Z}}(h) = 0.$$

In (a) and (b), w_i and w_j are the magnitudes of the additive outliers occurring at position i and j, respectively.

3 .2 Robust estimation method

It is known that a given zero-mean stationary univariate time series $\{Y_t\}_{t=1,\ldots,N}$ can be represented as a sum involving N sines and cosines at the Fourier frequencies $\lambda_k = 2\pi k/N, k = 0, \ldots, N-1$. The classical periodogram of $\{Y_t\}$ at frequency λ_k is

$$I_N^Y(\lambda_k) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} Y_t \exp(-\text{i}t\lambda_k)\right|^2.$$

As discussed in Reisen, Lévy-Leduc & Taqqu (2017), one alternative way to derive the periodogram function $I_N^Y(.)$ is based on the Least Square (LS) estimates of a bi-dimensional vector $\boldsymbol{\beta'} = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model

$$Y_i = c'_{Ni}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)}\cos(i\lambda_j) + \beta^{(2)}\sin(i\lambda_j) + \varepsilon_i \ , \ 1 \leq i \leq N, \ \boldsymbol{\beta} \in \mathbb{R}^2 \ , \tag{4.22}$$

where ε_i denotes the deviation of Y_i from $c'_{Ni}\boldsymbol{\beta}$ and $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] < \infty$. In the sequel, (ε_i) is assumed to be a function of a stationary Gaussian process.

It is supposed that

$$\varepsilon_i = G(\eta_i), \tag{4.23}$$

where G is a non null real-valued and skew symmetric measurable function (*i.e.* $G(-x) = -G(x)$, for all x) and $(\eta_i)_{i\geq 1}$ is a stationary Gaussian process with zero mean and unit variance. Additional assumption of $(\eta_i)_{i\geq 1}$ is given in (A19).

It can be shown that

$$I_N^Y(\lambda_k) = \frac{N}{8\pi}\,||\hat{\boldsymbol{\beta}}(\lambda_k)||^2 = \frac{N}{8\pi}\left(\hat{\beta}_1(\lambda_k)^2 + \hat{\beta}_2(\lambda_k)^2\right), \tag{4.24}$$

where $\hat{\boldsymbol{\beta}}(\lambda_k)$ is the least squares regression solution

$$\hat{\boldsymbol{\beta}}(\lambda_k) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{t=1}^{N} (Y_t - C_t'(\lambda_k)\boldsymbol{\beta})^2, \tag{4.25}$$

with the regressors $C_t(\lambda_k) = [\cos(t\lambda_k), \sin(t\lambda_k)]'$.

We robustify the periodogram by replacing the least squares regression with the M-regression.

Let now $\psi(.)$ be a function satisfying the following assumptions:

(**A15**) $0 < \mathbb{E}[\psi^2(\varepsilon_1)] < \infty$.

(**A16**) The function ψ is absolutely continuous with its almost everywhere derivative ψ' satisfying $\mathbb{E}[|\psi'(\varepsilon_1)|] < \infty$ and such that the function $z \mapsto \mathbb{E}[|\psi'(\varepsilon_1 - z) - \psi'(\varepsilon_1)|]$ is continuous at zero.

(**A17**) ψ is non-decreasing, $\mathbb{E}[\psi'(\varepsilon_1)] > 0$ and $\mathbb{E}[\psi'(\varepsilon_1)^2] < \infty$.

(**A18**) ψ is skew symmetric, *i.e.* $\psi(-x) = -\psi(x)$, for all x.

(**A19**) Let η_t, $t \in \mathbb{Z}$, be i.i.d. standard Gaussian random variables and let a_j be real numbers such that $\sum_{j\geq 0} |a_j| < \infty$ and $a_0 = 1$. Then,

$$\varepsilon_i = \sum_{j\geq 0} a_j \eta_{i-j}.$$

proposition

(**A20**) ψ is the Huber function that is $\psi(x) = \max[\min(x, c), -c]$, for all x in $\mathbb{R}$, where c is a positive constant.

The M-estimator $\hat{\boldsymbol{\beta}}_\psi(\lambda_k)$ is defined as the solution of

$$\sum_{t=1}^{N} C_t(\lambda_k)\psi(Y_t - C_t'(\lambda_k)\boldsymbol{\beta}) = 0, \tag{4.26}$$

where ψ is defined by

$$\psi(x) = \begin{cases} x, & \text{if } |x| \leq c, \\ c\,\text{sign}(x), & \text{if } |x| > c, \end{cases}$$

and c is some positive constant, see Reisen, Lévy-Leduc & Taqqu (2017). In the following, $c = 1.345$ is adopted to ensure an efficiency of 95% for the regression estimator in Gaussian case.

Similarly to (4.24), the robust M-periodogram is defined by

$$I^{Y}_{M,N}(\lambda_k) = \frac{N}{8\pi}||\hat{\boldsymbol{\beta}}_{\psi}(\lambda_k)||^2 = \frac{N}{8\pi}\left(\hat{\beta}_{1,\psi}(\lambda_k)^2 + \hat{\beta}_{2,\psi}(\lambda_k)^2\right). \tag{4.27}$$

For the univariate context, the asymptotic properties of $\boldsymbol{\beta}_{\psi}$ are established for the short and long-range dependence frameworks in Reisen, Lévy-Leduc & Taqqu (2017), Fajardo et al. (2018) and Reisen, Lévy-Leduc, Cotta, Bondon & Ispany (2019) .

Let now $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ be a sample observation of a bivariate second order stationary time series $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$. In view of (4.24), the cross-periodogram at frequency $\lambda_k = 2\pi k/N, k = 0, \ldots, N-1$. defined by (4.15) may be written as

$$I^{\boldsymbol{Y}}_{N,ij}(\lambda_k) = \begin{cases} \frac{N}{2\pi}\hat{\beta}_{1,Y_i}(\lambda_k)\hat{\beta}_{1,Y_j}(\lambda_k) & \lambda_k = 0 \\ \frac{N}{8\pi}(\hat{\beta}_{1,Y_i}(\lambda_k)\hat{\beta}_{1,Y_j}(\lambda_k) + \hat{\beta}_{2,Y_i}(\lambda_k)\hat{\beta}_{2,Y_j}(\lambda_k) - & \\ \qquad \text{i}(\hat{\beta}_{1,Y_i}(\lambda_k)\hat{\beta}_{2,Y_j}(\lambda_k) - \hat{\beta}_{1,Y_j}(\lambda_k)\hat{\beta}_{2,Y_i}(\lambda_k))) & \lambda_k \neq 0, \quad \text{i,j=1,2,} \end{cases}$$

where $\hat{\beta}_{1,Y_i}(\lambda_k)$ and $\hat{\beta}_{2,Y_i}(\lambda_k)$ are defined by (4.25) and $\{Y_t\}$ is replaced by $\{Y_{it}\}$, $i = 1, 2$.

Likewise, the M-cross-periodogram is defined by

$$I^{\boldsymbol{Y}}_{M,N,ij}(\lambda_k) = \begin{cases} \frac{N}{2\pi}\hat{\beta}_{1,Y_i,\psi}(\lambda_k)\hat{\beta}_{1,Y_j,\psi}(\lambda_k) & \lambda_k = 0 \\ \frac{N}{8\pi}(\hat{\beta}_{1,Y_i,\psi}(\lambda_k)\hat{\beta}_{1,Y_j,\psi}(\lambda_k) + \hat{\beta}_{2,Y_i,\psi}(\lambda_k)\hat{\beta}_{2,Y_j,\psi}(\lambda_k) - & \\ \qquad \text{i}(\hat{\beta}_{1,Y_i,\psi}(\lambda_k)\hat{\beta}_{2,Y_j,\psi}(\lambda_k) - \hat{\beta}_{1,Y_j,\psi}(\lambda_k)\hat{\beta}_{2,Y_i,\psi}(\lambda_k))) & \lambda_k \neq 0, \quad \text{i,j=1,2.} \end{cases}$$

where $\hat{\beta}_{1Y_i,\psi}(\lambda_k)$ and $\hat{\beta}_{2Y_i,\psi}(\lambda_k)$, are defined by (4.26) and $\{Y_t\}$ is replaced by $\{Y_{it}\}$, $i = 1, 2$.

Therefore, the M-periodogram matrix is defined by

$$\boldsymbol{I}^{\boldsymbol{Y}}_{M,N}(\lambda_k) = [I^{\boldsymbol{Y}}_{M,N,ij}(\lambda_k)]^2_{i,j=1} = \begin{bmatrix} I^{\boldsymbol{Y}}_{M,N,11}(\lambda_k) & I^{\boldsymbol{Y}}_{M,N,12}(\lambda_k) \\ I^{\boldsymbol{Y}}_{M,N,21}(\lambda_k) & I^{\boldsymbol{Y}}_{M,N,22}(\lambda_k) \end{bmatrix}. \tag{4.28}$$

As previously mentioned, the main objective of this paper ipropositions to obtain a robust ACOVF which satisfies all the assumptions of the definition of an ACOVF presented in Theorem 11.8.1 (non-negative definiteness) in Brockwell & Davis (2013). In order to obtain such estimator, in (4.13), the elements of $\hat{\text{D}}_N$ matrix are replaced by the robust spectral estimator $\boldsymbol{I}^{\boldsymbol{Y}}_{M,N}(.)$. Thus,

$$\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N} = 2\pi \boldsymbol{H}^{*}_{N}\hat{\text{D}}_{M,N}\boldsymbol{H}_N. \tag{4.29}$$

Statement 2. *Let $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$ be a sample observation of a second order stationary time series $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ satisfying (A11). Let $\boldsymbol{I}^{\boldsymbol{Y}}_{M,N}(.)$ be an estimator of the spectral density matrix of*

$\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ *and define* $\hat{\mathrm{D}}_{M,N}$ *as in* (4.8) *but replacing* $\boldsymbol{f}^{\boldsymbol{Y}}(.)$ *with* $\boldsymbol{I}^{\boldsymbol{Y}}_{M,N}(.)$. *Let* $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N}$ *as in* (4.29). *Suppose that* (*A*11) *to* (*A*20) *hold. Then,*

$$\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N} - 2\pi \boldsymbol{H}^{*}_{N}\mathrm{D}_{N}\boldsymbol{H}_{N} = o_p\left(\frac{1}{\sqrt{N}}\right) \quad as \quad N\to\infty. \tag{4.30}$$

Let $\boldsymbol{U}$ be a $2N\times 2N$,

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{u}_{22} \end{bmatrix}, \tag{4.31}$$

where

$$\boldsymbol{u}_{ii} = \left[(\hat{\gamma}^{\boldsymbol{Y}}_{N,ii,11})^{-1/2}\right] \tag{4.32}$$

and $\boldsymbol{u}_{ii}, i=1,2$, is an $n\times n$ matrix where all entries are the first element of $\boldsymbol{\Gamma}^{\boldsymbol{Y}}_{M,N,ii}$.

The correlation and the robust correlation matrices, namely, $\hat{\boldsymbol{\rho}}^{\star\boldsymbol{Y}}_{N}(h)$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$, are given byproposition

$$\hat{\boldsymbol{\rho}}^{\star\boldsymbol{Y}}_{N} = \boldsymbol{U}\hat{\boldsymbol{\Gamma}}^{\star\boldsymbol{Y}}_{N}\boldsymbol{U} \tag{4.33}$$

and

$$\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N} = \boldsymbol{U}_{M}\hat{\boldsymbol{\Gamma}}^{\star\boldsymbol{Y}}_{M,N}\boldsymbol{U}_{M}. \tag{4.34}$$

where $\boldsymbol{U}_M$ is defined as (4.31) but in (4.32) replacing the entries of $\hat{\boldsymbol{\Gamma}}^{\star\boldsymbol{Y}}_{N}$ with the ones from $\hat{\boldsymbol{\Gamma}}^{\star\boldsymbol{Y}}_{M,N}$.

The lag-h, $h=0,\ldots,N-1$, sample cross-covariance and cross-autocorrelation functions $\hat{\gamma}^{\star\boldsymbol{Y}}_{N,ij}(h)$ and $\hat{\rho}^{\star\boldsymbol{Y}}_{N,ij}(h)$ are, respectively, extracted from the first row of $\hat{\boldsymbol{\Gamma}}^{\star\boldsymbol{Y}}_{N,ij}$ and $\hat{\boldsymbol{\rho}}^{\star\boldsymbol{Y}}_{N,ij}$, for $i,j=1,2$. Their robust counterparts functions, $\hat{\gamma}^{\boldsymbol{Y}}_{M,N,ij}(h)$ and $\hat{\rho}^{\boldsymbol{Y}}_{M,Nij}(h)$ are respectively extracted from the first row of $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N,ij}$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N,ij}$, for $i,j=1,2$. Finally, the lag-h autocovariance and autocorrelation matrices functions in the sense of (4.2) and (4.3), are constructed estimating all (i,j)th elements for $i,j=1,2$. For example, the robust autocovariance and autocorrelation matrices functions are:

$$\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N}(h) = \begin{bmatrix} \hat{\gamma}^{\boldsymbol{Y}}_{M,N,11}(h) & \hat{\gamma}^{\boldsymbol{Y}}_{M,N,12}(h) \\ \hat{\gamma}^{\boldsymbol{Y}}_{M,N,21}(h) & \hat{\gamma}^{\boldsymbol{Y}}_{M,N22}(h) \end{bmatrix} \tag{4.35}$$

and

$$\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h) = \begin{bmatrix} \hat{\rho}^{\boldsymbol{Y}}_{M,N,11}(h) & \hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h) \\ \hat{\rho}^{\boldsymbol{Y}}_{M,N,21}(h) & \hat{\rho}^{\boldsymbol{Y}}_{M,N,22}(h) \end{bmatrix}. \tag{4.36}$$

4 Numerical experiments

The computational experiments were performed using the R programming language R Core Team (2019) and the estimators proposed are available in the *acfMperiod* package Cotta et al. (2019) on The Comprehensive R Archive Network (CRAN-R). In this section, the effect of additive outliers on the estimation of the autocovariance matrix function is investigated for finite sample size time series generated from a Gaussian VAR(1) model. The samples $\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\ldots,\boldsymbol{Y}_n\}$

were generated from a bivariate stationary VAR(1) model defined by the difference equation, $\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\mu} = 0$

$$\boldsymbol{\Phi} = \begin{bmatrix} 0.7 & 0.0 \\ 0.5 & 0.7 \end{bmatrix}$$

and $(\boldsymbol{\varepsilon}_t)$ is a zero-mean Gaussian white noise with covariance

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \begin{bmatrix} 1.00 & 0.0 \\ 0.0 & 1.00 \end{bmatrix}.$$

The sample sizes are $N = 200, 500, 800$ and 1000, and each experiment is replicated 1000 times. In this direction, two different scenarios are simulated: (i) uncontaminated scenario ($p = 0$) and (ii) the samples are contaminated ($p \neq 0$). The samples of $\{\boldsymbol{Z}_t\}$ are the contamination of $\{\boldsymbol{Y}_t\}$ and were contaminated according to (4.20) where, without loss of generality, only the first component $\{Y_{1t}\}$ is contaminated with additive outliers where the parameters are $\omega_1 = 15$ and $p_1 = 0.0, 0.01, 0.05$ and 0.1. For both scenarios, the effect of additive outliers on the sample estimates and the robustness property of the proposed robust estimators are verified by contrasting the empirical means as well as root mean squared errors (RMSE) and biases related to the theoretical values.

We focus on the autocorrelation matrix function since in accordance with the discussion presented in Section 3 , it is affected by additive outliers for all lags, which is in contrast with the covariance matrix function where only $h = 0$ is affected. In addition, the numerical experiments for univariate case, e.g, (i, i)th elements of the autocorrelation or autocovariance matrices functions when $i = 1, 2$, are reported in Cotta, A. Reisen & Bondon (2017). Thus, for simplicity, we present only the case when $i = 1$ and $j = 2$.

Figure 4.1 plots the true value $\rho^{\boldsymbol{Y}}_{N,12}(h)$ and the mean of the estimates $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ for $n = 800$, $p_1 = 0$ and $h = 0, \ldots, 8$. As it can be seen, all three estimators present similar behavior compared to the true values. Therefore, when handling real data and the analyst is uncertain of the presence of outliers, $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ may still be used. The contaminated scenario $p_1 = 0.01$ is shown in Figure 4.2. Now, the non-robust estimators $\hat{\rho}^{\boldsymbol{Z}}_{N,12}(h)$ are $\hat{\rho}^{\star\boldsymbol{Z}}_{N,12}(h)$ completely affected by outliers while the behavior of $\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$ remained slightly unchanged. Therefore, in the scenario of certain occurrence of outliers, the use of $\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$ is recommended.

We now direct our attention towards the RMSE and the BIAS of $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ as n increases for $p_1 = 0$ and $h = 0, \ldots, 8$, presented in Tables 4.1 and 4.2. From Table 4.1, we observe that the samples and the robust estimators have RMSE close to each other in the absence of contamination. As expected, for all lags, $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$ performed best followed by $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ and $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$. As the sample size n increases, we observe a reduction of the RMSE for all estimators. Thus, reinforcing that $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ is useful even in the context which the presence of additive outliers is uncertain.

The bias of the estimates are presented in Table 4.2. We can see that $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$ underestimates the cross-correlation structure while $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$ tends to overestimate it. For the case of $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ the values are overestimated for $n = 200$ and $n = 500$ and underestimated for $n = 800$ and $n = 1000$. Again, all estimators presented good performance.

We now consider the contaminated scenario when $p_1 = 0.1$. The cases when $p_1 = 0.01$ and $p_1 = 0.05$ presented similar conclusions and they are not shown here but available upon request. The RMSE of $\hat{\rho}^{\boldsymbol{Z}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Z}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$ as n increases for $h = 0, \ldots, 8$ is presented in Table 4.3. As one can see, the RMSE of $\hat{\rho}^{\boldsymbol{Z}}_{N,12}(h)$ and $\hat{\rho}^{\star\boldsymbol{Z}}_{N,12}(h)$ presented similar values but are

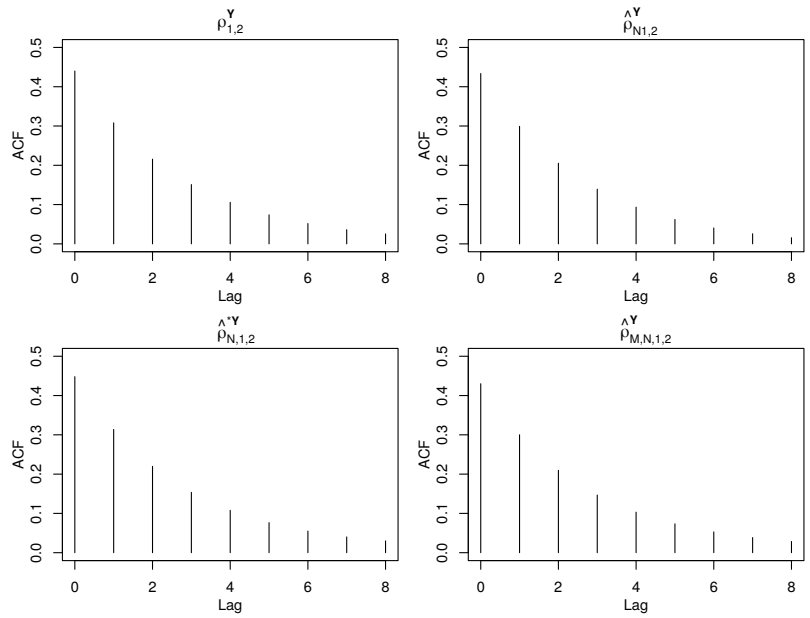


Figure 4.1: Simulation results: $\rho^{\boldsymbol{Y}}_{N,12}(h)$ and the mean of the estimates $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ for $n = 800$, $p_1 = 0$ and $h = 0, \ldots, 8$.

quite different from the ones of $\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$, indicating their destruction while the latter is more robust for 10% of occurrence of outliers.

5 Real data application

In this real data example, we consider the estimation of the sample autocorrelation matrix function of the particulate matter of diameter $\leq 10\mu g/m^3$ measured at Ibes and Vila Velha Centro (VVCentro) stations of the automatic air quality monitoring network (AAQMN) of the Great Vitória Region (GVR) stations from January 2005 to December 2009. This data set has already been considered as a real data example in Ispány et al. (2018) and Souza et al. (2018). Figure 4.3 shows the time series plots of the data. From the plots, it is possible to see the presence of outlying observations which justifies the application of the proposed methodology.

In Figures 4.4 and 4.5, we present the plots of $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$, respectively. Comparing both plots, one may see that vertical scale of both plots are equivalent. This might indicate that the presence of outlying observation is not strong enough in order to completely destroy the correlation structure of the data. However, since $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ behave similarly when the data is outlier free, $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ is used in the sub sequential analysis.

We focus on the estimation of $\boldsymbol{\Phi}$ matrix in (4.19). Thus, consider the following equations:

$$\Sigma_{\boldsymbol{\varepsilon}} = \boldsymbol{\Gamma}^{\boldsymbol{Y}}(0) + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}^{\boldsymbol{Y}}(-j) \tag{4.37}$$

and

$$\boldsymbol{\Gamma}^{\boldsymbol{Y}}(h) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}^{\boldsymbol{Y}}(h-j), \quad h = 1, \ldots, p. \tag{4.38}$$

Replacing $\boldsymbol{\Gamma}^{\boldsymbol{Y}}(h)$ in equations (4.37) and (4.38) by $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N}(h)$ leads to Yule-Walker estimators whose equation system is solved using the Whittle's algorithm. This estimation method might

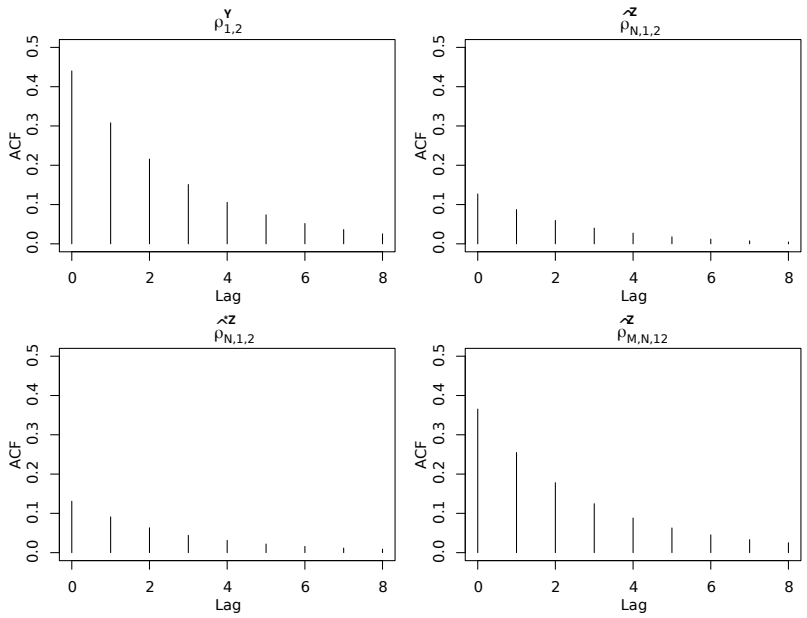


Figure 4.2: Simulation results: $\rho^{\boldsymbol{Y}}_{N,12}(h)$ and the mean of the estimates $\hat{\rho}^{\boldsymbol{Z}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Z}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$ for $n = 800$, $p_1 = 0.01$ and $h = 0, \ldots, 8$.

be affected if there are outliers among the data. Therefore, this estimation procedure may be robustified replacing $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{N}(.)$ by $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{Y}}_{M,N}(h)$ given in (4.35).

Returning to the model analysis, VAR models from first up to the forth order were fitted using the standard and robust Yule-Walker methods, and for parsimonious reasons, the working model is a VAR(1). Figures 4.6 and 4.7 present, respectively, the standard and robust autocorrelation matrix of the fitted VAR(1) model. Comparing the plots from Figures 4.6 and 4.7 with the ones of Figures 4.4 and 4.5, we see that the VAR(1) filters, apart from the seasonality component, were capable to capture the correlation structure of the pollutant series.

In addition, contrasting Figures 4.6 and 4.7, it is possible to observe that the $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ presents similar values as those of $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$. This is an expected result since the plots of Figures 4.4 and 4.5 also presented similar behaviour.

To end this example, we robustly estimated the parameters of the model. The plots of the residuals are shown in Figures 4.8 and 4.9 for $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$, respectively. Comparing the Figures 4.6 and 4.8 it is also noted a similarity between the values of $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$ and $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ which is also expected in view of the similarity between Figures 4.4 and 4.5.

6 Conclusions

The effect of additive outliers on the estimation of the covariance and correlation matrix functions of a stationary multivariate time series was addressed. Robust estimation methods for these matrices were proposed and their performance empirically investigated through Monte Carlo simulation. The numerical experiment results illustrated the good behavior in terms of mean square error of the proposed robust estimators even when the data contain a considerate number of atypical observations. A real data set was analyzed where the proposed robust covariance matrix estimator replaced the standard sample covariance estimator in the Yule-Walker equations.

Table 4.1: RMSE of $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ as n increases for $p_1 = 0$ and $h = 0, \ldots, 8$.

	n	0	1	2	3	4	5	6	7	8
$\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$	200	0.0980	0.1136	0.1225	0.1260	0.1262	0.1281	0.1302	0.1297	0.1302
	500	0.0596	0.0699	0.0764	0.0796	0.0807	0.0812	0.0824	0.0834	0.0845
	800	0.0453	0.0538	0.0595	0.0636	0.0662	0.0672	0.0678	0.0681	0.0684
	1000	0.0409	0.0478	0.0520	0.0548	0.0572	0.0589	0.0599	0.0602	0.0603
$\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$	200	0.1371	0.1477	0.1540	0.1576	0.1591	0.1627	0.1646	0.1652	0.1674
	500	0.0784	0.0865	0.0923	0.0961	0.0976	0.0986	0.1003	0.1015	0.1026
	800	0.0533	0.0611	0.0663	0.0701	0.0729	0.0746	0.0756	0.0763	0.0771
	1000	0.0463	0.0525	0.0564	0.0590	0.0614	0.0631	0.0643	0.0649	0.0652
$\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$	200	0.1328	0.1432	0.1492	0.1529	0.1549	0.1583	0.1602	0.1605	0.1626
	500	0.0766	0.0847	0.0907	0.0943	0.0957	0.0962	0.0970	0.0978	0.0988
	800	0.0547	0.0617	0.0665	0.0696	0.0716	0.0728	0.0733	0.0738	0.0751
	1000	0.0484	0.0534	0.0562	0.0579	0.0599	0.0618	0.0634	0.0638	0.0641

Table 4.2: BIAS of $\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$ as n increases for $p_1 = 0$ and $h = 0, \ldots, 8$.

	n	0	1	2	3	4	5	6	7	8
$\hat{\rho}^{\boldsymbol{Y}}_{N,12}(h)$	200	-0.0213	-0.0286	-0.0338	-0.0367	-0.0367	-0.0367	-0.0379	-0.0372	-0.0359
	500	-0.0050	-0.0077	-0.0096	-0.0092	-0.0093	-0.0095	-0.0104	-0.0110	-0.0129
	800	-0.0065	-0.0087	-0.0104	-0.0117	-0.0122	-0.0118	-0.0112	-0.0104	-0.0097
	1000	-0.0041	-0.0056	-0.0065	-0.0069	-0.0069	-0.0066	-0.0060	-0.0054	-0.0053
$\hat{\rho}^{\star\boldsymbol{Y}}_{N,12}(h)$	200	0.0323	0.0246	0.0193	0.0162	0.0163	0.0162	0.0151	0.0151	0.0162
	500	0.0202	0.0175	0.0156	0.0161	0.0162	0.0159	0.0152	0.0146	0.0127
	800	0.0081	0.0057	0.0041	0.0027	0.0023	0.0026	0.0032	0.0039	0.0047
	1000	0.0077	0.0061	0.0054	0.0050	0.0049	0.0051	0.0056	0.0063	0.0064
$\hat{\rho}^{\boldsymbol{Y}}_{M,N,12}(h)$	200	0.0114	0.0082	0.0074	0.0080	0.0101	0.0114	0.0121	0.0129	0.0152
	500	0.0018	0.0028	0.0047	0.0081	0.0099	0.0114	0.0116	0.0118	0.0103
	800	-0.0099	-0.0079	-0.0058	-0.0042	-0.0024	-0.0004	0.0012	0.0023	0.0032
	1000	-0.0112	-0.0085	-0.0051	-0.0023	0.0001	0.0016	0.0030	0.0043	0.0050

Table 4.3: RMSE of $\hat{\rho}^{\boldsymbol{Z}}_{N,12}(h)$, $\hat{\rho}^{\star\boldsymbol{Z}}_{N,12}(h)$ and $\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$ as n increases for $p_1 = 0.1$ and $h = 0, \ldots, 8$.

	n	0	1	2	3	4	5	6	7	8
$\hat{\rho}^{\boldsymbol{Z}}_{N,12}(h)$	200	0.3594	0.2594	0.1896	0.1415	0.1121	0.0936	0.0839	0.0776	0.0758
	500	0.3497	0.2458	0.1758	0.1278	0.0956	0.0746	0.0634	0.0572	0.0537
	800	0.3507	0.2487	0.1767	0.1274	0.0930	0.0697	0.0557	0.0480	0.0431
	1000	0.3478	0.2451	0.1739	0.1243	0.0908	0.0681	0.0542	0.0451	0.0400
$\hat{\rho}^{\star\boldsymbol{Z}}_{N,12}(h)$	200	0.3507	0.2524	0.1846	0.1391	0.1124	0.0969	0.0894	0.0852	0.0861
	500	0.3449	0.2412	0.1716	0.1243	0.0932	0.0736	0.0638	0.0591	0.0564
	800	0.3479	0.2461	0.1741	0.1252	0.0910	0.0684	0.0550	0.0478	0.0436
	1000	0.3453	0.2427	0.1716	0.1221	0.0888	0.0665	0.0530	0.0442	0.0397
$\hat{\rho}^{\boldsymbol{Z}}_{M,N,12}(h)$	200	0.1987	0.1602	0.1384	0.1233	0.1175	0.1158	0.1152	0.1166	0.1175
	500	0.1814	0.1334	0.1047	0.0871	0.0780	0.0740	0.0721	0.0719	0.0728
	800	0.1849	0.1345	0.1004	0.0803	0.0669	0.0597	0.0562	0.0552	0.0548
	1000	0.1803	0.1300	0.0958	0.0740	0.0609	0.0527	0.0492	0.0478	0.0473

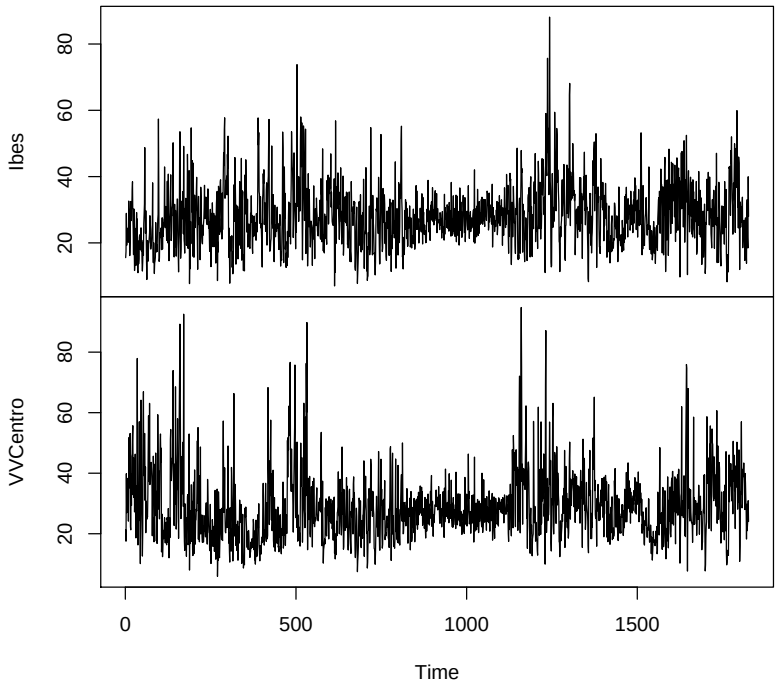


Figure 4.3: PM_{10} concentration measured at Ibes and VVCentro stations.

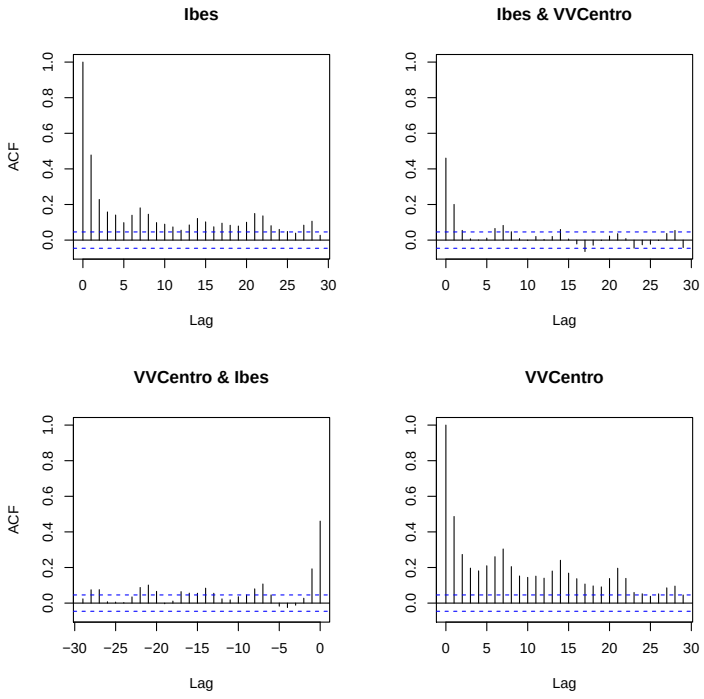


Figure 4.4: $\hat{\boldsymbol{\rho}}_N^{\boldsymbol{Y}}(h)$ of Ibes and VVCentro.

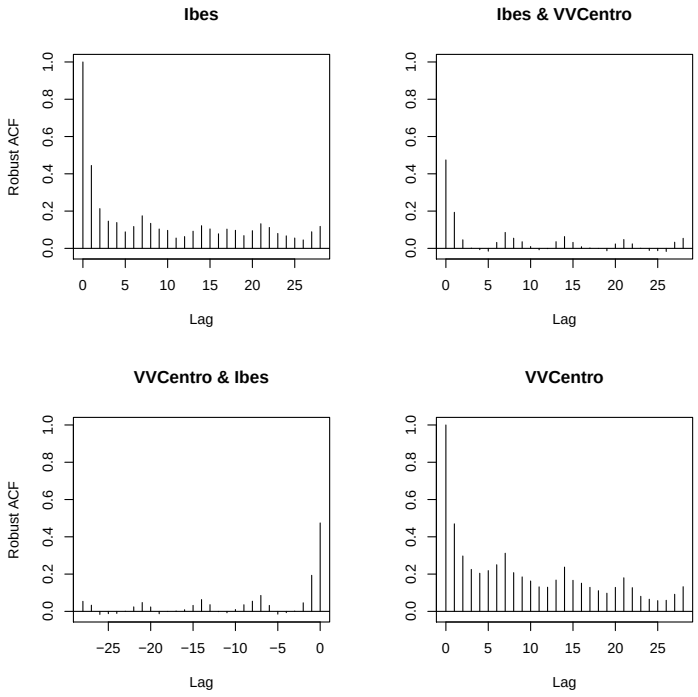


Figure 4.5: $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ of Ibes and VVCentro stations.

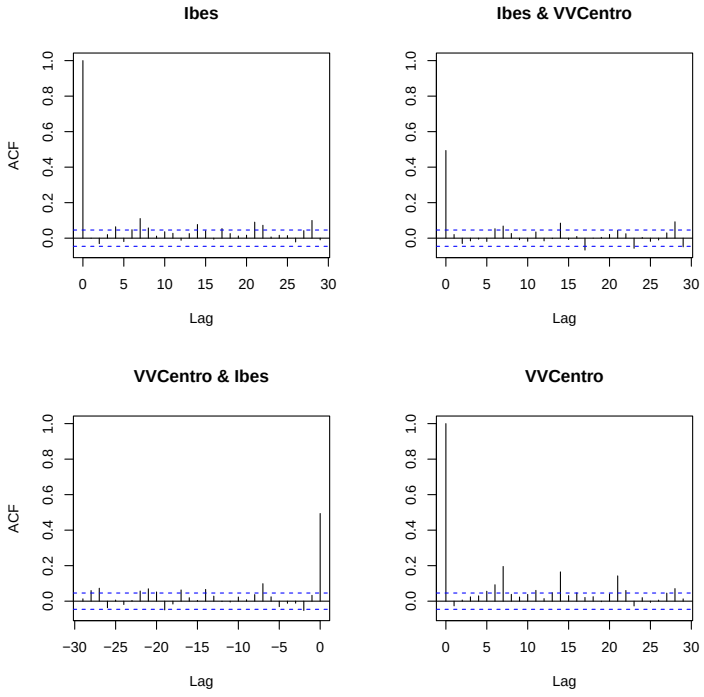


Figure 4.6: $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$ of the residuals the fitted VAR(1) via Yule-Walker.

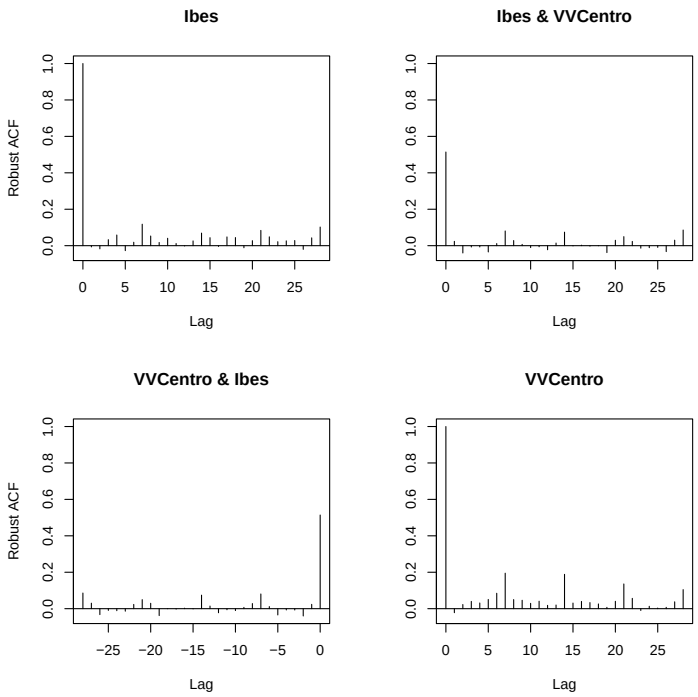


Figure 4.7: $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ of the residuals the fitted VAR(1) via Yule-Walker.

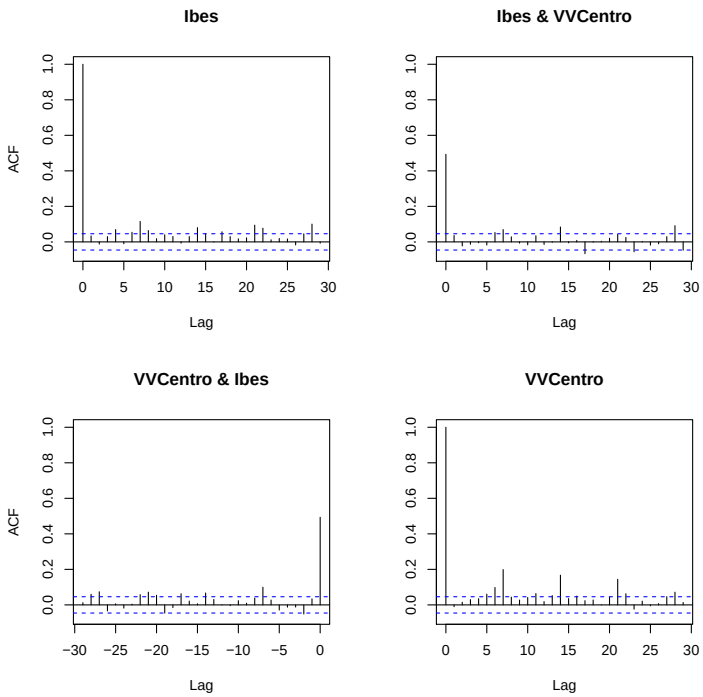


Figure 4.8: $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{N}(h)$ of the residuals of fitted VAR(1) via robustified Yule-Walker.

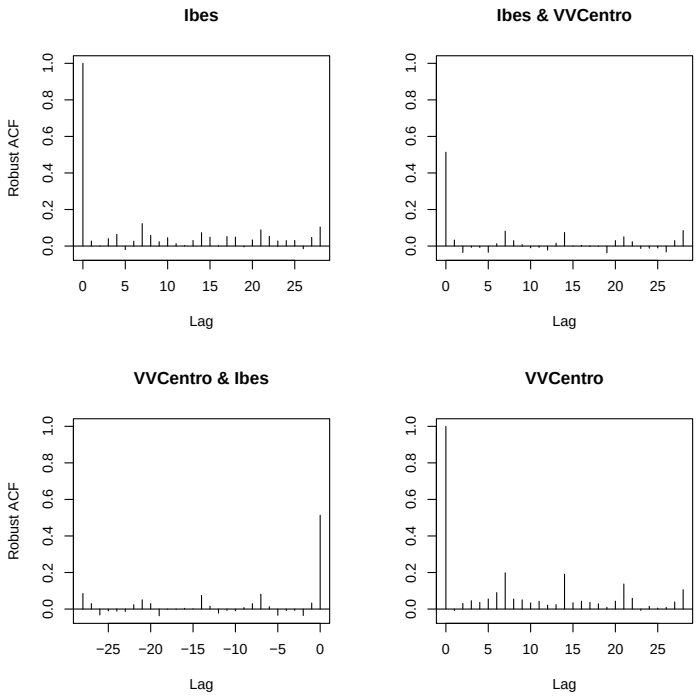


Figure 4.9: $\hat{\boldsymbol{\rho}}^{\boldsymbol{Y}}_{M,N}(h)$ of the residuals of fitted VAR(1) via robustified Yule-Walker.

Chapter 5

Paper 3: A robust method for estimating the number of factors in an approximate factor model

In the third paper, we pose a robust method for estimating the number of factors in approximate factor model. It is empirically demonstrated that additive outliers increase the number of estimated factors and a robust alternative suggested. The methodology consisted in replacing the lag-0 covariance matrix with the lag-0 robust covariance matrix proposed in the second paper. In the application section, we found that with the robust method only 1 factor can be used to explain the dynamic behavior of PM_{10} data.

Abstract

This paper considers the approximate factor model for high-dimensional time series with additive outliers for modeling the dynamic behavior of PM_{10} data measured at air quality monitoring stations of Île-de-France region. We propose a robustification procedure of the information criteria proposed by Bai & Ng (2002). The robust estimator of the number of factors is obtained by replacing the standard covariance matrix with M-covariance matrix. Simulations are carried out under the scenarios of multivariate time series with and without additive outliers to assess the impact of additive outliers on the standard information criteria and to analyze the finite sample size performance of the proposed robust estimator of the number of factors. In the application section, the robust factor analysis is performed to reduce the dimension of the data.

1 Introduction

Nowadays, thanks to the improvements in computer power and data storage capacity, data scientist have now the possibility to work and study high dimensional data sets. As time passes by more data is generated, the dimension increases and so does the number of parameters to be estimated of many statistical models. Therefore, new techniques that accommodate high dimensional data sets are needed. In this context, the factor analysis (FA) is, undoubtedly, one of the most used techniques employed by the analyst for summarizing information while reducing the dimension of a large amount of data.

The FA model assumes that the common factors are latent (not observed) and in order to the model be identifiable some assumptions about the underlying factor structure are required. This fact leads to the development of various factor models. In this direction, a common assumption is that the covariance matrix of the idiosyncratic component is diagonal and this is the starting point of the widely used orthogonal factor model. New factor models are created insofar this

basic assumption is relaxed. The approximate factor model in the sense of Connor & Korajczyk (1986) allows some correlation among the idiosyncratic component.

In this context, one possible approach to estimate the factor model is to assume normality and to use the maximum likelihood estimation or Kalman filter approaches. However, the assumption of normality may be too strong when working with applied data. Another drawback is the number of parameters to be estimated using the Kalman filter approach increases as the more variables are considered.

The approach considered here implements the framework of Bai & Ng (2002), which employs the principal component analysis (PCA) technique to estimate the latent factors. Nevertheless, the PCA tool is the most popular estimation method due to its performance and ease of use. As pointed out by many authors, the PCA method is sensitive to the occurrence of outliers among the collected data Bai & Ng (2017). For example, Reisen, Sgrancio, Lévy-Leduc, Bondon, Monte, Cotta & Ziegelmann (2019) showed that the number of factors is influenced by the presence of additive outliers and proposed the use of a robust autocovariance function estimator to mitigate the effect of additive outliers.

In addition, to be of high dimension due to a large number of variables measured at air pollution monitoring stations scattered over different regions, it is well-known that air pollution data may also present high peaks which may seem as outlying observations. In this scenario, the usual solution is to remove observations that are suspicious to be outliers, but doing so, one is tacitly implying the outliers to be errors which are not the case of most of the high peaks of air pollution time series as they may cause serious harms to the human health and environment.

Nonetheless, these high-level observations can be seen as aberrant values from a statistical point of view. In this direction, these outlying observation can directly affect the statistical properties of the standard estimates such as the sample mean and sample covariance which will affect any sub-sequential method, for example, the FA method making use of the standard PCA technique. In this scenario, the usual solution is to remove observations that are suspicious to be outliers but doing so, one is implicitly assuming the outliers to be errors. Thus, in this paper, robust estimators are proposed for tackling this common issue.

Therefore, this paper considers multivariate time series with additive outliers using the FA technique for dimension reduction where the number of factors is estimated using the criteria of Bai & Ng (2002). In this context, it is here proposed and studied a robust version of the estimators given in Bai & Ng (2002).

The paper is organized as follows: besides the introduction, Section 2 introduces the model and the estimation procedure here considered. Section 3 discuss the impact of additive outliers on the factor model and presents a robust methodology in order to mitigate the effect of the outlying observations. Some Monte Carlo experiments are presented in Section 4 . The application and the concluding remarks are in Sections 5 and 6 , respectively.

2 Model and estimation

Let N, $N \in \mathbb{N}$, denotes the number of variables and T, $T \in \mathbb{N}$, the sample size. For, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, the observation X_{it} is said to a have factor structure if it can be written as

$$X_{it} = \lambda_i' F_t + \epsilon_{it} = C_{it} + \epsilon_{it}, \tag{5.1}$$

where F_t is a vector of common factors, λ_i is a vector of factor loadings associated with F_t, and ϵ_{it} is the idiosyncratic component of X_{it}. C_{it} called the common component of X_{it}.

The model is latent, i.e, the factors and their corresponding loadings, and the idiosyncratic component are not observable. If the factors were to be observable, the model could be easily estimated, for example, by multiple linear regression. Note that the X_{it} has a contemporaneous relationship with F_t, thus (5.1) is referred as the static factor model. This is in contrast with the dynamic factor model, in which X_{it} does not have a contemporaneous relationship with the factors. Dynamic factor models are studied by Stock & Watson (2002) Geweke (1977) and Sargent et al. (1977), among others and are beyond the scope of this paper. In this paper, a robust estimator of the number of factors r, $r \in \mathbb{N}$ is proposed.

Let F_t^0 and λ_i^0 denote the true common factors and their corresponding factor loadings. Thus, (5.1) can be written as an N-dimensional time series with T observations. At a time $t = 1, \ldots, T$,

$$X_t = \Lambda^0 F_t^0 + \epsilon_t, \tag{5.2}$$

where $X_t = (X_{1t}, \ldots, X_{Nt})'$, $\Lambda^0 = (\lambda_1, \ldots, \lambda_N)'$ and $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$.

(5.1) can also be written as a $T-$dimensional vector of random variables. For a given i,

$$X_i = F^0 \lambda_i^0 + \epsilon_i, \tag{5.3}$$

where $X_i = (X_{i1}, \ldots, X_{iT})'$, $F_t^0 = (F_1^0, \ldots, F_T^0)'$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})'$.

Finally, in matrix form,

$$X = F^0 \Lambda^{0\prime} + \epsilon, \tag{5.4}$$

where $X = (X_1, \ldots, X_N)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ are $T \times N$ matrices.

Model in (5.4) has the covariance structure of the static factor model is given by

$$\Sigma_X = \Lambda^0 \Sigma_F \Lambda^{0\prime} + \Sigma_\epsilon, \tag{5.5}$$

where Σ_X, Σ_F and Σ_ϵ are $N \times N$ covariance matrices of X, F^0 and ϵ, respectively.

Let $\mathrm{Tr}(A)$ and $||A|| = (\mathrm{Tr}(A'A))^{1/2}$ denote the trace and the norm of a matrix A, respectively. According to Bai & Ng (2002), in order for the factor model be identifiable the following assumptions are made:

(**A21**) $\mathbb{E}(||F_t^0||^4) < \infty$ and $T^{-1} \sum_{t=1}^T F_t^0 F_t^{0\prime} \to \Sigma_F$ as $T \to \infty$ for some positive definite matrix Σ_F.

(**A22**) $||\lambda_i|| \leq \bar{\lambda} < \infty$, for some positive $\bar{\lambda}$ and $||\Lambda^{0\prime}\Lambda^0/N - D|| \to 0$ as $N \to \infty$ for some $r \times r$ positive definite matrix D.

(**A23**) There exists a positive constant $M < \infty$ such that for all N and T,

1. $\mathbb{E}(\epsilon_{it}) = 0$, $\mathbb{E}(|\epsilon_{it}|^8) \leq M$;
2. $\mathbb{E}(\epsilon_s'\epsilon_t/N) = \mathbb{E}(N^{-1}\sum_{i=1}^N \epsilon_{is}\epsilon_{it}) = \gamma_N(s,t)$,$|\gamma_N(s,s)| \leq M$ for all s, and $T^{-1}\sum_{i=1}^N \sum_{t=1}^T |\gamma_N(s,t)|$ M;
3. $\mathbb{E}(\epsilon_{it}\epsilon_{jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq |\tau_{ij}|$ for some τ_{ij} and for all t, and $N^{-1}\sum_{i=1}^N \sum_{j=1}^N |\tau_{ij}| \leq M$;
4. $\mathbb{E}(\epsilon_{it}\epsilon_{js}) = \tau_{ij,ts}$ and $(NT)^{-1}\sum_{i=1}^N \sum_{j=1} N \sum_{t=1}^T \sum_{s=1}^T |\tau_{ij,ts}| \leq M$;

5. for every (t,s), $\mathbb{E}(|N^{-1/2}\sum_{i=1}^{N}(\epsilon_{is}\epsilon_{it}-\mathbb{E}(\epsilon_{is}\epsilon_{it}))|^4)\leq M$.

(**A24**) $\mathbb{E}(N^{-1}\sum_{i=1}^{N}||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^0\epsilon_{it}||^2)\leq M$.

Under assumptions (A22) to (A24) the factor model here considered is the approximate factor model in the sense of Chamberlain & Rothschild (1982). The approximate factor model in contrast with the standard orthogonal factor model, allows some correlation in the idiosyncratic component.

Many different approaches have been proposed by the literature to estimate the factor model. In a small N setting, one may write the factor model in a state space form, assume normality and use the maximum likelihood approach. However, since the number of parameters increases with N, this approach requires an intensive computational effort. More details can be found in Stock & Watson (1989).

Another possible approach to estimate (5.4) is to consider the least square approach by minimizing the squared sum of the residuals. That is, the estimates of λ and F are obtained by solving the following optimization problem

$$V(\tilde{F},\tilde{\Lambda})=\underset{\Lambda,F}{\operatorname{argmin}}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\tilde{\lambda}_i\tilde{F}_t)^2, \tag{5.6}$$

where $\tilde{F}$ and $\tilde{\lambda}$ are the hypothetical values of the factors and their corresponding loadings. Minimizing (5.6) in respect to $\tilde{F}$ is equivalent to maximizing $\mathrm{Tr}(\tilde{\Lambda}'X'X\tilde{\Lambda})$ subject to $\tilde{\Lambda}'\tilde{\Lambda}/N=I$. In this context, the solution of (5.6) is obtained by setting $\hat{\Lambda}$ equal to the eigenvectors corresponding to the r largest eigenvalues of $X'X$. Thus, the PC estimator of F is

$$\hat{F}=X'\hat{\Lambda}/N. \tag{5.7}$$

A common issue related to the big data context arrives when the number of variables N is much larger than the number of samples T. In this scenario, the rank of $\hat{\Sigma}_X$ is no more than $\min\{N,T\}$. However, as noted by Connor & Korajczyk (1986), one might use the eigenvectors associated with the $T\times T$ XX' matrix. This approach is called asymptotic principal component analysis (APCA) and provides a consistent estimator of the common factors under the following additional assumptions

(**A25**)
1. $\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}\epsilon_{is}\to 0, n\neq s$;
2. $\frac{1}{N}\sum_{i=1}^{N}\epsilon_{it}^2\to\sigma^2$, for all t, as $N\to\infty$.

Thus, concentrating out $\hat{\Lambda}$, minimizing (5.6) in respect to $\tilde{\Lambda}$ is equivalent to maximizing $\mathrm{Tr}(\tilde{F}'X'X\tilde{F})$ subject to $\tilde{F}'\tilde{F}/T=I$. In this context, the solution of (5.6) is setting $\tilde{F}$ equal to the eigenvectors corresponding to the r largest eigenvalues of XX', yielding the APC estimator of F.

As in Bai (2003) and Bai & Ng (2002) this paper considers the setting when both N and $T\to\infty$. The space spanned by $\hat{F}$ and $\tilde{F}$ are equivalent. Therefore, they can be used interchangeably depending on the sizes of N and T to achieve a computationally simpler approach.

Now, the estimation of the number of factors is addressed. Supposing that the factors are observed, Bai & Ng (2002) proposed information criteria for the estimation of the number of factors in approximate factor models. They are

$$
\begin{aligned}
IC_{p1}(k) &= ln(\hat{V}_k) + k(\tfrac{N+T}{NT})\ln(\tfrac{NT}{N+T});\\
IC_{p2}(k) &= ln(\hat{V}_k) + k(\tfrac{N+T}{NT})\ln(\min\{N,T\}^2);\\
IC_{p3}(k) &= ln(\hat{V}_k) + k(\tfrac{\ln(\min\{N,T\}^2)}{\min\{N,T\}^2}),
\end{aligned}
\tag{5.8}
$$

where $\hat{V}_k$ is the minimized value of $V(.)$ in (5.6) and k is a number of estimated factors.

Since outliers are present in the data, any approach that makes use of (5.6) will also be affected. Next session presents the approximate factor model with additive outliers and develops a robust methodology to coherently estimate the number of factors when additive outliers are present.

3 Outliers and robust estimation

It is supposed that the observed process X_t results from the contamination of Z_t by additive random outliers, i.e.,

$$X_t = Z_t + \Omega\delta_t, \tag{5.9}$$

where $\Omega = \text{diag}(\omega_1, ..., \omega_N)$ and $\omega_i, i = 1, ..., N$, is the magnitude of the outliers which affects Z_{it}, $\delta_t = (\delta_{1t}, ..., \delta_{Nt})'$ is a random vector indicating the occurrence of an outlier at time t. It is assumed that X_t and δ_t are uncorrelated processes and that $\mathbb{P}(\delta_{it} = -1) = \mathbb{P}(\delta_{it} = 1) = p_i/2$, $\mathbb{P}(\delta_{it} = 0) = 1 - p_i$ for $i = 1, \ldots, N$ where $0 \leq p_i < 1$. Then $\mathbb{E}(\delta_{it}) = 0$ and $\mathbb{V}ar(\delta_{it}) = p_i$. It is also assumed that $\mathbb{C}ov(\delta_t, \delta_t) = \Sigma_\delta = \text{diag}(p_1, ..., p_N)$ and that $\mathbb{C}ov(\delta_t, \delta_{t+h}) = 0$ when $h \neq 0$.

It follows from (5.9) that the effects of additive outliers on the level of the process is $\mathbb{E}(Z_t) = \mathbb{E}(X_t)$. The effect of additive outliers on the autocovariance function of the process is $\Gamma_Z(0) = \Gamma_X(0) + \Omega\Sigma_\delta\Omega'$, with $\Sigma_X = \Gamma_X(0)$ and $\Sigma_Z = \Gamma_Z(0)$. $\Gamma_Z(h) = \Gamma_X(h)$ when $h \neq 0$.

In view of (5.4), the factor model with additive outliers is

$$Z = F\Lambda' + \epsilon + \Omega\delta. \tag{5.10}$$

As can be seen from (5.10), the outliers that additively influence X are not within the factors. The model under study here is in accord with the outlier models considered by Bai & Ng (2017) and Baragona et al. (2007).

The effect of the additive outliers on the covariance structure of the factor model is

$$\Sigma_Z = \Lambda\Sigma_F\Lambda' + \Sigma_\epsilon + \Omega\Sigma_\delta\Omega'. \tag{5.11}$$

However, it is not possible to decompose and correctly eliminate the occurrence of additive outliers from the observed series Z in a real data scenario. Therefore, the eigenvalues and their corresponding eigenvectors are affected, and, consequently, the number of factors as well the factors themselves. Therefore, in order to mitigate this issue, a robust methodology is here proposed.

Some approaches have been discussed in order to transform the standard factor model robust against additive outliers. From the optimization problem point of view, i.e., context of (5.6), one could replace the least square estimates by some robust alternative, e.g, a different loss function such as least absolute deviation (Kristensen (2014)), singular value threshold (Bai & Ng (2017)

and Fan et al. (2013)) or Huber loss function (Huber (1992)). The latter was considered by Fan et al. (2016) in factor models to perform a robust regression of the data onto the observed covariates before carrying out the PCA estimation procedure.

It is here proposed to robustify $\hat{F}$ and $\tilde{F}$ by replacing the traditional $N \times N$ or $T \times T$ covariance matrices by their corresponding robustified version.

3 .1 Robust estimation of the covariance matrix from the robust M-cross-periodogram

It is known that a given zero-mean stationary univariate time series $X_t, t = 1, \ldots, T$, can be represented as a sum involving T sines and cosines at the Fourier frequencies $\lambda_k = 2\pi k/T, k = 0, \ldots, T-1$. The classical periodogram of X_t at frequency λ_k is

$$I_T^X(\lambda_k) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} X_t \exp(-\mathrm{i}t\lambda_k) \right|^2 .$$

As discussed in Reisen, Lévy-Leduc & Taqqu (2017), one alternative way to derive the periodogram function $I_T^X(.)$ is based on the Least Square (LS) estimates of a bi-dimensional vector $\boldsymbol{\beta}' = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model

$$X_i = c'_{Ti}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)} \cos(i\lambda_j) + \beta^{(2)} \sin(i\lambda_j) + \varepsilon_i \ , \ 1 \leq i \leq T, \ \boldsymbol{\beta} \in \mathbb{R}^2 \ , \tag{5.12}$$

where ε_i denotes the deviation of X_i from $c'_{Ti}\boldsymbol{\beta}$, $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] < \infty$. In the sequel, (ε_i) is assumed to be a function of a stationary Gaussian process.

It is supposed that

$$\varepsilon_i = G(\eta_i), \tag{5.13}$$

where G is a non null real-valued and skew symmetric measurable function (*i.e.* $G(-x) = -G(x)$, for all x) and $(\eta_i)_{i\geq 1}$ is a stationary Gaussian process with zero mean and unit variance. Additional assumptions of $(\eta_i)_{i\geq 1}$ is given in (A9).

It can be shown that

$$I_T^X(\lambda_k) = \frac{T}{8\pi} ||\hat{\boldsymbol{\beta}}(\lambda_k)||^2 = \frac{T}{8\pi} \left(\hat{\beta}_1(\lambda_k)^2 + \hat{\beta}_2(\lambda_k)^2 \right), \tag{5.14}$$

where $\hat{\boldsymbol{\beta}}(\lambda_k)$ is the least squares regression solution

$$\hat{\boldsymbol{\beta}}(\lambda_k) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{t=1}^{T} (X_t - C'_t(\lambda_k)\boldsymbol{\beta})^2, \tag{5.15}$$

with the regressors $C_t(\lambda_k) = [\cos(t\lambda_k), \sin(t\lambda_k)]'$.

The periodogram is robustified by replacing the least squares regression with the M-regression.

The M-estimator $\hat{\boldsymbol{\beta}}_\psi(\lambda_k)$ is defined as the solution of

$$\sum_{t=1}^{T} C_t(\lambda_k)\psi(X_t - C'_t(\lambda_k)\boldsymbol{\beta}) = 0, \tag{5.16}$$

where ψ is defined by

$$\psi(x) = \begin{cases} x, & \text{if } |x| \leq c, \\ c\,\mathrm{sign}(x), & \text{if } |x| > c, \end{cases}$$

and c is some positive constant, see Reisen, Lévy-Leduc & Taqqu (2017). In the following, $c = 1.345$ is adopted to ensure an efficiency of 95% for the regression estimator in Gaussian case.

Similarly to (5.14), the robust M-periodogram is defined by

$$I^X_{M,T}(\lambda_k) = \frac{T}{8\pi}||\hat{\boldsymbol{\beta}}_\psi(\lambda_k)||^2 = \frac{T}{8\pi}\left(\hat{\beta}_{1,\psi}(\lambda_k)^2 + \hat{\beta}_{2,\psi}(\lambda_k)^2\right). \tag{5.17}$$

For the univariate context, the asymptotic properties of $\boldsymbol{\beta}_\psi$ are established for the short and long-range dependence frameworks in Reisen, Lévy-Leduc & Taqqu (2017), Fajardo et al. (2018) and Reisen, Lévy-Leduc, Cotta, Bondon & Ispany (2019) .

Let now $X_1, Y_2, \ldots, X_T$ be a sample observation of a bivariate, $N = 2$, second order stationary time series X_t. The cross-periodogram is defined by

$$I^X_{T,ij}(\lambda_k) = \frac{1}{2\pi}\sum_{h=-(T-1)}^{T-1}\hat{\gamma}^X_{T,ij}(h)\exp(-\mathrm{i}h\lambda_k), \tag{5.18}$$

where, $i, j = 1, 2$, and $\hat{\gamma}^X_{T,ij}(.)$ is the standard sample estimator of the cross-covariance function.

In view of (5.14), the cross-periodogram at frequency $\lambda_k = 2\pi k/T, k = 0, \ldots, T-1$. defined by (5.18) may be written as

$$I^X_{T,ij}(\lambda_k) = \begin{cases} \frac{T}{2\pi}\hat{\beta}_{1,X_i}(\lambda_k)\hat{\beta}_{1,X_j}(\lambda_k) & \lambda_k = 0 \\ \frac{T}{8\pi}(\hat{\beta}_{1,X_i}(\lambda_k)\hat{\beta}_{1,X_j}(\lambda_k) + \hat{\beta}_{2,X_i}(\lambda_k)\hat{\beta}_{2,X_j}(\lambda_k) - & \\ \qquad \mathrm{i}(\hat{\beta}_{1,X_i}(\lambda_k)\hat{\beta}_{2,X_j}(\lambda_k) - \hat{\beta}_{1,X_j}(\lambda_k)\hat{\beta}_{2,X_i}(\lambda_k))) & \lambda_k \neq 0, \quad \text{i,j=1,2,} \end{cases}$$

where $\hat{\beta}_{1,X_i}(\lambda_k)$ and $\hat{\beta}_{2,X_i}(\lambda_k)$ are defined by (5.15) and X_t is replaced by X_{it}, $i = 1, 2$.

Likewise, the M-cross-periodogram is defined by

$$I^X_{M,T,ij}(\lambda_k) = \begin{cases} \frac{T}{2\pi}\hat{\beta}_{1,X_i,\psi}(\lambda_k)\hat{\beta}_{1,X_j,\psi}(\lambda_k) & \lambda_k = 0 \\ \frac{T}{8\pi}(\hat{\beta}_{1,X_i,\psi}(\lambda_k)\hat{\beta}_{1,X_j,\psi}(\lambda_k) + \hat{\beta}_{2,X_i,\psi}(\lambda_k)\hat{\beta}_{2,X_j,\psi}(\lambda_k) - & \\ \qquad \mathrm{i}(\hat{\beta}_{1,X_i,\psi}(\lambda_k)\hat{\beta}_{2,X_j,\psi}(\lambda_k) - \hat{\beta}_{1,X_j,\psi}(\lambda_k)\hat{\beta}_{2,X_i,\psi}(\lambda_k))) & \lambda_k \neq 0, \quad \text{i,j=1,2.} \end{cases}$$

where $\hat{\beta}_{1X_i,\psi}(\lambda_k)$ and $\hat{\beta}_{2X_i,\psi}(\lambda_k)$, are defined by (5.16) and X_t is replaced by X_{it}, $i = 1, 2$.

Therefore, the M-periodogram matrix is defined by

$$I^X_{M,T}(\lambda_k) = [I^X_{M,T,ij}(\lambda_k)]^2_{i,j=1} = \begin{bmatrix} I^X_{M,T,11}(\lambda_k) & I^X_{M,T,12}(\lambda_k) \\ I^X_{M,T,21}(\lambda_k) & I^X_{M,T,22}(\lambda_k) \end{bmatrix}. \tag{5.19}$$

Let Γ^X_T be the covariance matrix of the first T, observations from X_t with absolutely summable autocovariance function and let $f^X(.)$ be its spectral density matrix. Let $\lambda_k = 2\pi k/T, k = 0, \ldots, T-1$, and D_T be an $2T \times 2T$ matrix,

$$D_T = [D_{T,ij}]^2_{i,j=1} = \begin{bmatrix} D_{T,11} & D_{T,12} \\ D_{T,21} & D_{T,22} \end{bmatrix}, \tag{5.20}$$

where

$$D_{T,ij} = \text{diag}[I^X_{M,T,ij}(\lambda_0), I^X_{M,T,ij}(\lambda_1), \ldots, I^X_{M,T,ij}(\lambda_{(T-1)})]. \tag{5.21}$$

Define a transformation matrix H_T by

$$H_T = \begin{bmatrix} G_T & 0 \\ 0 & G_T \end{bmatrix}, \tag{5.22}$$

where G_T is an $T \times T$ matrix with rows given by

$$g_{T,k} = T^{-1/2}[1, e^{-\mathrm{i}\pi k/T}, e^{-\mathrm{i}\pi 2k/T}, \ldots, e^{-\mathrm{i}\pi(T-1)k/T}], k = 0 \ldots, T-1. \tag{5.23}$$

Let H^*_T be the conjugate transpose of H_T. Thus, we robustly estimate Γ^X_T by

$$\hat{\Gamma}^X_{M,T} = 2\pi H^*_T \hat{D}_{M,T} H_T. \tag{5.24}$$

The lag-h, $h = 0, \ldots, N-1$, robust sample cross-covariance function $\hat{\gamma}^X_{M,T,ij}(h)$ is extracted from the first row of $\hat{\Gamma}^X_{M,T,ij}$ for $i, j = 1, 2$. Finally, the lag-h autocovariance and autocorrelation matrices are constructed estimating all (i, j)th elements for $i, j = 1, 2$. Thus, the robust autocovariance matrix function is:

$$\hat{\Gamma}^X_{M,T}(h) = \begin{bmatrix} \hat{\gamma}^X_{M,T,11}(h) & \hat{\gamma}^X_{M,T,12}(h) \\ \hat{\gamma}^X_{M,T,21}(h) & \hat{\gamma}^X_{M,T22}(h) \end{bmatrix} \tag{5.25}$$

It should be noted that the PCA or APCA procedure is calculated from the covariance matrix function at lag $h = 0$.

4 Simulation study

This section reports simulation results related to the performance of the proposed methodology for finite sample size. As in Bai & Ng (2002), the data generating process (DGP) is

$$X_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta}\epsilon_{it}, \tag{5.26}$$

where the factors are $T \times r$ matrices of $N(0, 1)$ random variables. The contaminated data generating process (CDGP) with additive outliers is

$$Z_{it} = X_{it} + \omega\delta_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + \sqrt{\theta}\epsilon_{it} + \omega\delta_{it}. \tag{5.27}$$

For the simulations, $r = 1, 3$ and 5 and the maximum number of factors is 8. $N = 50, 100, 200, 500$ and 1000. $T = 50, 100, 200$ and 500. Two scenarios are considered: (i) the samples are uncontaminated ($p_i = 0, i = 1, \ldots, N$), and (ii) the samples are contaminated ($p_i \neq 0$). When $p_i \neq 0$, $\omega_1 = 15$ and $\omega_i = 0$, $i = 2, \ldots, N$, i.e, the contamination occurs only in the first series of the random vector with the probability of occurrence given in the tables. The reported empirical results are based on 1000 replications. The simulations were performed using the `R` programming language R Core Team (2019).

The first objective of this empirical study is to verify the performance of the three information criteria for estimating the number of factors estimated by APCA method as given in (5.8) under influence of additive outlier model (5.9). In this scenario, the estimated number of factors is expected to increase. The averages of $\hat{r}$ are reported in Tables 5.1, 5.2 and 5.3, for $r = 1, 2$ and 5, respectively.

T	N	$p_i = 0$			$p_i = 0.01$			$p_i = 0.05$		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
50	50	1.00	1.00	4.90	1.21	1.18	5.38	1.70	1.66	6.16
50	100	1.00	1.00	1.00	1.30	1.26	1.38	1.88	1.85	1.93
50	200	1.00	1.00	1.00	1.40	1.38	1.45	1.97	1.96	1.98
50	500	1.00	1.00	1.00	1.49	1.47	1.52	2.00	2.00	2.00
50	1000	1.00	1.00	1.00	1.54	1.53	1.56	2.00	2.00	2.00
100	50	1.00	1.00	1.00	1.16	1.14	1.21	1.60	1.57	1.69
100	100	1.00	1.00	1.00	1.22	1.18	1.38	1.80	1.75	1.92
100	200	1.00	1.00	1.00	1.30	1.26	1.40	1.95	1.93	1.99
100	500	1.00	1.00	1.00	1.35	1.34	1.44	2.00	1.99	2.00
100	1000	1.00	1.00	1.00	1.41	1.40	1.46	2.00	2.00	2.00
200	50	1.00	1.00	1.00	1.10	1.10	1.13	1.46	1.44	1.51
200	100	1.00	1.00	1.00	1.16	1.13	1.24	1.68	1.64	1.81
200	200	1.00	1.00	1.00	1.19	1.15	1.44	1.84	1.79	1.97
200	500	1.00	1.00	1.00	1.28	1.25	1.40	1.99	1.98	2.00
200	1000	1.00	1.00	1.00	1.29	1.27	1.38	2.00	2.00	2.00
500	50	1.00	1.00	1.00	1.03	1.03	1.03	1.21	1.19	1.24
500	100	1.00	1.00	1.00	1.05	1.05	1.06	1.35	1.32	1.45
500	200	1.00	1.00	1.00	1.08	1.07	1.15	1.58	1.53	1.76
500	500	1.00	1.00	1.00	1.11	1.08	1.39	1.88	1.81	2.00
500	1000	1.00	1.00	1.00	1.14	1.12	1.30	1.98	1.97	2.00

Table 5.1: Averages of $\hat{r}$ for $p_i = 0, 0.01$ and 0.05 when $r = 1$.

From Tables 5.1, 5.2 and 5.3, the effect of additive outliers in factor models appears by comparing the estimated number of factors when $p_i = 0$ with the case $p_i \neq 0$. When $p_i = 0$, the results are in accord with the ones in Bai & Ng (2002). When $p_i \neq 0$, as expected, the increment of variability due to the presence of outliers leads to increase the number of estimated factors for all information criteria for the percentage of contamination of 1% and 5%. In general, it is noted that IC2 is less affected than the others.

The second objective is to verify and to compare the performance of the estimated number of factors using the information criteria when the standard APCA method is replaced by the robust methodology suggested in Section 3 . Let $\hat{r}^M$ denote the estimated number of factors considering the robust methodology. The primary interest here is to find out if the robust proposed methodology is competitive in the absence of contamination and if it still provides reliable results in a scenario where the data is contaminated. The results are reported in Tables 5.4, 5.5 and 5.6, for $r = 1, 2$ and 5, respectively.

From Tables 5.4, 5.5 and 5.6, when $p_i = 0$ it is noted that the reported values are in accord with the ones from Tables 5.1, 5.2 and 5.3. This indicates that the proposed robust method may still be considered in a scenario where the occurrence of outliers is uncertain. On the other hand, when there are outliers, i.e, $p_i \neq 0$ in the tables, the results are also close to the ones where $p_i = 0$ of Tables 5.1, 5.2 and 5.3. Thus, in a scenario where there are outliers present in the data, the robust methodology still provides useful results.

Other simulations with different degrees of contamination and data generating process present

T	N	$p_i = 0$			$p_i = 0.01$			$p_i = 0.05$		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
50	50	3.00	3.00	7.24	3.25	3.23	7.39	3.79	3.75	7.67
50	100	3.00	3.00	3.00	3.40	3.38	3.46	3.93	3.92	3.96
50	200	3.00	3.00	3.00	3.58	3.56	3.61	3.99	3.99	3.99
50	500	3.00	3.00	3.00	3.73	3.73	3.76	4.00	4.00	4.00
50	1000	3.00	3.00	3.00	3.85	3.84	3.86	4.00	4.00	4.00
100	50	3.00	3.00	3.00	3.22	3.19	3.25	3.74	3.71	3.80
100	100	3.00	3.00	3.00	3.35	3.32	3.47	3.90	3.88	3.97
100	200	3.00	3.00	3.00	3.46	3.44	3.54	3.98	3.97	4.00
100	500	3.00	3.00	3.00	3.64	3.62	3.69	4.00	4.00	4.00
100	1000	3.00	3.00	3.00	3.73	3.72	3.77	4.00	4.00	4.00
200	50	3.00	3.00	3.00	3.15	3.15	3.17	3.64	3.61	3.68
200	100	3.00	3.00	3.00	3.23	3.21	3.31	3.83	3.80	3.89
200	200	3.00	3.00	3.00	3.37	3.31	3.57	3.95	3.93	3.99
200	500	3.00	3.00	3.00	3.54	3.50	3.68	4.00	4.00	4.00
200	1000	3.00	3.00	3.00	3.61	3.59	3.70	4.00	4.00	4.00
500	50	3.00	3.00	3.00	3.10	3.09	3.10	3.43	3.42	3.45
500	100	3.00	3.00	3.00	3.12	3.12	3.17	3.65	3.63	3.71
500	200	3.00	3.00	3.00	3.20	3.18	3.30	3.83	3.82	3.92
500	500	3.00	3.00	3.00	3.34	3.28	3.66	3.98	3.98	4.00
500	1000	3.00	3.00	3.00	3.48	3.45	3.68	4.00	4.00	4.00

Table 5.2: Averages of $\hat{r}$ for $p_i = 0, 0.01$ and 0.05 when $r = 3$.

similar conclusions and are available upon request. The results presented in this section motive the application of the proposed methodology to a real data problem.

5 Application to PM_{10}

In this application section, 21 ($N = 21$) PM_{10} pollutant time series variables measured at the automatic air quality monitoring network (AAQMN) of Île-de-France (IDF) region are considered. Figure 5.1 presents the geographic localization of each station.

The data are collected hourly and the working series are the daily average from March 17th to June 11th of 2019 (91) days. Figure 5.2 shows the plots of PM_{10} concentrations for the 21 stations. We see that the series present high peaks of pollutant concentrations which can be view, from a statistical point of view, as outlying observations. Thus, the proposed robust information criteria is compared with the standard approach to verify whether these high levels influence the number of estimated factors or not.

For example, the classical and robust autocorrelation functions (ACF) of FR04156 station are displayed in Figure 5.3. Comparing the values of both plots, we see that the robust ACF values are greater than the ones from standard ACF. This indicates that the high levels of PM_{10} at FR04156 station reduce the standard sample ACF estimator values. Therefore, it is expected that the standard and robust FA estimated models might present diverging conclusions.

The estimation of the number of factors was performed accordingly to (5.8) using the standard and robust methodology. For the case of the standard estimator, all information criteria (IC_{p1}, IC_{p2} and IC_{p3}) found $\hat{r} = 3$. On the other hand, all three robust information criteria found $\hat{r}^M = 1$. Figure 5.4 presents $\hat{\boldsymbol{F}}_t$ of $\hat{r} = 3$, (a), (b) and (c), and $\hat{r}^M = 1$, (d).

In this context, we consider only the robust factor to construct the estimated concentrations

T	N	$p_i = 0$			$p_i = 0.01$			$p_i = 0.05$		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
50	50	5.00	5.00	7.93	5.27	5.25	7.95	5.84	5.82	7.98
50	100	5.00	5.00	5.00	5.43	5.42	5.49	5.96	5.95	5.98
50	200	5.00	5.00	5.00	5.63	5.62	5.66	6.00	6.00	6.00
50	500	5.00	5.00	5.00	5.84	5.83	5.86	6.00	6.00	6.00
50	1000	5.00	5.00	5.00	5.93	5.93	5.94	6.00	6.00	6.00
100	50	5.00	5.00	5.00	5.25	5.24	5.28	5.79	5.77	5.83
100	100	5.00	5.00	5.00	5.38	5.35	5.50	5.95	5.92	5.99
100	200	5.00	5.00	5.00	5.55	5.53	5.63	6.00	5.99	6.00
100	500	5.00	5.00	5.00	5.76	5.76	5.80	6.00	6.00	6.00
100	1000	5.00	5.00	5.00	5.87	5.86	5.90	6.00	6.00	6.00
200	50	5.00	5.00	5.00	5.21	5.20	5.23	5.67	5.66	5.71
200	100	5.00	5.00	5.00	5.30	5.29	5.38	5.89	5.88	5.94
200	200	5.00	5.00	5.00	5.47	5.43	5.63	5.98	5.97	6.00
200	500	5.00	5.00	5.00	5.69	5.67	5.79	6.00	6.00	6.00
200	1000	5.00	5.00	5.00	5.81	5.80	5.87	6.00	6.00	6.00
500	50	5.00	5.00	5.00	5.13	5.13	5.14	5.56	5.56	5.59
500	100	5.00	5.00	5.00	5.20	5.19	5.23	5.79	5.77	5.82
500	200	5.00	5.00	5.00	5.32	5.30	5.41	5.93	5.92	5.97
500	500	5.00	5.00	5.00	5.51	5.46	5.76	6.00	6.00	6.00
500	1000	5.00	5.00	5.00	5.69	5.66	5.84	6.00	6.00	6.00

Table 5.3: Averages of $\hat{r}$ for $p_i = 0, 0.01$ and 0.05 when $r = 5$.

of the 21 stations. Thus, Figure 5.5, presents the observed series (solid line) and estimated one (dashed line), by considering the linear combination of the only estimated robust factor of FR04156 (a) and FR04329 (b) stations. We observe that the measured data and the estimated one are alike, including the high volatility and large peaks periods of PM_{10} concentrations. Therefore, the estimated robust factor can be considered for forecasting in a context of smaller dimension. See, for instance, Stock & Watson (2002).

6 Conclusions

In this paper, a robust FA method for high-dimensional with additive outliers is proposed. The simulations show that additive outliers increase the number of factors estimated by the standard information criteria. The information criteria applied to the robustified estimation method presents better performance and is an alternative method when there is any evidence of atypical observations in the multivariate time series data. The proposed methodology was used to identify the number of factors of 21 PM_{10} pollutant time series obtained at the stations of Île-de-France region. It was found that a total of 1 factor may be used to summarize the information of all time series.

T	N	$p_i = 0$			$p_i = 0.01$			$p_i = 0.05$		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
50	50	1.00	1.00	1.02	1.01	1.01	1.04	1.07	1.05	1.27
50	100	1.00	1.00	1.00	1.00	1.00	1.01	1.06	1.04	1.15
50	200	1.00	1.00	1.00	1.00	1.00	1.00	1.05	1.04	1.10
50	500	1.00	1.00	1.00	1.00	1.00	1.00	1.08	1.07	1.09
50	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.18	1.18	1.19
100	50	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.01	1.03
100	100	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.01	1.08
100	200	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.05
100	500	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02
100	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02
200	50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01
200	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
200	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01
200	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
200	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	200	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 5.4: Averages of $\hat{r}^M$ for $p_i = 0, 0.01$ and 0.05 when $r = 1$.

T	N	$p_i = 0$			$p_i = 0.01$			$p_i = 0.05$		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
50	50	3.00	3.00	3.02	3.01	3.01	3.04	3.07	3.05	3.07
50	100	3.00	3.00	3.00	3.00	3.00	3.01	3.06	3.04	3.05
50	200	3.00	3.00	3.00	3.00	3.00	3.00	3.05	3.04	3.07
50	500	3.00	3.00	3.00	3.00	3.00	3.00	3.08	3.07	3.09
50	1000	3.00	3.00	3.00	3.00	3.00	3.00	3.08	3.07	3.09
100	50	3.00	3.00	3.00	3.00	3.00	3.00	3.02	3.01	3.03
100	100	3.00	3.00	3.00	3.00	3.00	3.00	3.02	3.01	3.08
100	200	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.01	3.05
100	500	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.01	3.02
100	1000	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.01	3.02
200	50	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.01
200	100	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.00	3.00
200	200	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.01
200	500	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
200	1000	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
500	50	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
500	100	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
500	200	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
500	500	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
500	1000	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00

Table 5.5: Averages of $\hat{r}^M$ for $p_i = 0, 0.01$ and 0.05 when $r = 3$.

T	N	$p_i = 0$			$p_i = 0.01$			$p_i = 0.05$		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
50	50	5.00	5.00	5.02	5.01	5.01	5.04	5.07	5.05	5.23
50	100	5.00	5.00	5.00	5.00	5.00	5.01	5.06	5.04	5.05
50	200	5.00	5.00	5.00	5.00	5.00	5.00	5.05	5.04	5.10
50	500	5.00	5.00	5.00	5.00	5.00	5.00	5.08	5.07	5.10
50	1000	5.00	5.00	5.00	5.00	5.00	5.00	5.18	5.08	5.89
100	50	5.00	5.00	5.00	5.00	5.00	5.00	5.02	5.01	5.03
100	100	5.00	5.00	5.00	5.00	5.00	5.00	5.02	5.01	5.08
100	200	5.00	5.00	5.00	5.00	5.00	5.00	5.01	5.01	5.05
100	500	5.00	5.00	5.00	5.00	5.00	5.00	5.01	5.01	5.02
100	1000	5.00	5.00	5.00	5.00	5.00	5.00	5.01	5.01	5.02
200	50	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.01
200	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
200	200	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.01
200	500	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
200	1000	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	50	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	100	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	200	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	500	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
500	1000	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00

Table 5.6: Averages of $\hat{r}^M$ for $p_i = 0, 0.01$ and 0.05 when $r = 5$.

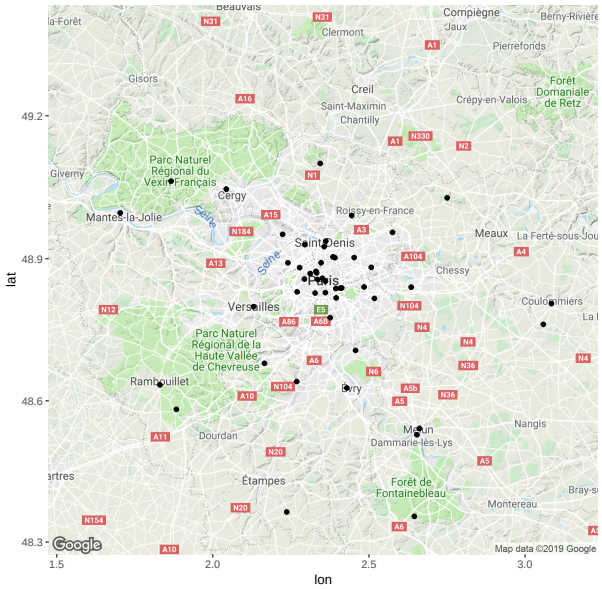


Figure 5.1: Geographical location of the stations from IDF.

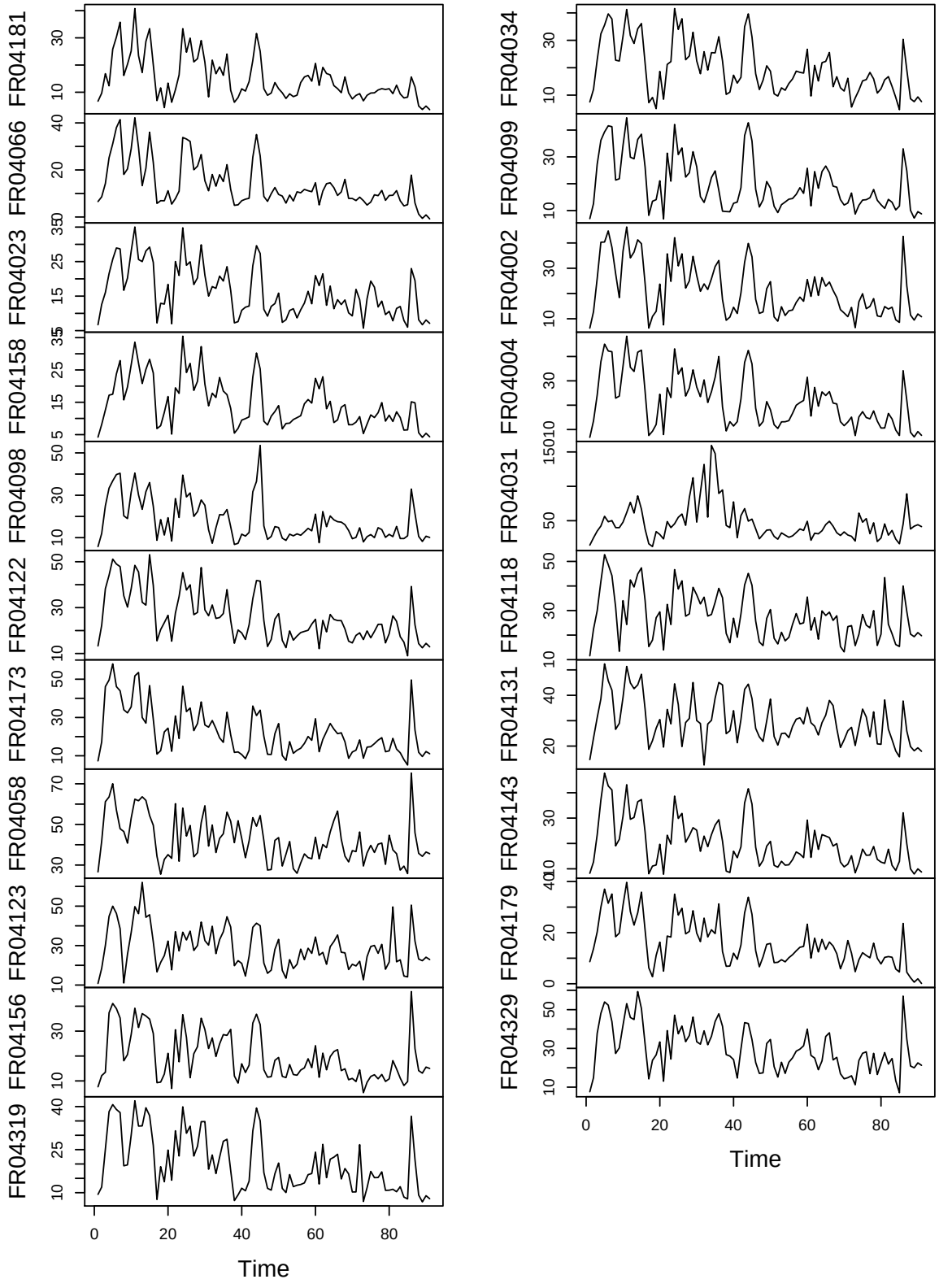


Figure 5.2: Plots of the PM_{10} pollutant concentrations of the 21 stations of the AAQMN of IDF ($N = 21$).

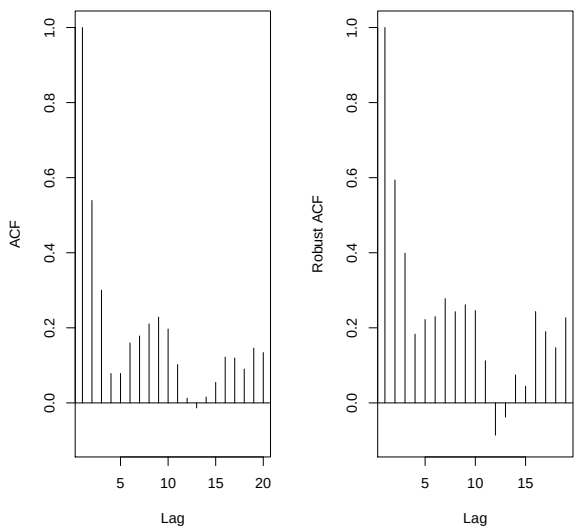
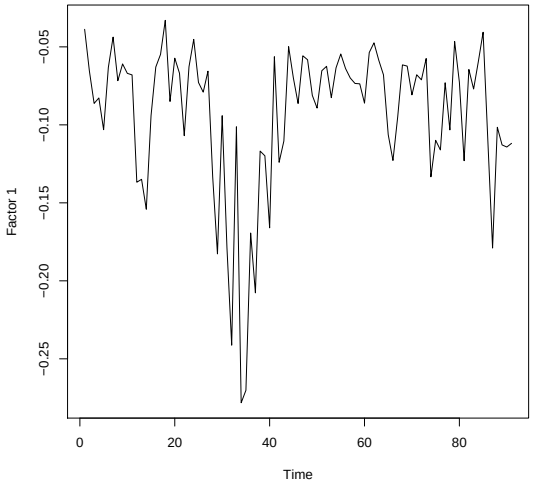
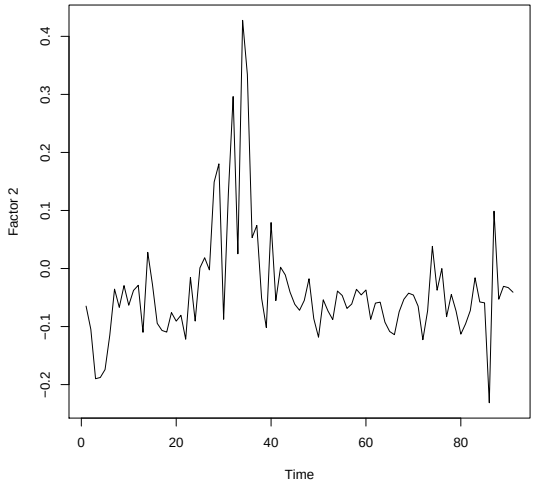


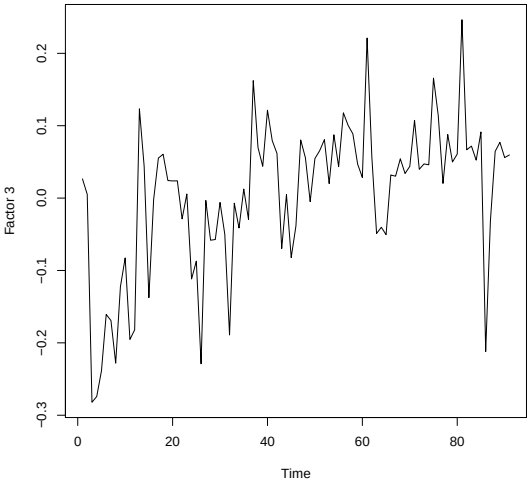
Figure 5.3: Classical and robust autocorrelation functions of FR04156 station.



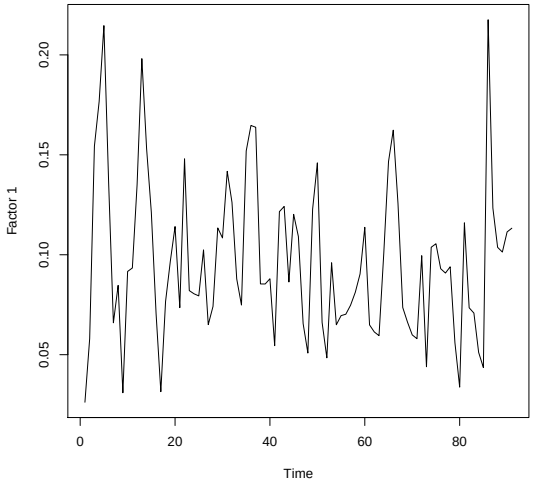
(a) Estimated $\hat{\boldsymbol{F}}_{1,t}$, standard approach.



(b) Estimated $\hat{\boldsymbol{F}}_{2,t}$, standard approach.

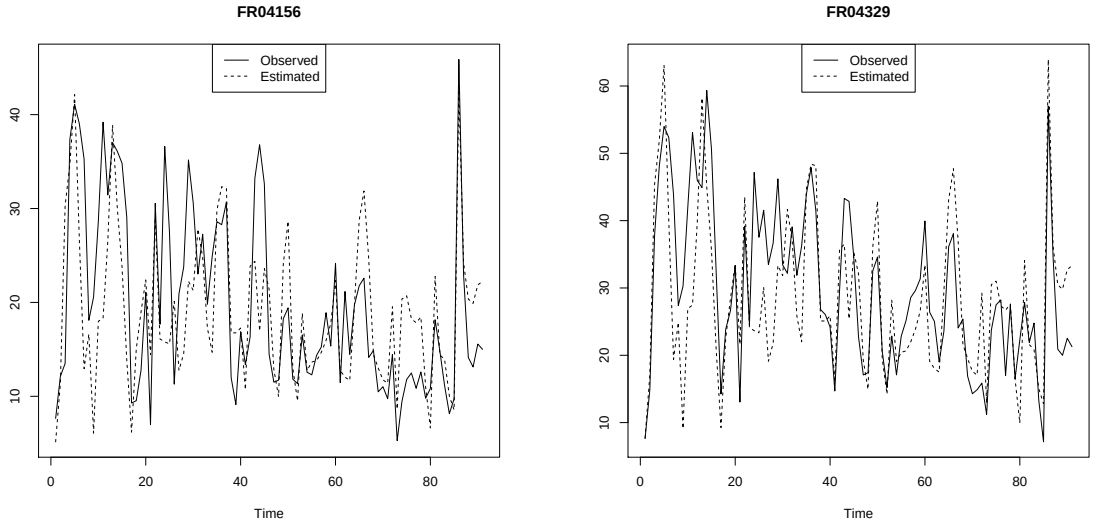


(c) Estimated $\hat{\boldsymbol{F}}_{3,t}$, standard approach.



(d) Estimated $\hat{\boldsymbol{F}}_{1,t}$, robust approach.

Figure 5.4: Time series plots of the three estimated factors by means of standard method, (a), (b) and (c), respectively. Time series plots of the only estimated factor considering the robust approach (d).



(a) FR04156 station.

(b) FR04329 station.

Figure 5.5: Time series plots of observed series and estimated one, solid and dashed lines, respectively, of FR04156 (a) and FR04329 (b) stations.

Chapter 6

Conclusions

1 General conclusions

This thesis considered the study of additive outliers in multivariate stationary time series. The scientific contribution of this thesis is presented in three papers.

The first paper dealt with the robust estimation of the autocovariance and autocorrelation functions of stationary univariate time series. The robust methodology is achieving by replacing the periodogram with the robust M-periodogram. The M-periodogram is obtained by replacing the standard least square estimates by the robust M-estimated in the computation of the Fourier coefficients of the periodogram. The proposed model details were presented and the consistency and asymptotic normality obtained. Then a comprehensive simulation study with finite samples was conducted providing good robust results. Thus, this proposed methodology is useful when dealing with additive outliers in stationary univariate time series processes.

The good performance and the results obtained by the robust estimation methodology proposed by the first paper motivated it extension to stationary multivariate time series processes. In this paper, the cross-periodogram was replaced with the robust M-cross-periodogram. The M-cross-periodogram is obtained as in the first paper. That is, from the robust Fourier coefficients from the univariate periodograms. The consistency and asymptotically normality of the estimators was also studied. Their performance are also investigated by means of numerical experiments and a real data set measured at the stations of the Automatic Air Quality Monitoring Network (AAQMN) of Greater Vitória was analyzed. The robust methodology provided a better fit to the data with additive outliers.

In the third paper, it was proposed a robust method for estimating the number of factors in approximate factor model. It was empirically demonstrated that additive outliers increase the number of estimated factors and a robust alternative was suggested. The methodology consisted in replacing the lag-0 covariance matrix with the lag-0 robust covariance matrix proposed in the second paper. The dynamic behavior of PM_{10} pollutant data measured at the stations of Île-de-France region was studied. It was found that only one factor can summarize the pollution behavior of the 21 stations.

The computer codes developed for this thesis are grouped in the package named **acfMPeriod** readily available in the CRAN-R repository.

2 Perspectives

Some suggestion of future investigation lines are:

- In the computation of M-periodogram, it could be interesting to replace the Huber's loss function with different loss functions and compare their performance under different scenarios. This is the case for the univariate and multivariate papers.
- Related to the estimation of the factor model, it could be of interest to replace the squared loss function by some robust alternative, i.e, Huber's loss function. This approach will also lead to the robustification of the estimation method.
- Another investigation line, is to consider the dynamic factor model where the static factor relationship is relaxed.
- Still related to factor models, one could use the suggested methodology with other pollutant variables.
- One might be interested to consider the proposed estimators with other statistical tools such as principal component analysis, cluster analysis, canonical correlation analysis, among others.

References

Bai, J. (2003), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**(1), 135–171.

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bai, J. & Ng, S. (2017), 'Principal components and regularized estimation of factor models', *arXiv preprint arXiv:1708.08137* .

Baragona, R., Battaglia, F. et al. (2007), 'Outliers in dynamic factor models', *Electronic Journal of Statistics* **1**, 392–432.

Brockwell, P. J. & Davis, R. A. (2013), *Time series: theory and methods*, Springer Science & Business Media.

Chamberlain, G. & Rothschild, M. (1982), 'Arbitrage, factor structure, and mean-variance analysis on large asset markets'.

Chan, W. (1992), 'A note on time series model specification in the presence of outliers', *Journal of Applied Statistics* **19**(1), 117–124.

Chan, W. (1995), 'Outliers and financial time series modelling: a cautionary note', *Mathematics and Computers in Simulation* **39**(3), 425–430.

Chang, C. C. & Politis, D. N. (2016), 'Robust autocorrelation estimation', *Journal of Computational and Graphical Statistics* **25**(1), 144–166.

Connor, G. & Korajczyk, R. A. (1986), 'Performance measurement with the arbitrage pricing theory: A new framework for analysis', *Journal of financial economics* **15**(3), 373–394.

Cotta, H., A. Reisen, V. & Bondon, P. (2017), Robust autocovariance estimation from the frequency domain, *in* 'International Work-Conference on Time Series', Granada, Spain, pp. 1073–1074.

Cotta, H. H. A. (2014), Análise de componentes principais robusta em dados de poluição do ar: aplicação à otimização de uma rede de monitoramento, Master's thesis, Programa de Pós-graduação em Engenharia Ambiental - Universidade Federal do Espírito Santo., Vitória.

Cotta, H., Reisen, V., Bondon, P. & Lévy-Leduc, C. (2019), *acfMPeriod: Applications of the M-periodogram and M-cross-periodogram to the robust estimation of the autocovariance and autocorrelation functions(Univariate and Multivariate)*. R package version 1.0.0.

Cotta, H., Reisen, V., Bondon, P. & Stummer, W. (2017), Robust estimation of covariance and correlation functions of a stationary multivariate process, *in* 'Proc. Int. Conf. Time Series, Granada'.

Dürre, A., Fried, R. & Liboschik, T. (2015), 'Robust estimation of (partial) autocorrelation', *Wiley Interdisciplinary Reviews: Computational Statistics* **7**(3), 205–222.

Fajardo, F., Reisen, V. A., Lévy-Leduc, C. & Taqqu, M. (2018), 'M-periodogram for the analysis of long-range-dependent time series', *Statistics* **52**(3), 665–683.

Fan, J., Ke, Y. & Liao, Y. (2016), 'Robust factor models with covariates'.

Fan, J., Liao, Y. & Mincheva, M. (2013), 'Large covariance estimation by thresholding principal orthogonal complements', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4), 603–680.

Fox, A. J. (1972), 'Outliers in time series', *Journal of the Royal Statistical Society* **34**, 350–363.

Fuller, W. A. (1996), *Introduction to statistical time series*, Vol. 428, John Wiley & Sons.

Geweke, J. (1977), 'The dynamic factor analysis of economic time series', *Latent variables in socio-economic models* .

Huber, P. J. (1964), 'Robust estimation of a location parameter', *Annals of Mathematical Statistics* **35**(1), 73–101.

Huber, P. J. (1992), Robust estimation of a location parameter, *in* 'Breakthroughs in statistics', Springer, pp. 492–518.

Huber, P. & Ronchetti, E. (2009), *Robust Statistics*, Wiley.

IAU (2018), *Chiffres-clés de la région Île-de-France*. Available at http://www.cci-paris-idf.fr/sites/default/files//crocis/wysiwyg/Chiffres-cles-2018derlight.pdf.

Ispány, M., Reisen, V. A., Franco, G. C., Bondon, P., Cotta, H. H. A., Filho, P. R. P. & Serpa, F. S. (2018), On generalized additive models with dependent time series covariates, *in* I. Rojas, H. Pomares & O. Valenzuela, eds, 'Time Series Analysis and Forecasting', Springer International Publishing, Cham, pp. 289–308.

Koul, H. L. & Surgailis, D. (2000), 'Second order behavior of M-estimators in linear regression with long-memory errors', *Journal of Statistical Planning and Inference* **91**, 399–412.

Kristensen, J. T. (2014), 'Factor-based forecasting in the presence of outliers: Are factors better selected and estimated by the median than by the mean?', *Studies in Nonlinear Dynamics & Econometrics* **18**(3), 309–338.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. & Reisen, V. A. (2011*a*), 'Large sample behavior of some well-known robust estimators under long-range', *Statistics* **45**(1), 59–71.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M. S. & Reisen, V. A. (2011*b*), 'Robust estimation of the scale and of the autocovariance function of Gaussian short-and long-range dependent processes', *Journal of Time Series Analysis* **32**(2), 135–156.

Li, T.-H. (2008), 'Laplace periodogram for time series analysis', *Journal of the American Statistical Association* **103**(482), 757–768.

Li, T.-H. (2010), 'A nonlinear method for robust spectral analysis', *IEEE Transactions on Signal Processing* **58**(5), 2466–2474.

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

Maronna, R. A. (1976), 'Robust M-estimators of multivariate location and scatter', *Annals of Statistics* **4**, 51–67.

Maronna, R., Martin, R. D. & Yohai, V. (2006), *Robust Statistics*, Wiley.

Molinares, F. F., Reisen, V. A. & Cribari-Neto, F. (2009), 'Robust estimation in long-memory processes under additive outliers', *Journal of Statistical Planning and Inference* **139**(8), 2511–2525.

Priestley, M. B. (1981), *Spectral analysis and time series*, Vol. 1, Academic press London.

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

Reisen, V. A., Lévy-Leduc, C. & Taqqu, M. S. (2017), 'An M-estimator for the long-memory parameter', *Journal of Statistical Planning and Inference* **187**, 44 – 55.

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

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

Reisen, V., Lévy-Leduc, C., Cotta, H., Bondon, P. & Ispany, M. (2019), 'An overview of robust spectral estimators'.

Reisen, V., Lévy-Leduc, C. & Taqqu, M. (2017), 'An M-estimator for the long-memory parameter', *Journal of Statistical Planning and Inference* **187**, 44 – 55.

Rousseeuw, P. J. & Croux, C. (1993), 'Alternatives to the median absolute deviation', *Journal of the American Statistical Association* **88**(424), 1273–1283.

Rousseeuw, P. & Van Zomeren, B. (1990), 'Unmasking multivariate outliers and leverage points', *Journal of the American Statistical Association* **85**, 633–639.

Sargent, T. J., Sims, C. A. et al. (1977), 'Business cycle modeling without pretending to have too much a priori economic theory', *New methods in business cycle research* **1**, 145–168.

Sarnaglia, A. J. Q., Reisen, V. A., Bondon, P. & Lévy-Leduc, C. (2016), A robust estimation approach for fitting a PARMA model to real data, *in* '2016 IEEE Statistical Signal Processing Workshop (SSP)'.

Seber, G. A. (2008), *A matrix handbook for statisticians*, Vol. 15, John Wiley & Sons.

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

Stock, J. H. & Watson, M. W. (1989), 'New indexes of coincident and leading economic indicators', *NBER macroeconomics annual* **4**, 351–394.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American statistical association* **97**(460), 1167–1179.

Tsay, R. S., Peña, D. & Pankratz, A. E. (2000), 'Outliers in multivariate time series', *Biometrika* **87**(4), 789–804.

Tukey, J. (1975), Useable resistant/robust techniques of analysis, *in* 'Proceedings of the first ERDA Symposium'.

Zhang, Z. & Chan, S.-C. (2005), Robust adaptive lomb periodogram for time-frequency analysis of signals with sinusoidal and transient components, *in* 'Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on', Vol. 4, IEEE, pp. iv–493.

Zoubir, A. M., Koivunen, V., Chakhchoukh, Y. & Muma, M. (2012), 'Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts', *IEEE Signal Processing Magazine* **29**(4), 61–80.

A Linear Algebra

A.1 Definitions

This Appendix presents some matrix definition used in this thesis.

Definition 1. *An $N \times N$ real or complex matrix $\boldsymbol{A}$ is a (regular) circulant if it has the form*

$$\boldsymbol{A} = \begin{bmatrix} a_0 & a_1 & a_2 & \dots & a_{N-1} \\ a_{N-1} & a_0 & a_1 & \dots & a_{N-1} \\ a_{N-2} & a_{N-1} & a_0 & \dots & a_{N-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_1 & a_2 & a_3 & \dots & a_0 \end{bmatrix}. \tag{1}$$

The notation $\boldsymbol{A} = circ[a_0, a_1, \dots, a_{N-1}]$ is also used since only the first row is necessary to define a circulant matrix.

Definition 2. *An $N \times N$ real or complex matrix $\boldsymbol{A}$ is a (regular) symmetric circulant if it has the form*

$$\boldsymbol{A} = \begin{bmatrix} a_0 & a_1 & a_2 & \dots & a_2 & a_1 \\ a_1 & a_0 & a_1 & \dots & a_3 & a_2 \\ a_2 & a_1 & a_0 & \dots & a_4 & a_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_1 & a_2 & a_3 & \dots & a_1 & a_0 \end{bmatrix}. \tag{2}$$

The notation $\boldsymbol{A} = circ[a_0, a_1, \dots, a_2, a_1]$ is also used.

Definition 3. *Let $a_1, a_2 \dots, a_N$ be a set of real numbers, V and V' are called $N \times N$ Vandermonde matrices if they have the form*

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ a_1 & a_2 & a_3 & \dots & a_N \\ a_1^2 & a_2^2 & a_3^2 & \dots & a_N^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_1^{N-1} & a_2^{N-1} & a_3^{N-1} & \dots & a_N^{N-1} \end{bmatrix}. \tag{3}$$

Definition 4. *Let $\omega = \exp^{2\pi \mathrm{i}/N} = \cos(2\pi/N) + \mathrm{i}\sin(2\pi/N)$, where $i = \sqrt{-1}$, so that $\bar{\omega} = \exp^{-2\pi \mathrm{i}/N}$. Also $\omega^r = \cos(2\pi r/N) + \mathrm{i}\sin(2\pi r/N)$. Then, the Fourier matrix $\boldsymbol{F}$ is defined by*

$$\boldsymbol{F} = N^{-1/2}\boldsymbol{V}(1, \bar{\omega}, \bar{\omega}^2, \dots, \bar{\omega}^{N-1}). \tag{4}$$

$$\boldsymbol{F}^\star = \bar{\boldsymbol{F}} = (N)^{-1/2} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \dots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \dots & \omega^{2N-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \omega^{N-1} & \omega^{2N-2} & \dots & \omega^{(N-1)(N-1)} \end{bmatrix}. \tag{5}$$

B Co-authored papers

In this Appendix, it is displayed papers that I have made contributions during my Ph.D. period. Note that these papers are directly connected, from a theoretical and empirical point of views, with the thesis. I take this opportunity to thanks my supervisors Professor Valderio Reisen and Pascal Bondon to invite me to join these additional and interesting works which in some way helped me in many directions of my thesis. The papers are:

1. Robust factor modelling for high-dimensional time series: An application to air pollution data. Applied Mathematics and Computation, 2019.
2. On generalized additive models with dependent time series covariates- Time Series and Forecasting: Contributions to Statistics. Springer series, 2018.
3. Principal component analysis with autocorrelated data. Submitted.
4. An overview of robust spectral estimators Time Series and Cyclostationary process. Springer series, 2019.

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, Espíto 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, $\boldsymbol{\varepsilon}_t \sim WN(\boldsymbol{0}, \boldsymbol{\Sigma}_\varepsilon)$, $\boldsymbol{X}_t$ 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

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

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.

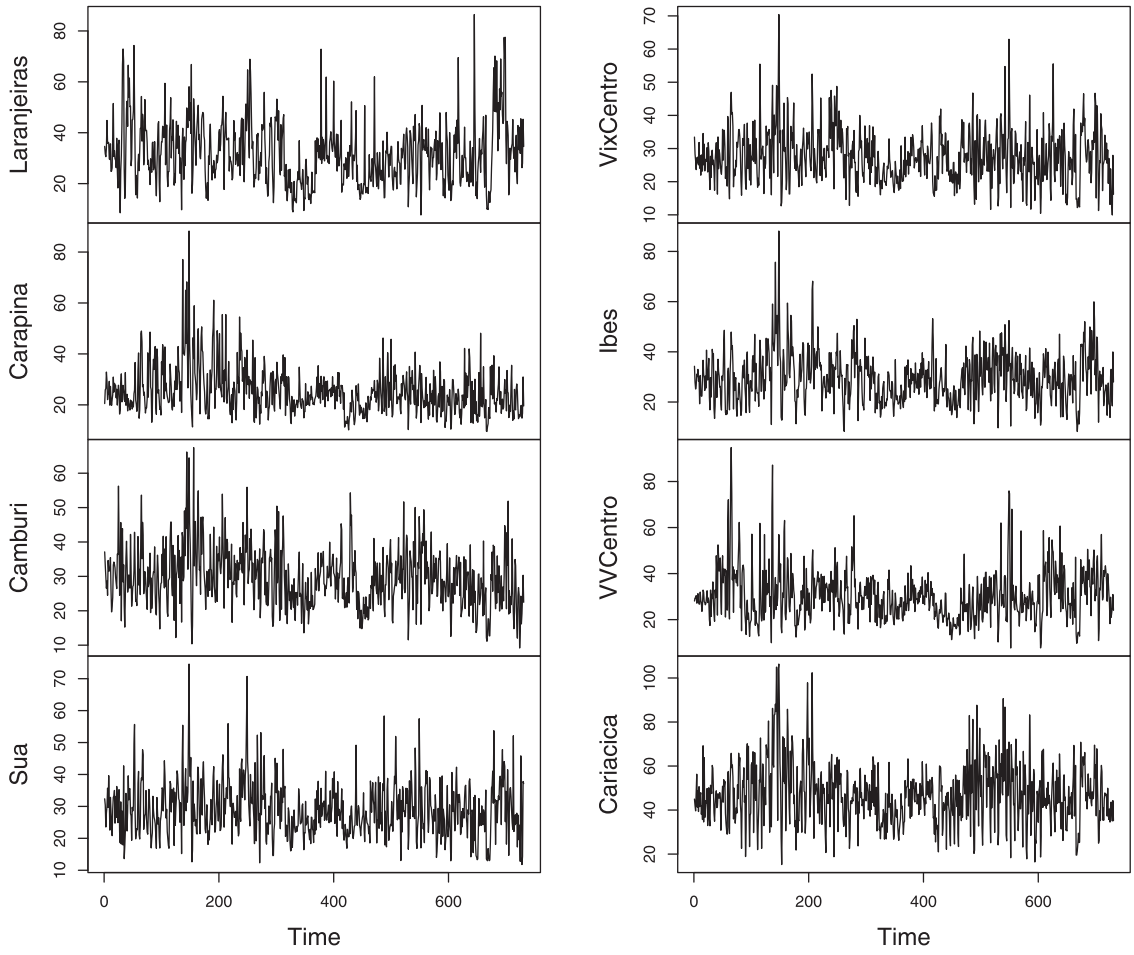


Fig. 1. Plots of the PM_{10} pollutant concentrations of the eight stations of AAQMN ($k = 8$).

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.

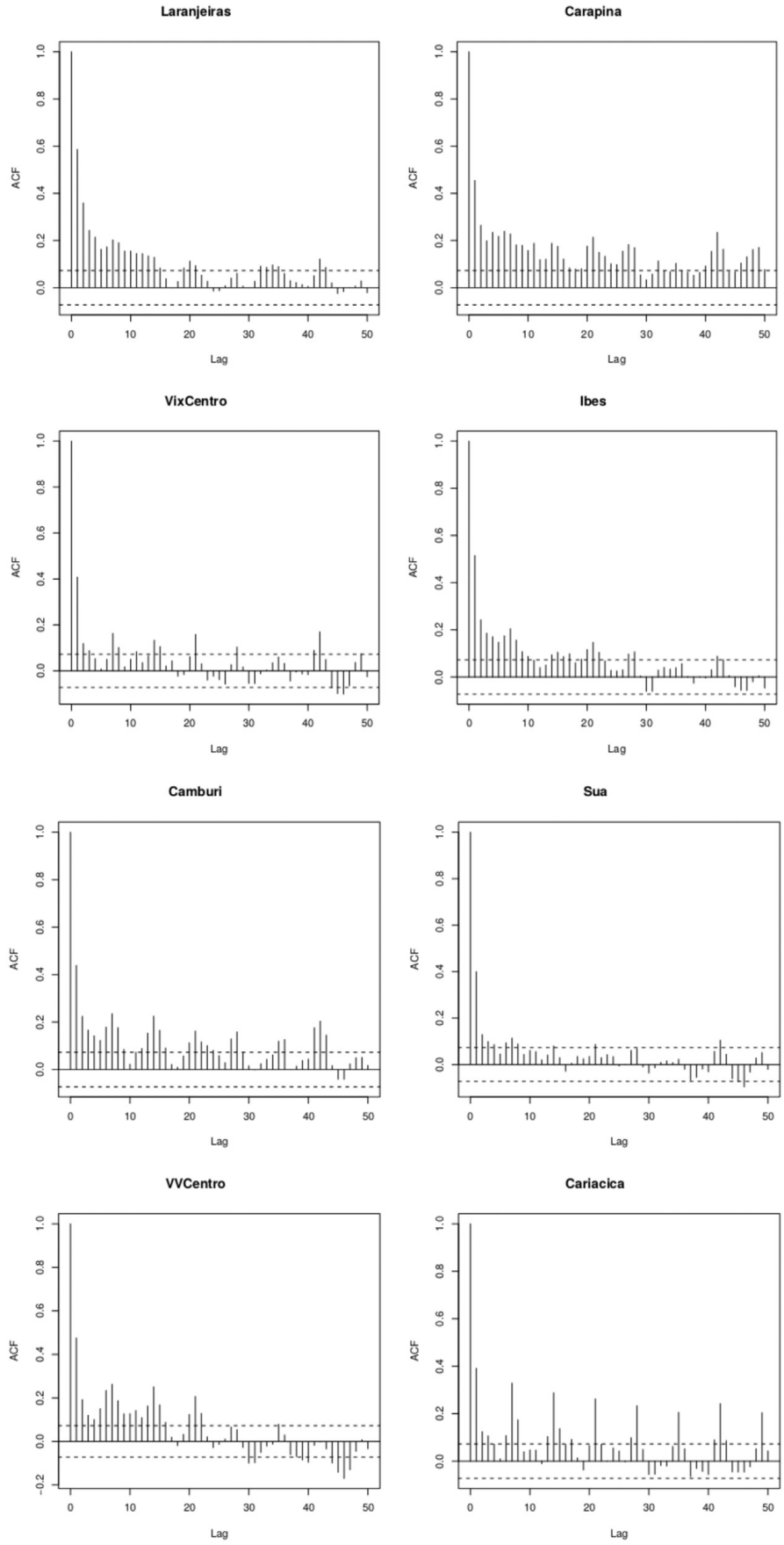


Fig. 2. Classical ACF estimates of the PM_{10} pollutant concentrations.

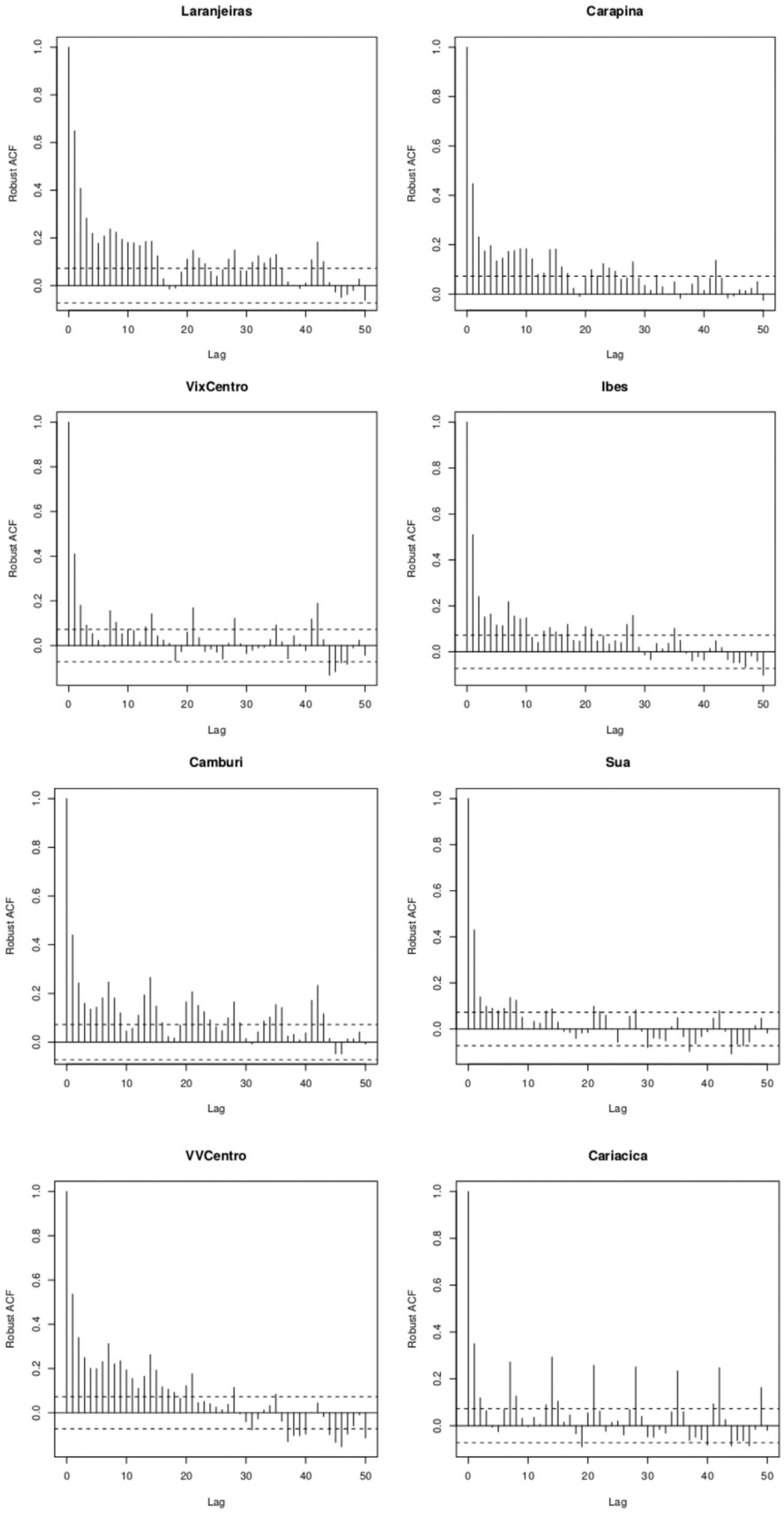


Fig. 3. Robust ACF estimates of the PM_{10} pollutant concentrations.

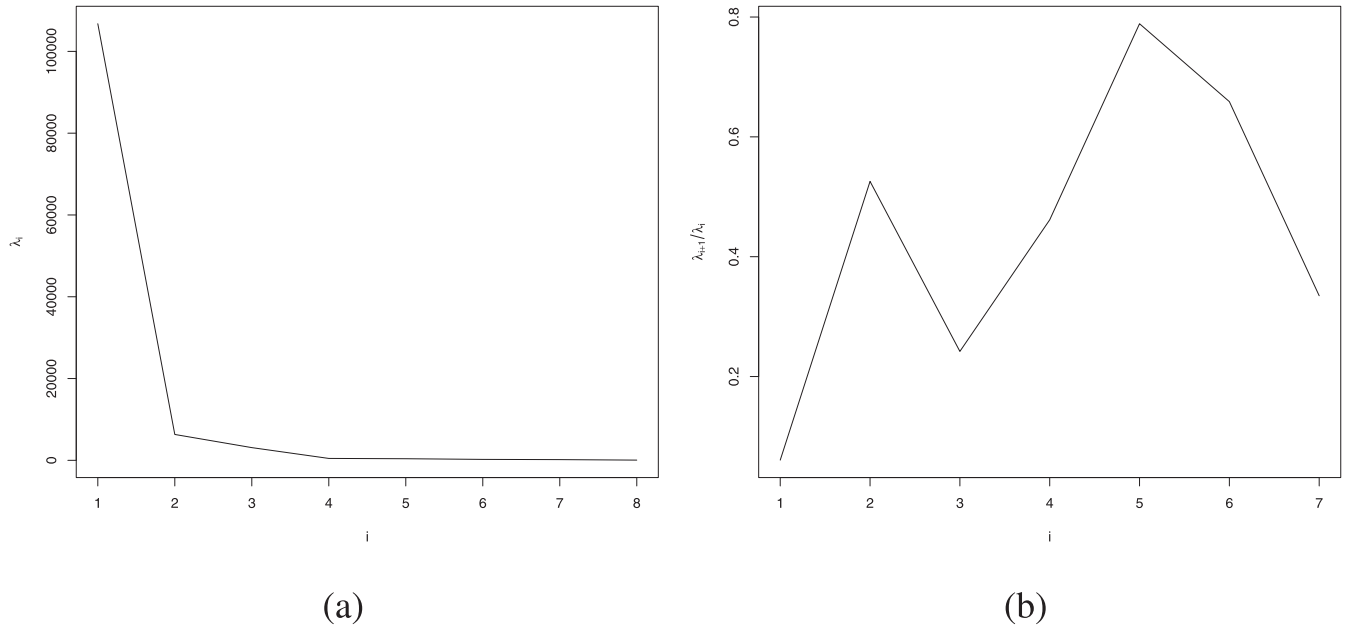


Fig. 4. A scree plot (a) and the plot of the ratios (b) of the eigenvalues of $\hat{\boldsymbol{M}}$.

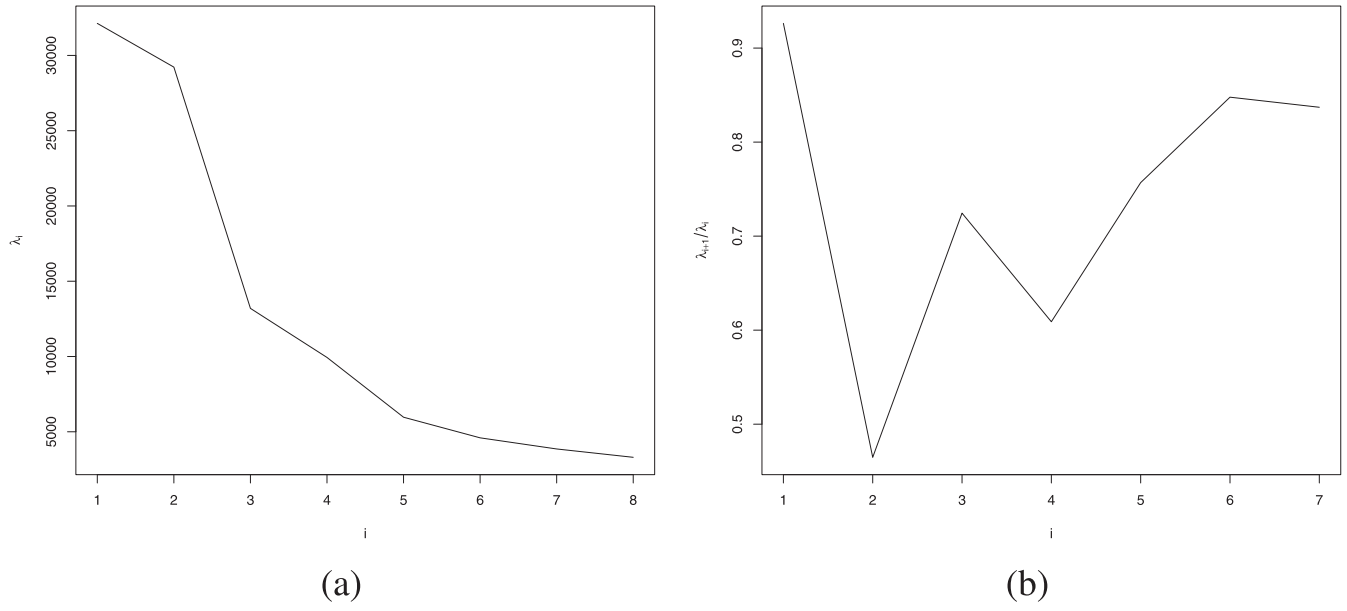
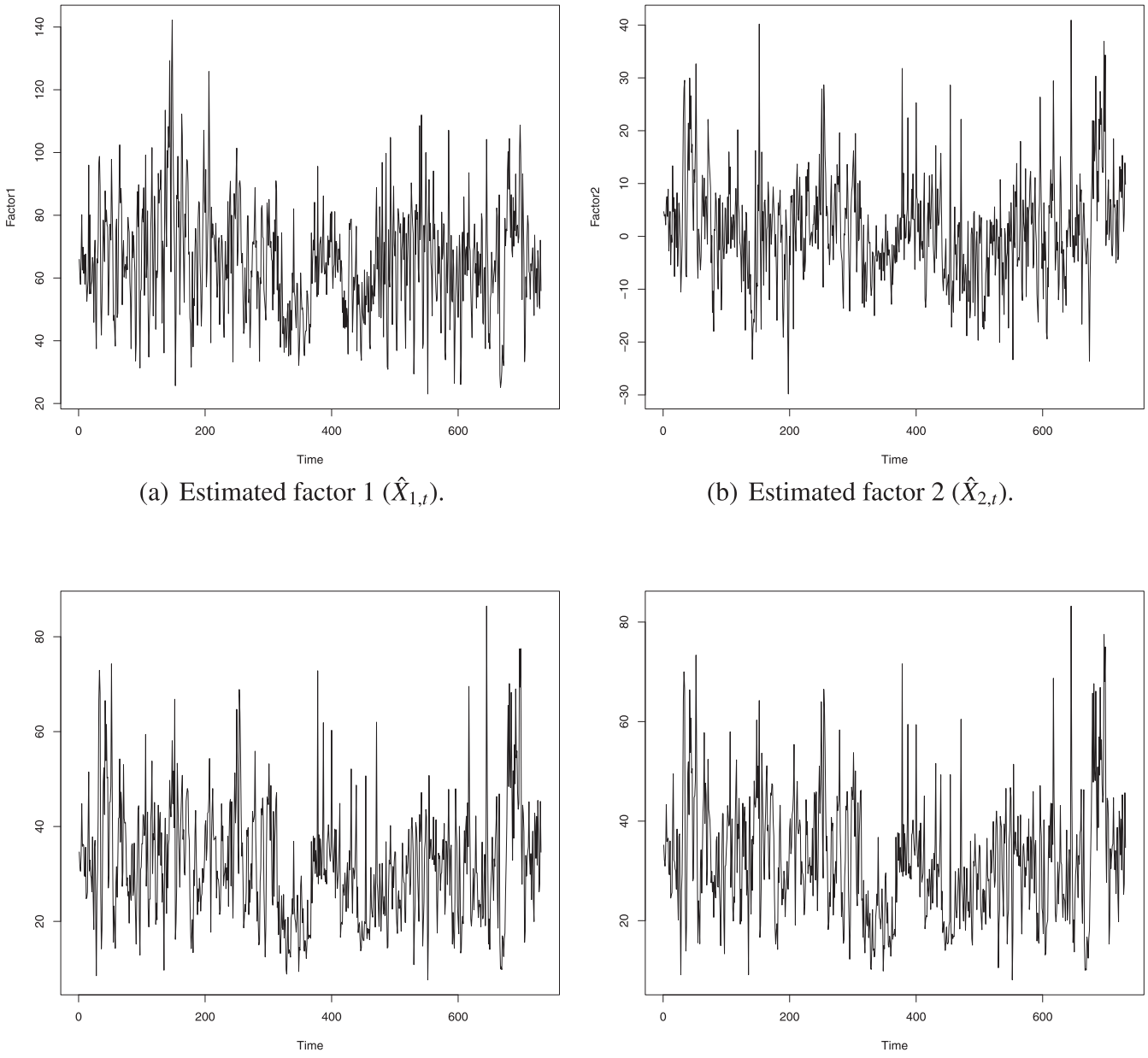


Fig. 5. A scree plot (a) and the plot of the ratios (b) of the eigenvalues of $\hat{\boldsymbol{M}}^{Q}$.

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 $\hat{r}^{Q}=2$ orthonormal eigenvectors of $\hat{\boldsymbol{M}}^{Q}$ corresponding to its $\hat{r}^{Q}=2$ 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

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

where

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

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

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.

On generalized additive models with dependent time series covariates

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

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

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

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

1 Introduction

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

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

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

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

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

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

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

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

2 The GAM-PCA-VAR model

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

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

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

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

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

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

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

with link

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 Theoretical results

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

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

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

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

Proof. By conditioning we have that

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

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

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

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

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

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

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

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

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

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

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

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

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

From Equation (10), we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Since

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

similarly to equation (9) we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4 Simulation study

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

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

Table 1. Simulation results for model in Example 1

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

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

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

5 Application to air pollution data

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

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

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

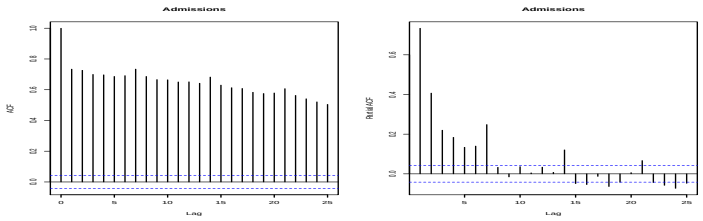


Fig. 1. Sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the response variable.

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

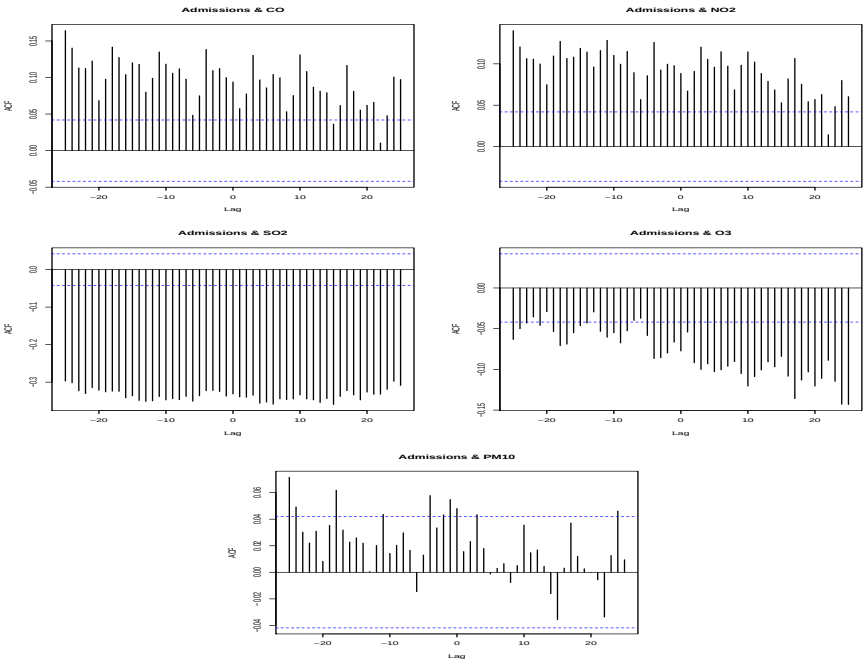


Fig. 2. Sample cross-correlation function (CCF) of the response and pollutant variables.

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

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

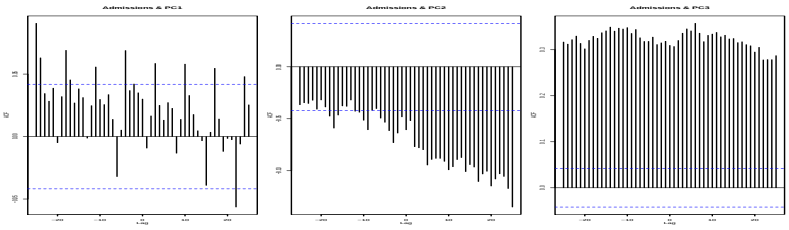


Fig. 3. Sample cross-correlation function (CCF) of the response and first three PCs.

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

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

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

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

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

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

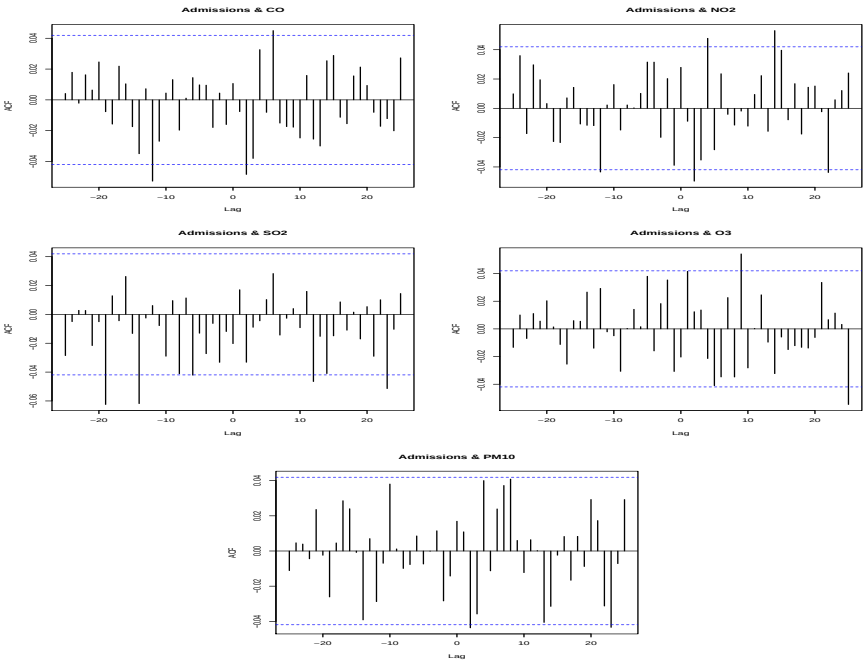


Fig. 4. Sample cross-correlation function (CCF) between the response and pollutant variables after the filtering.

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

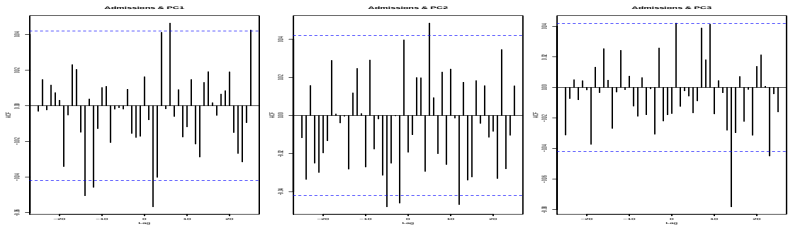


Fig. 5. Sample cross-correlation function (CCF) between the response and PCs after the filtering.

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

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

6 Conclusions

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

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

Acknowledgments

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

References

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

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

Principal component analysis with autocorrelated data

Bartolomeu Zamprogno[1,2], Valdério A. Reisen[1,2,3], Pascal Bondon[3], Higor H. Aranda Cotta[1,2,3] and Neyval C. Reis. Jr.[2]

[1]*NuMEs-DEST-CCE,* [2]*PPGEA-CT – Federal University of Espírito Santo, Brazil*

[3] *Laboratoire des Signaux et Systèmes – CNRS, CentraleSupélec, Univ. Paris-Sud, Université Paris-Saclay, France*

Abstract

This paper contributes to the analysis, interpretation and use of the principal component analysis (PCA) of the variance-covariance matrix of a multivariate time-correlated linear process. The effect of ignoring the autocorrelation structure of the process when applying and interpreting PCA is investigated. The spurious impact of the time-correlation on the eigenvalues is studied. To mitigate this impact, a pre-filtering procedure to whiten the data is suggested. The results are justified theoretically and empirically. The proposed methodology is used to identify redundant particulate matter measurements in the Great Vitória Region (GVR) in Brazil. Among the eight considered monitoring stations, it is found that three are needed to characterize the region.

Keywords: Principal component analysis, autocorrelation, cross-correlation, eigenvalue, air pollution.

1 Introduction

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

The use of PCA not focuses only on reducing the dimension of the data. For example, Karar & Gupta (2007) used PCA as a cluster analysis to identify sources of pollution and Romero et al. (1999), White et al. (1991) and Cohen (1983) applied PCA to identify homogeneous sub-regions of climatic stations in a large geographical area. Besides the use of PCA as a cluster analysis, several studies used the technique to extenuate the multicollinearity in a regression analysis context and to detect outliers, see e.g., Liu (2009), Wang & Pham (2011*a*), Souza et al. (2014), Souza et al. (2018) and Reisen et al. (2019). PCA is also a step procedure in other multivariate techniques such as factor analysis, canonical correlation analysis and discriminant analysis, among others.

In the air quality area, the identification of pollution sources using PCA has been considered by many authors, for example, Statheropoulos et al. (1998), Borbon et al. (2002), Wang & Shooter (2004), Karar & Gupta (2007) and Shi et al. (2009), among others. In the network management context, Pires et al. (2008*a*,*b*) used PCA with monitored pollutant concentrations data to manage the monitoring network of the metropolitan area of Porto (Portugal) to reduce costs. The authors proposed to select only one station among those belonging to a same cluster and having similar concentrations behaviours. They concluded that six stations instead of ten are sufficient to measure

the level of concentration of sulphur dioxide (SO_2), and no more than two stations are required for monitoring the particulate matter less than 10 μm in diameter (PM_{10}). Lu et al. (2011) evaluated the performance of PCA and cluster analysis for the management of the local air quality monitoring network of Hong Kong (China) with the aim to identify city areas with similar air pollution behaviour and to locate emission sources. They found that the monitoring stations could be grouped into different classes based on the air pollution behaviours.

One of the usual assumptions of PCA technique is the data independence. Nevertheless, PCA is widely used with time series which are time-correlated, without justification. For example, the pollution data considered in the above cited papers are time-dependent. Not taking into account the temporal structure of the observations may lead to misleading analysis and interpretations. It is important to recognize that the use of standard statistical methods like PCA, neglecting the required data assumption may produce biased estimates and spurious results. The recent work of Vanhatalo & Kulahci (2016) discuss the impact of temporal correlation in the statistical process control with PCA. The authors presented some practical insights of ignoring the correlation structure of the data in PCA-based control charts. The effect of time-correlation on model estimation using PCA is also one of the main contribution of Souza et al. (2018), where the multicollinearity issue when using pollutants as covariates in the generalized additive model is solved using PCA, and where it is suggested to use a multivariate time series model to remove the temporal correlation of the covariates. Wang & Pham (2011*b*) also considered PCA in the regression model to quantify the relationship between morbidity and pollutants; however, the temporal correlation of the variables was ignored by the authors.

The purpose of this paper is to fill some gaps when applying PCA technique to multivariate time series data. In this context, the objective is to evaluate the effect of different correlation structures of a multivariate stationary process in the interpretation and inference of the principal component (PCs) computed from the variance-covariance matrix. The study is justified empirically and theoretically, and a real data set of pollutant concentrations is considered as an example of application. Proposition 1 shows that the PCs are autocorrelated and cross-correlated. Thus, the paper suggests to pre-whitening the data with a linear model to attenuate the time-correlation before applying PCA. This whitening technique has been considered by some authors in the econometric area, but without discussing the consequence of neglecting the temporal correlation. For example, Matteson & Tsay (2011) and Hu & Tsay (2014) applied vector autoregressive (VAR) models to remove the serial correlation of time series of stock returns before carrying out PCA of the residuals.

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

2 PCA with time series data

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

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

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

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

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

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

for all $h \geq 0$. Although the elements of ε_t at different times are uncorrelated, they may be contemporaneously correlated when Σ_ε is not diagonal. In the following, we assume without loss of generality that $\mu = 0$.

In the analysis of a multivariate data set, PCA looks for linear combinations of the components capturing the highest percentage of variation of the data. This technique depends exclusively on the covariance or the correlation matrix of the data, see e.g., Anderson (2003), Jollife (2002) and Johnson & Wichern (1998). PCA is well suited for time-independent observations since it explains only the contemporaneous correlation of the data and does not take into account the time-correlation. Specifically, PCA calculates the characteristic roots and vectors of $\Gamma_X(0)$. Let $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$ be the non necessarily distinct eigenvalues of $\Gamma_X(0)$ with corresponding orthonormal (with respect to the usual inner product) eigenvectors $p_1, \ldots, p_k$. Then $\Gamma_X(0) p_i = \lambda_i p_i$ for $i = 1, \ldots, k$, and $P' \Gamma_X(0) P = \Lambda$ where P is the $k \times k$ matrix whose ith column is p_i and Λ is the $k \times k$ diagonal matrix whose ith diagonal element is λ_i, i.e., $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, see e.g. Banerjee & Roy (2014, Theorem 11.27). Equivalently, $\Gamma_X(0)$ admits the so-called spectral decomposition

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

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

Proposition 1. Let X_t be defined by (1) where, without loss of generality, it is assumed that $\mu = 0$. Let $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$ be the eigenvalues of $\Gamma_X(0)$ with corresponding orthonormal eigenvectors $p_1, \ldots, p_k$, and $Y_{it} = p_i' X_t$ be the ith PC for $i = 1, \ldots, k$. Then,

a) $\mathrm{Var}(Y_{it}) = p_i' \Gamma_X(0) p_i = \lambda_i$,

b) $\mathrm{Cov}(Y_{it}, Y_{jt}) = p_i' \Gamma_X(0) p_j = 0$ when $i \neq j$,

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

Remark 1. Propositions 1a,b) appear, for example, in Anderson (2003) and Johnson & Wichern (1998) in the particular case of an uncorrelated process, that is when $X_t = \varepsilon_t$ in (1). Proposition 1c) shows that the autocovariances ($i = j$) and the cross-covariances ($i \neq j$) of the PCs are non zero. This induces some issues discussed below in descriptive and inferential procedures of PCA in the case of time series data.

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

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

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

Let $A_n = \sum_{j=1}^{n} \Psi_j \Sigma_\varepsilon \Psi_j'$. We have,

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

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

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

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

A parametric class of models satisfying (1) is the k-dimensional vector seasonal autoregressive fractionally integrated moving average (VSARFIMA) model with season $s \in \mathbb{N}$, non-seasonal orders $(p, d_1, \ldots, d_k, q)$ and seasonal orders $(P, D_1, \ldots, D_k, Q)$. This model is defined by the difference equations

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

$$Z_{it} = (1-B)^{d_i}(1-B^s)^{D_i} X_{it}, \tag{8}$$

for $i = 1, \ldots, k$, where ε_t is a vector white noise with $\mathrm{E}(\varepsilon_t) = 0$ and $\Gamma_\varepsilon(h)$ given by (2), and B is the backward operator, i.e., $BX_t = X_{t-1}$ for any process X_t. For any $d \in \mathbb{R} \setminus \mathbb{Z}$, the time series $(1-B)^d X_t$ is defined by

$$(1-B)^d X_t = \sum_{k=0}^{\infty} b_k X_{t-k},$$

where

$$b_k = \prod_{j=1}^{k} \frac{j-1-d}{j} = \frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}$$

are the coefficients in the Taylor series for $(1-z)^d$ when $|z| < 1$ and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the Gamma function. It is assumed that (p, q, P, Q) are positive integers, $0 \leq d_i + D_i < 1/2$ and $0 \leq D_i < 1/2$ for $i = 1, \ldots, k$. The matrix-valued polynomials $\phi(\cdot)$, $\theta(\cdot)$, $\Phi(\cdot)$ and $\Theta(\cdot)$ given by

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

satisfy that $\det(\phi(z)\Phi(z^s)) \neq 0$ and $\det(\theta(z)\Theta(z^s)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. These two conditions are known as the causality and invertibility properties, respectively. Additional conditions have to be imposed in order to obtain an identifiable model, see e.g. Brockwell & Davis (2006, page 431) and Reinsel (1997, section 2.3). In (7)–(8), the matrix parameters ϕ_i's, θ_i's, Φ_i's, Θ_i's, and the fractional orders d_i's, D_i's are unknown and have to be estimated from the observed data $X_1, \ldots, X_n$. When all d_i's and D_i's are zero, X_t reduces to a VSARMA process and has a short-memory correlation structure in the sense that the sequence of matrices $\Gamma_X(h)$ for $h \in \mathbb{Z}$ is summable. Otherwise, X_t has a long-memory behaviour in the sense that the matrices $\Gamma_X(h)$ are only square summable, see Chung (2012). Reisen, Zamprogno, Palma & Arteche (2014) and Reisen, Sarnaglia, Reis Jr, Lévy-Leduc & Santos (2014) discussed the univariate SARFIMA model and its estimation methods.

As mentioned in Remark 3, when X_t is time-correlated, the PCs of X_t have a larger variance than the ones of ε_t. One way to mitigate this effect is to apply to X_t a multivariate linear filter, such as the VSARFIMA filter before applying PCA. In this context, PCA tools are applied to ε_t in place of X_t in (7). This issue is one of the contribution of this paper.

The VAR(1) model is the particular case of (7)–(8) where X_t satisfies the difference equation $X_t = \Phi X_{t-1} + \varepsilon_t$ with Φ a matrix parameter. This model is very simple and is widely used in modelling multivariate time series. Proposition 2 illustrates the effect of temporal correlation on the PCs Y_t when X_t is a VAR(1) process. This result can be extended to more general processes. For example, this is well-known that the VAR(p) model can be written as a VAR(1) process, see e.g. Lutkepohl (2005, page 15) and Hamilton (1994, page 259).

Proposition 2. Let X_t be a VAR(1) process where all the eigenvalues of Φ are less than one in modulus. Let the vector process $Y_t = (Y_{1t}, \cdots, Y_{kt})'$ where Y_{it}, $i = 1, \cdots, k$ are defined as in Proposition 1. Then $\Gamma_X(h) = \Gamma_X(0)(\Phi^h)'$ and $\Gamma_Y(h) = \Lambda P'(\Phi^h)'P$ for all $h \geq 0$.

Proof. It follows from Brockwell & Davis (2006, Example 11.3.1) that $\Psi_j = \Phi^j$ in (1). Then (3) implies that $\Gamma_X(h) = \Gamma_X(0)(\Phi^h)'$ for all $h \geq 0$. Thus, $\Gamma_Y(h) = P'\Gamma_X(h)P = P'\Gamma_X(0)(\Phi^h)'P = P'P\Lambda P'(\Phi^h)'P = \Lambda P'(\Phi^h)'P$ for all $h \geq 0$. □

Remark 4. In the particular case where $\Phi = \mathrm{diag}(\phi_1, \ldots, \phi_k)$ with $|\phi_i| < 1$ for $i = 1, \ldots, k$, we deduce from (3) that the (i,j)th element of $\Gamma_X(h)$, $\Gamma_X^{ij}(h)$ is given by

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

for all $h \geq 0$. If Σ_ε is diagonal, consequently, $\Gamma_X(h)$ is also diagonal, the eigenvectors of Σ_ε and $\Gamma_X(h)$ are the vectors $(e_1, \ldots, e_k)$ of the natural basis of $\mathbb{R}^k$, and the eigenvalue of Σ_ε, resp. $\Gamma_X(h)$, associated to e_i is Σ_ε^{ii}, resp. $\phi_i^h/(1-\phi_i^2)\Sigma_\varepsilon^{ii}$. Nevertheless, if some of the ϕ_i's are distinct, we deduce from Proposition 2 that $\Gamma_Y(h)$ is not necessarily diagonal for $h > 0$. This is an interesting case where the components of X_t are not cross-correlated but the components of the PCs are. If, $\phi_i = \phi$ for $i = 1, \ldots, k$ and Σ_ε is any nonnegative definite matrix, $\Gamma_X(h) = \phi^h/(1 - \phi^2)\Sigma_\varepsilon$ for all $h \geq 0$, the eigenvectors of $\Gamma_X(h)$ and Σ_ε are the same, while the eigenvalues of $\Gamma_X(h)$ are the ones of $\Gamma_X(0)$ multiplied by $\phi^h/(1 - \phi^2)$. Furthermore, it follows from Proposition 2 that $\Gamma_Y(h) = \phi^h\Lambda$ for all $h \geq 0$. In this case, the components of X_t are cross-correlated when Σ_ε is not diagonal, but the components of the PCs are not.

Example 1. Consider the case where $k = 2$ and X_t is the VAR(1) process defined by

$$\Phi = \begin{bmatrix} 0 & \phi \\ 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \Sigma_\varepsilon = \begin{bmatrix} \alpha & 0 \\ 0 & 1-\alpha \end{bmatrix}$$

where $0 < \alpha < 1$. Then $\det(I - \Phi z) = 1 \neq 0$ for all $z \in \mathbb{C}$ for any $\phi \in \mathbb{R}$. We have

$$\begin{aligned} X_{1t} &= \phi\, \varepsilon_{2(t-1)} + \varepsilon_{1t}, \\ X_{2t} &= \varepsilon_{2t}. \end{aligned}$$

Then

$$\Gamma_X(0) = \begin{bmatrix} \alpha + \phi^2(1-\alpha) & 0 \\ 0 & 1-\alpha \end{bmatrix} \qquad \text{and} \qquad \Gamma_X(h) = \phi(1-\alpha) \begin{bmatrix} 0 & \delta_{-1}(h) \\ \delta_1(h) & 0 \end{bmatrix}$$

for all $h \neq 0$, where

$$\delta_a(h) = \begin{cases} 1 & \text{if } h = a, \\ 0 & \text{otherwise.} \end{cases}$$

We have $\lambda_1 = \alpha + \phi^2(1-\alpha)$ if $2\alpha - 1 + \phi^2(1-\alpha) \geq 0$, which is the case when $\alpha \geq 1/2$ for any $\phi \in \mathbb{R}$, and if $\alpha \leq 1/2$ and $\phi^2 \geq (1-2\alpha)/(1-\alpha)$. In the other cases, $\lambda_1 = 1 - \alpha$. This simple example illustrates the effect of the time-correlation which is related to parameter ϕ on the eigenvalues of $\Gamma_X(0)$. In particular, when $\alpha \leq 1/2$, the eigenvector with the largest eigenvalue of Σ_ε is $(0,1)'$, whereas the eigenvector with the largest eigenvalue of $\Gamma_X(0)$ is $(1,0)'$ if $\phi^2 \geq (1-2\alpha)/(1-\alpha)$.

The MA(1) model is the particular case of (7)–(8) where X_t satisfies the difference equation $X_t = \varepsilon_t + \Theta\varepsilon_{t-1}$ with Θ a matrix parameter. Proposition 3 gives the expressions of $\Gamma_X(h)$ and $\Gamma_Y(h)$ when X_t is a MA(1) process. As for the VAR(1) model, this result can be extended to more complicated processes.

Proposition 3. Let X_t be a MA(1) process where all the eigenvalues of Θ are less than one in modulus and Y_t defined as previously. Then

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

Proof. The results follow readily from the difference equation $X_t = \varepsilon_t + \Theta\varepsilon_{t-1}$. □

Remark 5. In the particular case where $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_k)$ with $|\theta_i| < 1$ for $i = 1, \ldots, k$, we deduce from Proposition 3 that

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

If Σ_ε is diagonal, so is $\Gamma_X(h)$ and the eigenvalue of Σ_ε, $\Gamma_X(0)$ and $\Gamma_X(1)$ associated to e_i is Σ_ε^{ii}, $(1 + \theta_i^2)\Sigma_\varepsilon^{ii}$ and $\theta_i\Sigma_\varepsilon^{ii}$, respectively. If, $\theta_i = \theta$ for $i = 1, \ldots, k$, Σ_ε, $\Gamma_X(0)$ and $\Gamma_X(1)$ have the same eigenvectors, while the eigenvalues of $\Gamma_X(0)$ and $\Gamma_X(1)$ are the ones of Σ_ε multiplied by $1 + \theta^2$ and θ, respectively. Furthermore, in this case, we deduce from Proposition 3 that $\Gamma_Y(1) = \theta/(1+\theta^2)\Lambda$.

Example 2. Consider the case where $k = 2$ and X_t is the MA(1) process defined by

$$\Theta = \begin{bmatrix} \theta_1 & \theta_2 \\ 0 & 0 \end{bmatrix} \qquad \text{and} \qquad \Sigma_\varepsilon = \begin{bmatrix} \alpha & 0 \\ 0 & 1-\alpha \end{bmatrix}$$

where $0 < \alpha < 1$. Then $\det(I + \Theta z) = 1 + \theta_1 z \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$ if and only if $|\theta_1| < 1$. We have

$$\begin{aligned} X_{1t} &= \varepsilon_{1t} + \theta_1\, \varepsilon_{1(t-1)} + \theta_2\, \varepsilon_{2(t-1)}, \\ X_{2t} &= \varepsilon_{2t}. \end{aligned}$$

Thus

$$\Gamma_X(0) = \begin{bmatrix} \alpha(1+\theta_1^2) + \theta_2^2(1-\alpha) & 0 \\ 0 & 1-\alpha \end{bmatrix} \quad \text{and} \quad \Gamma_X(h) = \theta_2(1-\alpha) \begin{bmatrix} 0 & \delta_{-1}(h) \\ \delta_1(h) & 0 \end{bmatrix}$$

for all $h \neq 0$. We have $\lambda_1 = \alpha(1+\theta_1^2) + \theta_2^2(1-\alpha)$ if $2\alpha - 1 + \alpha\theta_1^2 + \theta_2^2(1-\alpha) \geq 0$, which is the case when $\alpha \geq 1/2$ for any θ_1, θ_2, and if $\alpha \leq 1/2$ and $\alpha\theta_1^2 + \theta_2^2(1-\alpha) \geq 1 - 2\alpha$. In the other cases, $\lambda_1 = 1 - \alpha$. As in Example 1, the eigenvector with the largest eigenvalue of $\Gamma_X(0)$ and Σ_ε may be different when the time-correlation if large enough. Finally, observe that when $\theta_1 = 0$, X_t is the VAR(1) process of Example 1 where ϕ is replaced by θ_2, and then the same results can be derived by replacing ϕ by θ_2.

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

$$\hat{\Gamma}_X(0) = \frac{1}{n} \sum_{t=1}^{n} X_t X_t', \tag{9}$$

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

$$\hat{\Gamma}_X(0) = BLB', \tag{10}$$

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

A major concern about using PCA is how many PCs should be selected. Several criteria are proposed in the literature such as the eigenvalues plot of Jollife (2002) and the mean eigenvalue test of Perez-Neto et al. (2005). Assume that the random variables X_t are mutually independent and identically distributed with finite moments and $\lambda_1 > \ldots > \lambda_k > 0$. Fujikoshi (1980, Theorem 1) generalized Anderson (2003, Theorem 13.5.1) to non Gaussian data and showed that $\sqrt{n}(l_i - \lambda_i)$ has the limiting normal distribution $\mathrm{N}(0, 2\lambda_i^2 + \kappa_4^i)$, where κ_4^i is the fourth-order cumulant of the ith component X_{it} of X_t for all $i = 1, \ldots, k$. Therefore, an asymptotic confidence interval (ACI) of significance level α for λ_i is

$$l_i - \sqrt{\frac{2l_i^2 + \hat{\kappa}_4^i}{n}}\, z_{\frac{\alpha}{2}} \leq \lambda_i \leq l_i + \sqrt{\frac{2l_i^2 + \hat{\kappa}_4^i}{n}}\, z_{\frac{\alpha}{2}}, \tag{11}$$

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

$$R_m - \sqrt{\frac{\sum_{i=1}^k \hat{T}_i^2(2l_i^2+\hat{\kappa}_4^i)}{n}}\, z_{\frac{\alpha}{2}} \leq \tau_m \leq \; R_m + \sqrt{\frac{\sum_{i=1}^k \hat{T}_i^2(2l_i^2+\hat{\kappa}_4^i)}{n}}\, z_{\frac{\alpha}{2}}, \tag{12}$$

where $\hat{T}_i = (c_i - R_m)/(l_1+\cdots+l_k)$. In Section 3, the ACI's (11) and (12) are used.

3 Numerical experiments

This section presents finite sample size studies to illustrate and to quantify empirically the effect of time-correlation on the interpretation and testing of PCA. First, we consider the simple Examples 1 and 2 and show the behavior of the estimates l_1 and l_2 of λ_1 and λ_2, respectively, for different values of ϕ, θ_1, θ_2 and for the sample sizes $n = 30, 100, 500$. Then we study more complex VAR(1) models with four components. The number of replications in our Monte Carlo simulations is 1000.

We consider Example 1 where X_t is a Gaussian process and we set $\alpha = 0.4$. If $\phi = 0$, $X_t = \varepsilon_t$, $\lambda_1 = 1-\alpha = 0.6$ and the eigenvector associated to λ_1 is $(0,1)'$. For each replicate $m = 1,\ldots,1000$, we denote by l_{1m} and l_{2m} the eigenvalues of $\hat{\Gamma}_X(0)$ with $l_{1m} \geq l_{2m}$ and we build the ACI of significance level 0.95 for λ_2 given by (11), i.e., $[l_{2m}(1 \pm 1.96\sqrt{2/n})]$. We report in Table 1 the percentage of replicates for which 0.4 is outside this ACI as n and ϕ increase. According to the results in Example 1, $\lambda_2 = 0.4 + 0.6\phi^2$ if $|\phi| \leq 1/\sqrt{3}$, and $\lambda_2 = 0.6$ if $|\phi| \geq 1/\sqrt{3}$. Therefore, for a given n large enough, for l_{2m} to be a good estimate of λ_2, the percentage in Table 1 increases as $|\phi|$ increases. When n is small, the results becomes unreliable because l_{2m} may be severely biased and the length of the ACI is large.

Table 1: Percentage of replicates for which 0.4 is outside the 95 % ACI of λ_2 for the Gaussian AR(1) process.

ϕ	n		
	30	100	500
0	10	5.5	5.1
0.2	10	6.3	5.2
0.5	4.1	21	100
0.6	3.5	40	100
0.8	5	61	100
0.9	5.2	65	100

We consider Example 2 where X_t is a Gaussian process and we set $\alpha = 0.4$. If $\theta_1 = \theta_2 = 0$, $X_t = \varepsilon_t$, $\lambda_1 = 1-\alpha = 0.6$ and the eigenvector associated to λ_1 is $(0,1)'$. As before, we build the ACI of significance level 0.95 for λ_2, $[l_{2m}(1 \pm 1.96\sqrt{2/n})]$. We report in Table 2 the percentage of replicates for which 0.4 is outside this ACI as n, θ_1 and θ_2 increase. According to the results in Example 2, $\lambda_2 = 0.4(1+\theta_1^2) + 0.6\theta_2^2$ if $0.4\theta_1^2 + 0.6\theta_2^2 \leq 0.2$, and $\lambda_2 = 0.6$ if $0.4\theta_1^2 + 0.6\theta_2^2 \geq 0.2$.

Therefore, as in Table 1, the percentage increases as $|\theta_1|$ and $|\theta_2|$ increase for n large enough, and the results are unreliable when n is small.

Table 2: Percentage of replicates for which 0.4 is outside the 95 % ACI of λ_2 for the Gaussian MA(1) process.

(θ_1,θ_2)	n		
	30	100	500
(0,0)	12	6.3	5.1
(0.1,0.1)	15	6.4	5.4
(0.25,0.25)	10	5.5	50
(0.4,0.4)	3	24.8	100
(0.5,0.5)	5	50	100
(0.1,0.8)	5.7	60	100
(0.8,0.8)	5.6	65	100

We consider different VAR(1) models $X_t = \Phi X_{t-1} + \varepsilon_t$ with the same matrix Σ_ε given by

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

and the matrix parameters Φ displayed in Table 3. The white noise model $X_t = \varepsilon_t$ is denoted by Model 1. It follows from Remark 4 that Models 1, 2 and 3 have the same eigenvectors which correspond to the natural basis of $\mathbb{R}^4$, have a diagonal matrix $\Gamma_X(h)$ for all $h \in \mathbb{Z}$, and $\Gamma_X(h) = 0$ for all $h \neq 0$ in the case of Model 1. Furthermore, the eigenvalue of $\Gamma_X(0)$ in Models 2 and 3 associated to e_i is $\Sigma_\varepsilon^{ii}/(1-\phi_i^2)$, where $\Phi = \text{diag}(\phi_1, \ldots, \phi_4)$. Since all ϕ_i's are equal in Model 2, the eigenvalues of $\Gamma_X(0)$ in Models 1 and 2 are proportional, which is not the case in Models 1 and 3. Contrarily to the three first models, Models 4 and 5 present cross-correlations between the components of X_t at different lags h. According to Proposition 2, $\Gamma_X(h) = \Gamma_X(0)(\Phi^h)'$ for all $h \geq 0$. Therefore, if the entries of $\Gamma_X(0)$ are nonnegative, large positive entries of Φ implies large positive cross-covariances. In this sense, Model 5 presents stronger cross-covariances than Model 4.

Table 3: Matrix parameters Φ of VAR(1) Models 2 to 5.

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

The matrices $\Gamma_X(0)$ of the VAR(1) models are presented in Table 4. For each model, we see that $\text{tr}(\Gamma_X(0)) \geq \text{tr}(\Sigma_\varepsilon)$, as mentioned in Remark 3.

Table 4: Covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 2 to 5.

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

Table 5 shows for each VAR(1) model, the eigenvalues λ_i's of $\Gamma_X(0)$ with their respective percentage of variability $\lambda_i/(\lambda_1 + \cdots + \lambda_4)$. These percentages are the same for Models 1 and 2 since the λ_i's are proportional. Model 3 presents more variability than Models 1 and 2 because λ_1 is much larger than the other eigenvalues. Since the parameters Φ of Models 2 and 4 are close, the associated eigenvalues of $\Gamma_X(0)$ and their percentages of variability are similar. Now, the large positive cross-covariance in Model 5 compared to Model 2 increases drastically the variability of the eigenvalues of $\Gamma_X(0)$ and the first PC captures almost all the variability. This is a problem of great practical relevance, for instance in the context of reducing the data dimension. These results are comparable to the simulations presented by Vanhatalo & Kulahci (2016) on the impact of different degrees of correlation on the statistical process chart control.

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

Model	λ_1	λ_2	λ_3	λ_4	% λ_1	% λ_2	% λ_3	% λ_4
1	10.00	5.00	3.00	1.00	52.63	26.32	15.79	5.26
2	10.99	5.49	3.30	1.10	52.63	26.32	15.79	5.26
3	27.78	6.67	3.30	1.01	71.68	17.20	8.51	2.61
4	11.21	5.50	3.79	1.16	51.73	25.39	17.51	5.37
5	60.09	8.29	5.44	3.24	77.98	10.75	7.06	4.21

Let $\tau_2 = (\lambda_1 + \lambda_2)/(\lambda_1 + \cdots + \lambda_4)$ be the fraction of the variance explained by the first two PCs, and $R_2 = (l_1 + l_2)/(l_1 + \cdots + l_4)$ an estimate of τ_2. For Models 1 and 2, we have $\tau_2 = 15/19$. For each VAR(1) model and for each replicate $m = 1, \ldots, 1000$, we simulate a Gaussian process X_t, we denote by $l_{1m} \geq \cdots \geq l_{4m}$ the eigenvalues of $\hat{\Gamma}_X(0)$, $R_{2m} = (l_{1m} + l_{2m})/(l_{1m} + \cdots + l_{4m})$, and we build the ACI of significance level 0.95 for τ_2 given by (12), i.e., $[R_{2m} \pm 1.96\sqrt{2\sum_{i=1}^4 \hat{T}_{im}^2 l_{im}^2/n}]$ where $\hat{T}_{im} = (c_i - R_{2m})/(l_{1m} + \cdots + l_{4m})$, $c_1 = c_2 = 1$ and $c_3 = c_4 = 0$. We report in Table 6 the percentage of replicates for which 15/19 is outside this ACI for different sample sizes n. As expected, Models 1 and 2 present similar results and the percentage gets closer to 5% as n increases. Model 4 is very interesting because although it is quite close to Model 2, the percentage for $n = 500$ is 36 instead of 5. This is due to the length of the ACI which also depends on R_{2m}. Therefore, even

a moderate cross-covariance may induce a large difference in the rejection rate when the eigenvalues are estimated. For Models 3 and 5, the variability of the eigenvalues concentrate on the first PC, we have $\tau_2 \simeq 0.888$ which explains why the percentages are very high, even with a small sample size. Therefore, as already mentioned, in practical situations, when the vector of observations is time-dependent, PCA methodology must be used with caution. The cross-covariance between the components of the observations cannot be neglected.

Table 6: Percentage of replicates for which 15/19 is outside the 95 % ACI of τ_2 for the Gaussian VAR(1) Models 1 to 5.

Model	n		
	30	100	500
1	10.2	7.5	5
2	11.6	8	6
3	70	97.5	100
4	11	7.4	36
5	70	100	100

We have seen that nonnegative entries in $\Gamma_X(0)$ associated to large positive entries of Φ imply large positive cross-covariances in $\Gamma_X(h)$ for all $h \geq 0$ and increase the percentage of variability of the first PC. Now, we present some simulations with more general VAR(1) models. Specifically, we consider the VAR(1) models $X_t = \Phi X_{t-1} + \varepsilon_t$ with the same matrix Σ_ε given by

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

and the matrix parameters Φ displayed in Table 7. The white noise model $X_t = \varepsilon_t$ is denoted by Model 6.

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

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

The covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 7 to 10 are presented in Table 8. For each model, we have $\mathrm{tr}(\Gamma_X(0)) \geq \mathrm{tr}(\Sigma_\varepsilon)$ in agreement with Remark 3. The trace of $\Gamma_X(0)$ represents the total variability of the PCs of X_t and increases from Model 6 to Model 10.

Table 8: Covariance matrices $\Gamma_X(0)$ of the VAR(1) Models 7 to 10.

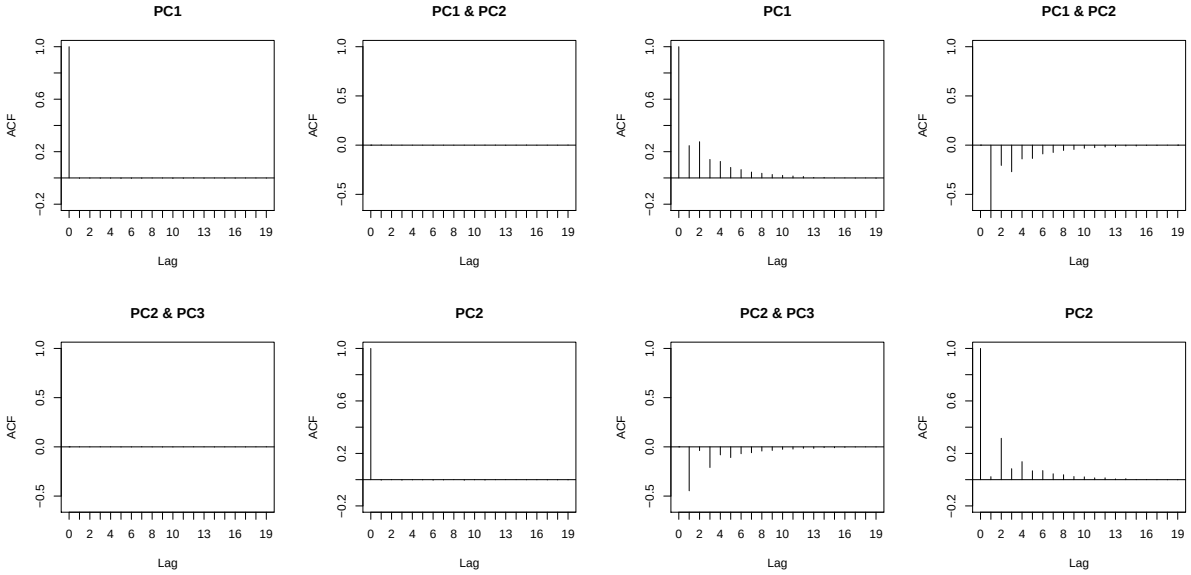
Model 7				Model 8			
132.29	30.00	42.73	54.39	169.33	24.00	49.47	44.29
30.00	58.00	33.00	70.00	24.00	77.33	31.43	116.67
42.73	33.00	85.33	89.23	49.47	31.43	64.65	53.70
54.39	70.00	89.23	337.25	44.29	116.67	53.70	477.78
Model 9				Model 10			
240.04	193.95	104.52	81.18	575.20	44.82	183.86	120.35
193.95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

Table 9 shows the eigenvalues λ_i's of the matrices $\Gamma_X(0)$ for each VAR(1) model and their respective percentage of variability $\lambda_i/(\lambda_1 + \cdots + \lambda_4)$. Comparing with Table 5, we see that the cross-covariances in Models 7 to 10 do not have drastic effects in the interpretation of PCA compared to Model 6.

Table 9: Eigenvalues of $\Gamma_X(0)$ of the VAR(1) Models 6 to 10 with their percentages of variability.

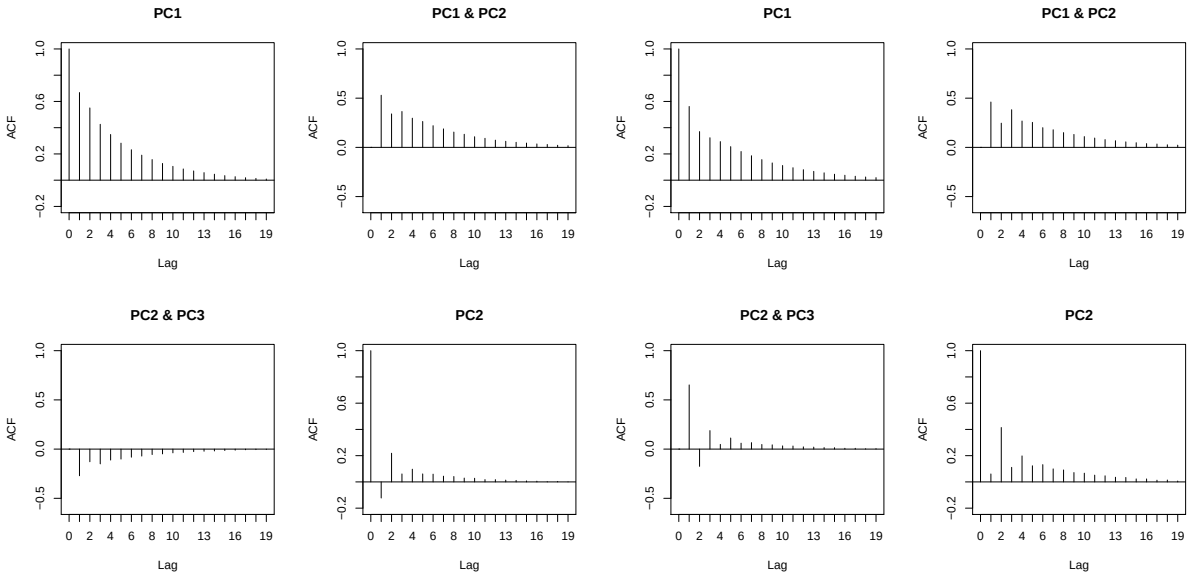
Model	λ_1	λ_2	λ_3	λ_4	% λ_1	% λ_2	% λ_3	% λ_4
6	276.42	87.71	34.22	22.65	65.66	20.83	8.13	5.38
7	402.11	125.27	50.32	35.17	65.61	20.44	8.21	5.74
8	525.90	177.62	54.26	31.32	66.65	22.51	6.88	3.97
9	626.19	164.90	112.36	34.31	66.78	17.58	11.98	3.66
10	690.65	204.24	115.34	49.78	65.16	19.27	10.88	4.70

Figures 1 and 2 plot sample autocorrelation and cross-correlation functions for $n = 1000$ of some PCs in the cases of Models 6 and 8, and Models 9 and 10, respectively. Figure 1a) shows that the PCs are neither autocorrelated nor cross-correlated in the case of a white noise. Figure 1b) shows that the PCs may be cross-correlated when the matrix parameter Φ is diagonal but the diagonal elements are not all equal. Figure 2 shows that the full correlation structure of the data is transferred to the PCs in the case of general matrices Φ and Σ_ε. These observations corroborate and illustrate Proposition 2 and Remark 4.



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

Figure 1: Sample autocorrelation and cross-correlation functions of some PCs in Models 7 and 8.



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

Figure 2: Sample autocorrelation and cross-correlation functions of some PCs in Models 9 and 10.

The above numerical experiments confirm that positive temporal cross-correlations between the components of X_t have an impact on PCA. Therefore, it is necessary to introduce procedures that allow the use of PCA with multivariate time-correlated data. This paper proposes to pre-processing the data with a multivariate linear filter in order to whiten the data before applying

PCA. The linear filter belongs to the parametric class of VSARFIMA models, for example. Note that transforming the data with linear filters to attenuate the temporal structure in multivariate techniques was also addressed in the recent work of Jaimungal & Ng (2007), Greenaway-McGrevy et al. (2012) and Hu & Tsay (2014).

4 Application to PM_{10} data

We consider PM_{10} concentrations with the aim to manage the air quality monitoring using PCA and cluster analysis, without neglecting the temporal correlation structure of the data. We investigate whether or not the temporal correlation of the variables affects PCA and its interpretation. This issue is not addressed by Pires et al. (2008*a*,*b*), for example. Furthermore, we identify the cities areas with similar PM_{10} behaviours.

The data set is collected at the automatic air quality monitoring network (AAQMN) in the GVR in Brazil and is composed by the observations of eight monitoring stations located in urban areas of four cities in the GVR. Additionally to PM_{10} concentrations, the AAQMN monitors the total suspended particles (TSP), ozone (O_3), nitrogen oxides (NO_x), carbon monoxide (CO), hydrocarbons (HC) and meteorological variables. The PM_{10} concentrations were measured from January 2005 to December 2009. Their daily averages at the eight stations constitute the time series X_t which is plotted in Figure 3.

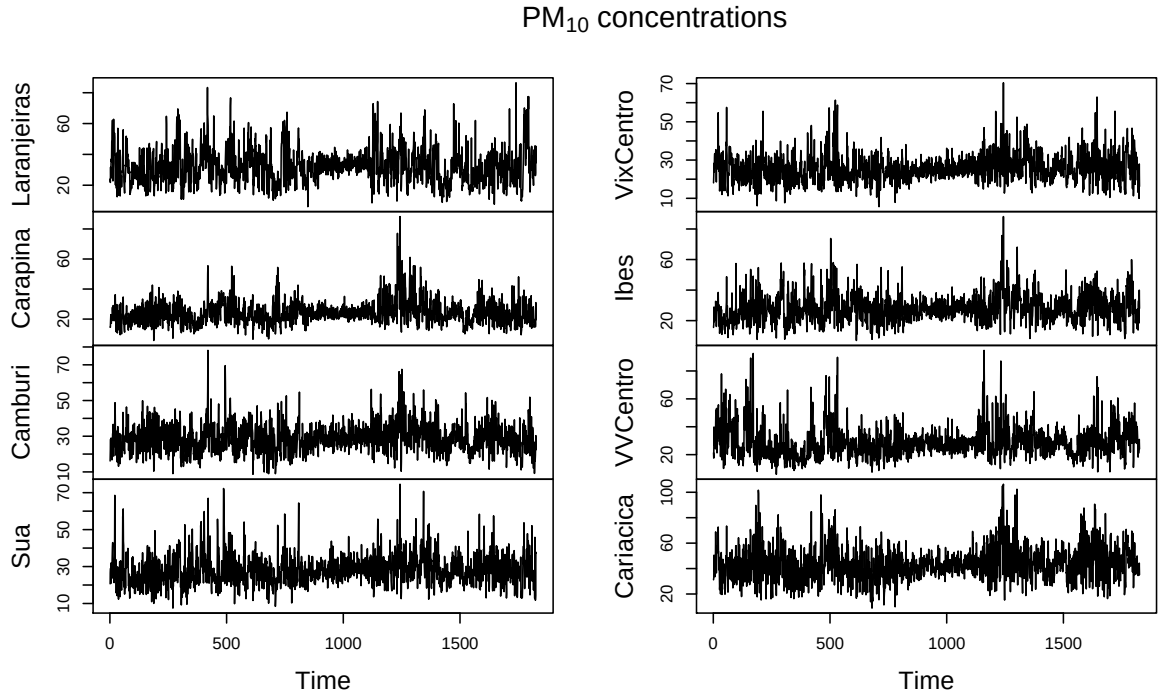


Figure 3: PM_{10} concentrations of the AAQMN.

The sample autocorrelation functions of each component of X_t are plotted in Figure 4. This

figure shows a strong weekly seasonal behaviour which is expected with daily pollution data. In addition, the sample autocorrelations are positive, decrease slowly in the first lags, in the lags multiple of seven and in the lags between the seasonal periods, which is typical of a long memory seasonal time series.

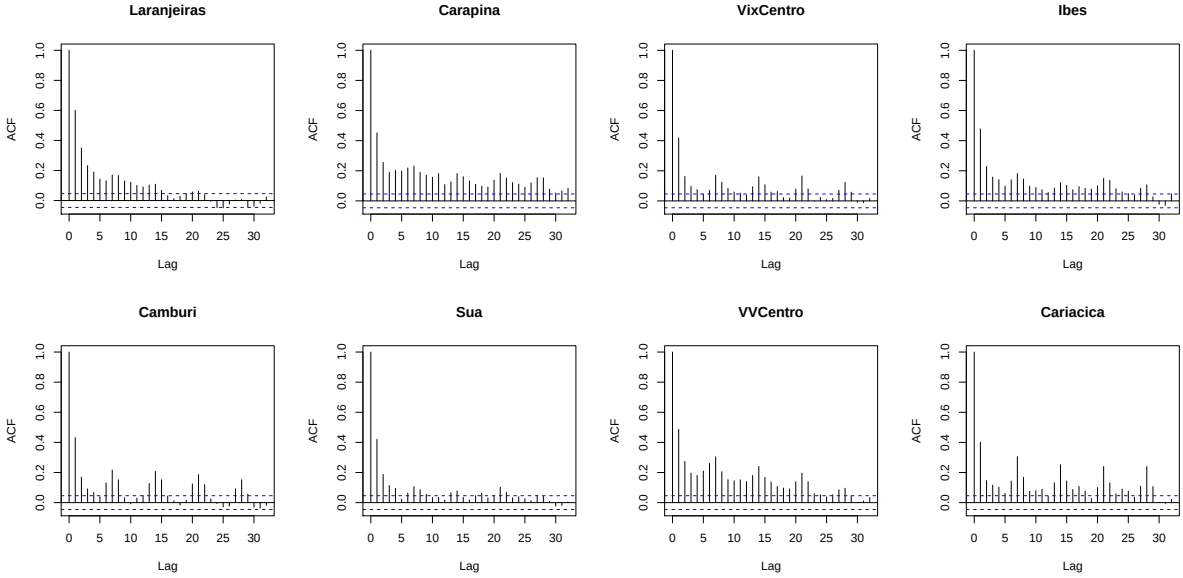


Figure 4: Sample autocorrelation functions of PM_{10} concentrations.

We fit a VSARFIMA model with season $s = 7$ to X_t. We use the estimator proposed by Reisen, Zamprogno, Palma & Arteche (2014) with the bandwidth $m = n^{0.5}$ to estimate the eight fractional parameters d_i's and D_i's in (8). These estimates, $\hat{d}_i$'s, $\hat{D}_i$'s, and their estimated standard deviations, $\hat{\sigma}(\hat{d}_i)$'s, $\hat{\sigma}(\hat{D}_i)$'s are displayed in Table 10. We see that these parameters are significant for each station.

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

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

For each $i = 1, \ldots, 8$, we build the series $\hat{Z}_{it} = (1 - B)^{\hat{d}_i}(1 - B^s)^{\hat{D}_i} X_{it}$ and we fit a VSARMA model (7) to $\hat{Z}_t$. Following the standard methodology, we choose the orders (p, q, P, Q) with an information criterion, namely the bias-corrected Akaïke information criterion (AICC), see Brockwell & Davis (2006, Section 9.2). This criterion selects a simple VAR(1) model with the following

matrix parameter

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

Apart the first diagonal element, all the coefficients of $\hat{\Phi}$ are quite small, which indicates that the fractional filtering giving $\hat{Z}_t$ extracts almost all the temporal correlation of X_t. Figure 5 plots the sample autocorrelation functions of each component of the residual $\hat{\varepsilon}_t = \hat{Z}_t - \hat{\Phi}\hat{Z}_{t-1}$ and clearly shows that these components are white noises.

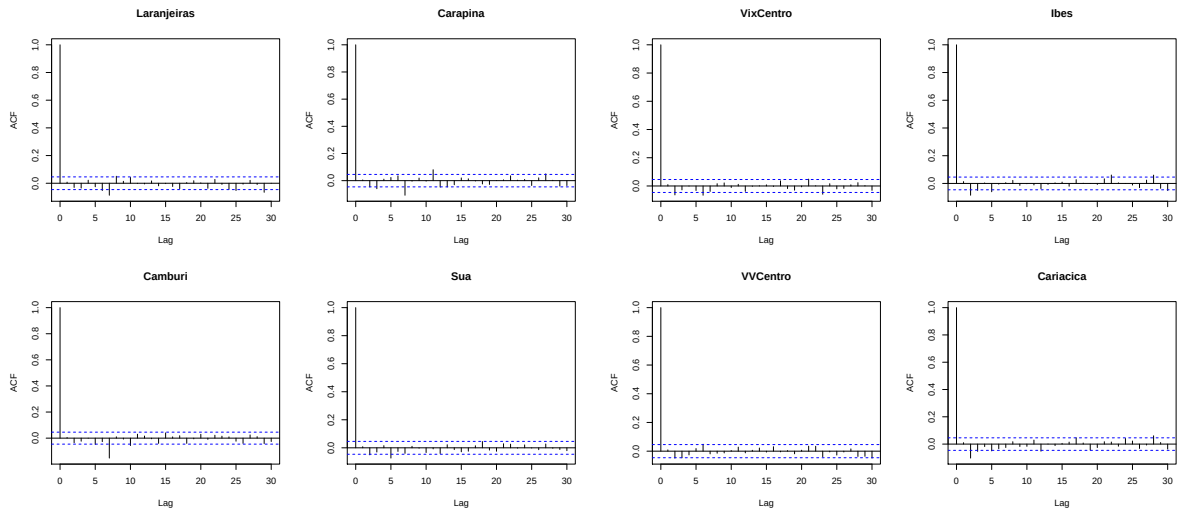


Figure 5: Sample autocorrelation functions of the residuals of the fitted VSARFIMA model to PM_{10} concentrations.

Now, we investigate the temporal correlation effect in the analysis and interpretation of PCA applied to PM_{10} data. The sample estimate $\hat{\Gamma}_X(0)$ of $\Gamma_X(0)$ is given by (9) and its spectral decomposition is (10). Let $\hat{\Gamma}_{\hat{\varepsilon}}(0) = (1/n)\sum_{t=1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_t'$ with the spectral decomposition $\hat{\Gamma}_{\hat{\varepsilon}}(0) = CMC'$, where $M = \text{diag}(m_1, \ldots, m_k)$, $m_1 \geq \ldots \geq m_k \geq 0$ are the eigenvalues of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, and C is an orthonormal matrix whose ith column c_i is an eigenvector associated to m_i for $i = 1, \ldots, k$.

In Table 11, the four columns corresponding to the PCA of $\hat{\Gamma}_X(0)$ display the eigenvectors b_i's, the eigenvalues l_i's, the proportions $l_i/(l_1 + \cdots + l_8)$'s and the cumulative proportions $(l_1 + \cdots + l_i)/(l_1 + \cdots + l_8)$'s for $i = 1, \ldots, 4$. The four columns corresponding to the PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$ display the eigenvectors c_i's, the eigenvalues m_i's, the proportions $m_i/(m_1 + \cdots + m_8)$'s and the cumulative proportions $(m_1 + \cdots + m_i)/(m_1 + \cdots + m_8)$'s for $i = 1, \ldots, 4$. For both PCA, the main part of the variability is captured by the first PC, namely 61% for the PCA of $\hat{\Gamma}_X(0)$ and 57% for the PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$. The proportions for the other PCs are quite similar for both PCA.

To classify the monitoring stations in clusters, we select for each PC the stations with the highest factor loading in absolute value. The coefficients in bold are larger than 0.37 in absolute value. Selecting these coefficients, we retain the cluster CL1 : VixCentro, Ibes and Cariacica for the 1st PC of $\hat{\Gamma}_X(0)$, the cluster CL2 : Laranjeiras and Carapina for the 2nd PC of $\hat{\Gamma}_X(0)$, the cluster CL3 : VVCentro for the 3rd PC of $\hat{\Gamma}_X(0)$, the cluster CL4 : Camburi and Sua for the 4th PC of $\hat{\Gamma}_X(0)$, and the cluster CL1 : Sua, VixCentro and Ibes for the 1st PC of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, the cluster CL2 : Laranjeiras, Carapina and Cariacica for the 2nd PC of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$, and the cluster CL3 : CL4 : Camburi and VVCentro for the 3rd and the 4th PC of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$. Note that four PCs are necessary in the PCA of $\hat{\Gamma}_X(0)$ to encompass the eight stations, while three PCs are enough in the PCA of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$.

Table 11: PCA of original and filtered PM_{10} concentrations.

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

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

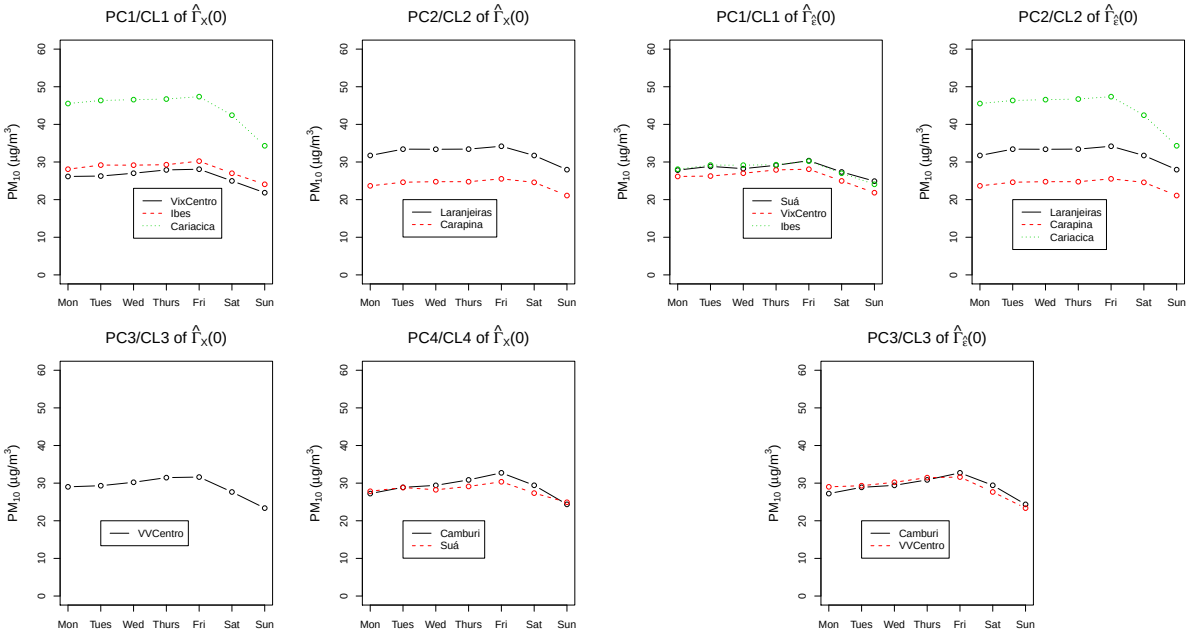
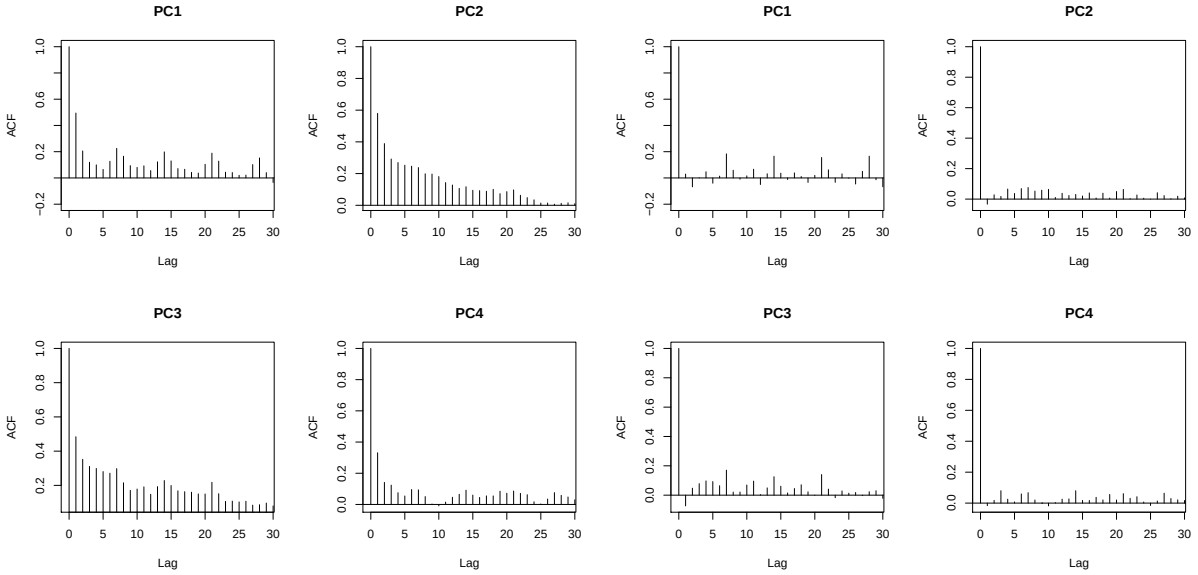


Figure 6: Average daily profile of PM_{10} concentrations grouped by the PC/CL category.

Following the same approach as Pires et al. (2008*a*,*b*), the number of monitoring stations that should be maintained among the eight corresponds to the maximum number of selected PCs. Based on the PCs of $\hat{\Gamma}_X(0)$, the four stations Ibes, Laranjeiras, VVCentro and Camburi are maintained, while the analysis of the PCs of $\hat{\Gamma}_{\hat{\varepsilon}}(0)$ leads to retain only the three stations, Ibes, Laranjeiras and Camburi. The equipment of the others stations may be moved to alternative areas of interest to cover a larger area of the GVR.

Figure 7 plots the sample autocorrelation functions of the PCs of original and filtered PM_{10} concentrations. Figure 7a) shows that the PCs are autocorrelated in the case of a correlated time series which is in accordance with the results of Proposition 1. Since the filtered time series $\hat{\varepsilon}_t$ is almost a white noise, the autocorrelations in Figure 7b) are very small.



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

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

5 Conclusion

This paper has investigated the effect of time-correlation on PCA technique. It was shown that the PCs are generally cross-correlated and that the variability of a linear process is larger than the variability of its innovation. Explicit calculations were presented in the case of simple linear parametric models. The theoretical results were illustrated empirically by means of Monte Carlo simulations.

It was found that PCA tool can still be used when a weak autocorrelation is present, since from a descriptive and an inferential point of view, the time-correlation does not affect drastically the final results. However, when a strong correlation structure is present, it is recommended to apply a linear filter for whitening the data before PCA.

An application of the proposed methodology to the identification of redundant air quality measurements was considered. It was pointed out that the data science practitioner must proceed with caution when interpreting the cluster analysis.

6 Acknowledgements

Some results of this paper appear in the PhD thesis of Zamprogno (2013) under the supervision of 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. V. A. Reisen was visiting CentraleSupélec (12/2018 to 01/2019). This author is indebted to CentraleSupélec for its financial support. This research was also partially supported by the iCODE Institute, research project of the IDEX Paris-Saclay, and by the Hadamard Mathematics LabEx (LMH) through the grant number ANR-11-LABX-0056-LMH in the “Programme des Investissements d’Avenir”.

References

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

Banerjee, S. & Roy, A. (2014), *Linear algebra and matrix analysis for statistics*, Chapman & Hall/CRC Texts in Statistical Science Series, CRC Press, Boca Raton, FL.

Borbon, A., Locoge, N., Veillerot, M., Galloo, J. C. & Guillermo, R. (2002), 'Characterisation of NMHCs in a french urban atmosphere: overview of the main sources', *the Science of the Total Environment* **292**, 177–191.

Brockwell, P. J. & Davis, R. A. (2006), *Time Series: Theory and Methods*, Springer Series in Statistics, 2nd edn, Springer Science, New York, NY.

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

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

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

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

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

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

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

Jaimungal, S. & Ng, E. K. H. (2007), 'Consistent functional PCA for financial time-series', *Tecnical-Report, University of Toronto* pp. 1–6.

Johnson, R. A. & Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, 6th edn, Prentice Hall.

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

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

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

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

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

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

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

Pires, J. C. M., Souza, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M. & Martins, F. G. (2008*a*), 'Management of air quality monitoring using principal component and cluster analysis — part I: SO_2 and PM_{10}', *Atmospheric Environment* **42**, 1249–1260.

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

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

Reisen, V. A., Sarnaglia, A. J. Q., Reis Jr, N. C., Lévy-Leduc, C. & Santos, J. M. (2014), 'Modeling and forecasting daily average PM_{10} concentrations by a seasonal long-memory model with volatility', *Environmental Modelling & Software* **51**, 286–95.

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

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

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

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

Shi, G.-L., Li, X., Yin-Chang, F., Wang, Y.-Q., Wu, J.-H., Jun, L. & Tan, Z. (2009), 'Combined source apportionment, using positive matrix factorization-chemical mass balance and principal component analysis/multiple linear regression-chemical mass balance models', *Atmospheric Environment* **43**, 2929–2937.

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

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

Statheropoulos, M., Vassiliadis, N. & Pappa, A. (1998), 'Principal component and canonical correlation analysis for examining air pollution and meteorological data', *Atmospheric Environment* **32**(6), 1087–1095.

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

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

Wang, H. & Shooter, D. (2004), 'Source apportionment of fine and coarse atmospheric particles in Auckland, New Zealand', *Science of the Total Environment* **340**, 189–198.

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

Wang, Y. & Pham, H. (2011*b*), 'Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components', *Int. J. Syst. Assur. Eng. Manag.* pp. 253–259.

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

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

An overview of robust spectral estimators

Valdério Anselmo Reisen[1,3], Céline Lévy-Leduc[2], Higor Henrique Aranda Cotta[1,3], Pascal Bondon[3], Marton Ispany[4], and Paulo Roberto Prezotti Filho[3,5]

[1] DEST and PPGEA-Universidade Federal do Espirito Santo-UFES, Vitória, Brazil
valderioanselmoreisen@gmail.com, valderio.reisen@ufes.br
[2] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France
[3] Laboratoire des Signaux et Systémes (L2S), CNRS-CentraleSupélec-Université Paris-Sud, Gif sur Yvette, France
[4] University of Debrecen, Debrecen, Hungary
[5] PPGEA-UFES and IFES, Brazil

Abstract. The periodogram function is widely used to estimate the spectral density of time series processes and it is well-known that this function is also very sensitive to outliers. In this context, this paper deals with robust estimation functions to estimate the spectral density of univariate and periodic time series with short and long-memory properties. The two robust periodogram functions discussed and compared here were previously explicitly and analytically derived in Fajardo et al. (2018), Reisen et al. (2017) and Fajardo et al. (2009) in the case of long-memory processes. The first two references introduce the robust periodogram based on $M-$regression estimator. The third reference is based on the robust autocovariance function introduced in Ma and Genton (2000) and studied theoretically and empirically in Lévy-Leduc et al. (2011). Here, the theoretical results of these estimators are discussed in the case of short and long-memory univariate time series and periodic processes. A special attention is given to the $M-$ periodogram for short-memory processes. In this case, Theorem 1 and Corollary 1 derive the asymptotic distribution of this spectral estimator. As the application of the methodologies, robust estimators for the parameters of AR, ARFIMA and PARMA processes are discussed. Their finite sample size properties are addressed and compared in the context of absence and presence of atypical observations. Therefore, the contributions of this paper come to fill some gaps in the literature of modeling univariate and periodic time series to handle additive outliers.

Time series, M-estimation, Q_N-estimation, long-memory, periodic processes, robustness.

1 Introduction

It is well known that outlying observations may completely destroy most of the standard estimators and several authors developed robust approaches in order to mitigate the impact of additive outliers, specially in time series models which is the process considered in this paper. However, most of the work is devoted to the robust estimation of the location, scale and other statistical tools. In this direction, the classical periodogram is the natural estimator of the spectral density of a time series and recent studies indicate that the periodogram is highly sensitive to the presence of outliers, and, thus, it becomes useless in any sub-sequential analysis. As a viable approach to attenuate this issue, the $M-$regression method applied to build alternative spectral estimators given in Fajardo et al. (2018) and Reisen et al. (2017)) and the Q_N-periodogram introduced in Fajardo et al. (2009) are some methodologies proposed recently in the literature of time series to handle additive outliers.

The M-periodogram is discussed in Fajardo et al. (2018) and Reisen et al. (2017) for the long-memory time series. The short-range process was still an open problem and is one main contribution of this paper. The asymptotic property of the $M-$periodogram is derived for the process which is identified to have short-memory property such as an ARMA model (Theorem 1). As a second contribution of this paper, the recent results given Fajardo et al. (2018) and Reisen et al. (2017), for long-memory model, are summarized and these methods are compared empirically

with Q_N-periodogram and the classical periodogram which is widely used in modelling time series data. Here, these methods are empirically studied and compared in time series with and without additive outliers with the aim to verify their finite sample size robustness properties, that is, to verify their capacity to accommodate the additive outlier's effect.

The use of $M-$ and Q_N-periodograms in periodic ARMA (PARMA) models is also discussed here in the context of handling atypical or aberrant observations (additive outliers). This becomes the third contribution of this paper.

This paper is organized as follows: Section 2 discusses robust periodograms based on $M-$regression method and Q_N function for short and long-memory time series. Section 3 presents some simulation results for the methods discussed in Section 2. Section 4 gives some applications of the alternative periodograms in short and long-memory and periodic processes.

2 Robust periodograms

Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a second order stationary process. Since this paper deals with short and long-memory processes, additional assumptions on the process $\{Y_t\}_{t\in\mathbb{Z}}$ will be given in the sequel of the paper. For a sample $\{Y_1, Y_2, ..., Y_N\}$, the classical periodogram function, at the Fourier frequency $\lambda_j = 2\pi j/N, j = 1, \ldots, [N/2]$, is defined as

$$I_N(\lambda_j) = \frac{1}{2\pi N}\left|\sum_{k=1}^{N} Y_k \exp(\mathrm{i}k\lambda_j)\right|^2 . \tag{1}$$

Next subsections deal with alternative periodogram functions which present similar performance (from theoretical and empirical meaning) to $I_N(\lambda)$, $\lambda \in (-\pi, \pi)$, but with robustness property against additive outliers and asymmetric and heavy-tail distributions.

2.1 *M*-periodogram

One alternative way to derive the periodogram function $I_N(\lambda_j)$ is based on the Least Square (LS) estimates of a two-dimensional vector $\boldsymbol{\beta}' = (\beta^{(1)}, \beta^{(2)})$ in the linear regression model

$$Y_i = c'_{Ni}\boldsymbol{\beta} + \varepsilon_i = \beta^{(1)}\cos(i\lambda_j) + \beta^{(2)}\sin(i\lambda_j) + \varepsilon_i \ , \ 1 \le i \le N, \ \boldsymbol{\beta} \in \mathbb{R}^2 \ , \tag{2}$$

where ε_i denotes the deviation of Y_i from $c'_{Ni}\boldsymbol{\beta}$ and $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) < \infty$. In the sequel (ε_i) is assumed to be a function of a stationary Gaussian process, see (10) for a precise definition. Then,

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\mathrm{Arg\ min}} \sum_{i=1}^{N}(Y_i - c'_{Ni}(\lambda_j)\boldsymbol{\beta})^2 \ , \tag{3}$$

where

$$c'_{Ni}(\lambda_j) = (\cos(i\lambda_j) \ \sin(i\lambda_j)) \ . \tag{4}$$

The solution of (3) is

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (C'C)^{-1}C'\mathbf{Y} \ , \tag{5}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)'$, C and $C'C$ are defined by

$$C = \begin{pmatrix} \cos(\lambda_j) & \sin(\lambda_j) \\ \cos(2\lambda_j) & \sin(2\lambda_j) \\ \vdots & \vdots \\ \cos(N\lambda_j) & \sin(N\lambda_j) \end{pmatrix} \tag{6}$$

and

$$C'C = \begin{pmatrix} \sum_{k=1}^{N}\cos(k\lambda_j)^2 & \sum_{k=1}^{N}\cos(k\lambda_j)\sin(k\lambda_j) \\ \sum_{k=1}^{N}\cos(k\lambda_j)\sin(k\lambda_j) & \sum_{k=1}^{N}\sin(k\lambda_j)^2 \end{pmatrix} = \frac{N}{2}\,\mathrm{Id}_2 \tag{7}$$

where Id_2 is the identity matrix 2 by 2. Hence,

$$\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = \frac{2}{N} C'\mathbf{Y} = \frac{2}{N}\left(\sum_{k=1}^{N} Y_k \cos(k\lambda_j) \quad \sum_{k=1}^{N} Y_k \sin(k\lambda_j)\right)' = (\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j), \hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))'. \quad (8)$$

In view of (1),

$$I_N(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j)\|^2 = \frac{N}{8\pi}\left((\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j))^2 + (\hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))^2\right) =: I_N^{\mathrm{LS}}(\lambda_j)\ , \quad (9)$$

where $\|\cdot\|$ denotes the classical Euclidean norm and $\hat{\boldsymbol{\beta}}_N^{\mathrm{LS}}(\lambda_j) = (\hat{\beta}_N^{\mathrm{LS},(1)}(\lambda_j), \hat{\beta}_N^{\mathrm{LS},(2)}(\lambda_j))'$ is the least square estimates of $\boldsymbol{\beta}' = (\beta^{(1)}, \beta^{(2)})$ see, for example, Fajardo et al. (2018) and Reisen et al. (2017) and references therein. Note that $I_N(\lambda_j)$ (9) can be derived for different choices of ε_i, $i = 1, \ldots, N$.

It is supposed here that

$$\varepsilon_i = G(\eta_i)\ . \quad (10)$$

In (10), G is a non null real-valued and skew symmetric measurable function (*i.e.* $G(-x) = -G(x)$, for all x) and $(\eta_i)_{i\geq 1}$ is a stationary Gaussian process with zero mean and unit variance. Additional assumptions of $(\eta_i)_{i\geq 1}$ will be given in the sequel of the paper.

Let $\psi(.)$ be a function satisfying the following assumptions.

(**A1**) $0 < \mathbb{E}[\psi^2(\varepsilon_1)] < \infty$.

(**A2**) The function ψ is absolutely continuous with its almost everywhere derivative ψ' satisfying $\mathbb{E}[|\psi'(\varepsilon_1)|] < \infty$ and such that the function $z \mapsto \mathbb{E}[|\psi'(\varepsilon_1 - z) - \psi'(\varepsilon_1)|]$ is continuous at zero.

(**A3**) ψ is nondecreasing, $\mathbb{E}[\psi'(\varepsilon_1)] > 0$ and $\mathbb{E}[\psi'(\varepsilon_1)^2] < \infty$.

(**A4**) ψ is skew symmetric, *i.e.* $\psi(-x) = -\psi(x)$, for all x.

It is now introduced the M-periodogram based on the M-estimator $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}$ of the parameter $\boldsymbol{\beta}$ defined in Equation (2). The M-estimator $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}} = (\hat{\beta}_N^{(1)}, \hat{\beta}_N^{(2)})'$ is defined as the solution (t_1, t_2) of

$$\sum_{i=1}^{N} \cos(i\lambda_j)\, \psi(Y_i - \cos(i\lambda_j)t_1) = 0 \text{ and } \sum_{i=1}^{N} \sin(i\lambda_j)\, \psi(Y_i - \sin(i\lambda_j)t_2) = 0. \quad (11)$$

$\hat{\beta}_N^{(1)}$ and $\hat{\beta}_N^{(2)}$ can be also seen as the minimizers with respect to t_1 and t_2, respectively, of

$$\left|\sum_{i=1}^{N} \cos(i\lambda_j)\, \psi(Y_i - \cos(i\lambda_j)t_1)\right| \text{ and } \left|\sum_{i=1}^{N} \sin(i\lambda_j)\, \psi(Y_i - \sin(i\lambda_j)t_2)\right|, \quad (12)$$

where ψ satisfies the same assumptions as in Koul and Surgailis (2000). By analogy to (9), the robust periodogram $I_N^M(\lambda_j)$ at $\lambda_j = 2\pi j/N, j = 1, \ldots, [N/2]$, is defined by

$$I_N^M(\lambda_j) = \frac{N}{8\pi}\|\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}(\lambda_j)\|^2 = \frac{N}{8\pi}\left((\hat{\beta}_N^{(1)}(\lambda_j))^2 + (\hat{\beta}_N^{(2)}(\lambda_j))^2\right)\ . \quad (13)$$

2.1.1 $M-$Periodogram in short-memory processes In this subsection the asymptotic properties of $\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}$ are established in the short-range dependence framework. For this, the following assumptions are introduced. This result helps to establish the theoretical properties of the robust periodogram I_N^M given in Corollary 1.

(**A5**) Let η_t, $t \in \mathbb{Z}$, be i.i.d. standard Gaussian random variables and let a_j be real numbers such that $\sum_{j\geq 0}|a_j| < \infty$ and $a_0 = 1$. Then,

$$\varepsilon_i = \sum_{j\geq 0} a_j \eta_{i-j}.$$

(**A6**) ψ is the Huber function that is $\psi(x) = \max[\min(x, c), -c]$, for all x in $\mathbb{R}$, where c is a positive constant.

Theorem 1. *Assume that (A5) and (A6) hold and that $\boldsymbol{\beta} = \mathbf{0}$ in (2) so that $Y_i = \varepsilon_i$. Then, for any fixed j, $\hat{\boldsymbol{\beta}}_N^M$ defined by (12) satisfies*

$$\sqrt{\frac{N}{2}}(F(c) - F(-c))\hat{\boldsymbol{\beta}}_N^M(\lambda_j) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Delta}^{(j)}\right), \ N \to \infty ,$$

where F is the c.d.f. of ε_1 and

$$\boldsymbol{\Delta}^{(j)} = \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\} \begin{pmatrix} \cos(k\lambda_j) & \sin(k\lambda_j) \\ -\sin(k\lambda_j) & \cos(k\lambda_j) \end{pmatrix}.$$

Theorem 1 is proved in Section 5.

Corollary 1. *Under the assumptions of Theorem 1, $I_N^M(\lambda_j)$ defined in (13) satisfies for any fixed j,*

$$I_N^M(\lambda_j) \xrightarrow{d} \frac{X^2 + Y^2}{4\pi(F(c) - F(-c))^2}, \quad \text{as } N \to \infty ,$$

where

$$X \sim \mathcal{N}\left(0, \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\cos(k\lambda_j)\right), \ Y \sim \mathcal{N}\left(0, \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\cos(k\lambda_j)\right)$$

and

$$\mathrm{Cov}(X, Y) = \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\sin(k\lambda_j).$$

The proof of Corollary 1 is a straightforward consequence of Theorem 1 and (13).

2.1.2 *M*-periodogram for long-memory processes Now, consider the following assumption for $(\eta_i)_{i\geq 1}$ in the case of long-memory process. The results in this subsection are well detailed in Fajardo et al. (2018).

(**A7**) $(\eta_i)_{i\geq 1}$ is a stationary zero-mean Gaussian process with covariances $\rho(k) = \mathbb{E}(\eta_1\eta_{k+1})$ satisfying:

$$\rho(0) = 1 \text{ and } \rho(k) = k^{-D}L(k), \ 0 < D < 1 ,$$

where the function L is slowly varying at infinity and is positive for large k. Recall that a slowly varying function $L(x)$, $x > 0$ is such that $L(xt)/L(x) \to 1$, as $x \to \infty$ for any $t > 0$. Constants and logarithms are example of slowly varying functions.
Moreover, the spectral density f of $(\eta_i)_{i\geq 1}$ can be expressed as:

$$f(\lambda) = |1 - \exp(-\mathrm{i}\lambda)|^{-2d} f^*(\lambda) , \tag{14}$$

where $d \in (0, 1/2)$ and f^* is an even, positive, continuous function on $(-\pi, \pi]$, bounded above and bounded away from zero.

Note that

$$D = 1 - 2d , \tag{15}$$

where D is defined in Assumption (A7) and d is the standard long-memory parameter notation given in the literature of long-memory models. The fact that $(\eta_i)_{i\geq 1}$ is required to satisfy (A7) essentially means that both $L(x)$, $x \geq 1$ and $f^*(\lambda)$, λ in $(-\pi, \pi]$ satisfy some smoothness properties.

Theorem 2. *Assume that (A7), (A1), (A2), (A3) and (A4) hold and that $\boldsymbol{\beta} = 0$ in (2) so that $Y_i = \varepsilon_i$. Then, for any fixed j, $\hat{\boldsymbol{\beta}}_N^M(\lambda_j)$ defined by (12) satisfies*

$$\sqrt{\frac{N}{2}}\hat{\boldsymbol{\beta}}_N^M(\lambda_j) = \frac{J_1}{\mathbb{E}[\psi'(\varepsilon_1)]}\left\{\sqrt{\frac{2}{N}}\sum_{i=1}^{N}\begin{pmatrix}\cos(i\lambda_j)\\ \sin(i\lambda_j)\end{pmatrix}\eta_i\right\} + o_p(N^{(1-D)/2})\ ,\ \textit{as } N\to\infty\ , \tag{16}$$

where $J_1 = \mathbb{E}[\psi(G(\eta))\eta] \neq 0$, η being a standard Gaussian random variable and $D = 1-2d$. Moreover,

$$N^{D/2}\hat{\boldsymbol{\beta}}_N^M(\lambda_j) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{J_1^2}{(\mathbb{E}[\psi'(\varepsilon_1)])^2}\widetilde{\Gamma}\right)\ ,\ N\to\infty\ , \tag{17}$$

where

$$\widetilde{\Gamma} = \lim_{N\to\infty}\frac{4}{N^{2-D}}\sum_{1\le k,\ell\le N} c_{Nk}(\lambda_j)c_{N\ell}^T(\lambda_j)\rho(k-\ell) \tag{18}$$

$$= 8\pi\times(2\pi j)^{-2d}f^*(0)\begin{pmatrix}\mathcal{L}_1 & 0\\ 0 & \mathcal{L}_2\end{pmatrix}\ . \tag{19}$$

In Relation (18), the vector $c_{Nk}(\lambda_j)$ is defined in (4),

$$\mathcal{L}_1 = \frac{1}{\pi}\int_{\mathbb{R}}\frac{\sin^2(\lambda/2)}{(2\pi j-\lambda)^2}\left|\frac{\lambda}{2\pi j}\right|^{-2d}\mathrm{d}\lambda - \frac{1}{\pi}\int_{\mathbb{R}}\frac{\sin^2(\lambda/2)}{(2\pi j-\lambda)(2\pi j+\lambda)}\left|\frac{\lambda}{2\pi j}\right|^{-2d}\mathrm{d}\lambda\ , \tag{20}$$

and

$$\mathcal{L}_2 = \frac{1}{\pi}\int_{\mathbb{R}}\frac{\sin^2(\lambda/2)}{(2\pi j-\lambda)^2}\left|\frac{\lambda}{2\pi j}\right|^{-2d}\mathrm{d}\lambda + \frac{1}{\pi}\int_{\mathbb{R}}\frac{\sin^2(\lambda/2)}{(2\pi j-\lambda)(2\pi j+\lambda)}\left|\frac{\lambda}{2\pi j}\right|^{-2d}\mathrm{d}\lambda\ . \tag{21}$$

Corollary 2. *Under the assumptions of Theorem 2, the periodogram I_N^M defined in (13) satisfies*

$$N^{D-1}I_N^M(\lambda_j) \xrightarrow{d} (Z_1^2+Z_2^2)\ ,\ \textit{as } N\to\infty\ , \tag{22}$$

where (Z_1, Z_2) is a zero-mean uncorrelated Gaussian vector with covariance matrix equal to

$$\frac{J_1^2}{8\pi(\mathbb{E}[\psi'(\varepsilon_1)])^2}\widetilde{\Gamma}\ , \tag{23}$$

with $\widetilde{\Gamma}$ defined in (18).

Theorem 2 and Corollary 2 are proved in Fajardo et al. (2018).

2.2 Q_N-periodogram

Another possible approach to obtain the classical periodogram (1) is to write it in terms of the sample autocovariance function

$$I_N(\lambda_j) = \frac{1}{2\pi}\sum_{h=-(N-1)}^{N-1}\widehat{\gamma}(h)\cos(h\lambda_j), \tag{24}$$

where $\lambda_j = 2\pi j/N, j = 1,\ldots,[N/2]$ and $\widehat{\gamma}(h)$ is the classical sample autocovariance function for a sample $\{Y_1,...,Y_N\}$.

A straightforward approach to robustify 24 is to plug in a robust autocovariance function replacing the classical one. This methodology is now addressed.

For a sample $x_1,...,x_N$ Rousseeuw and Croux (1993) proposed a robust scale estimator function $Q_N(\cdot)$ which is based on the τth order statistic of $\binom{N}{2}$ distances $\{|x_j - x_k|, j<k\}$, and can be written as

$$Q_N(x) = \kappa\times\{|x_j-x_k|; j<k\}_{(\tau)}, \tag{25}$$

where κ is a constant used to guarantee consistency ($\kappa = 2.2191$ for the Gaussian distribution) and $\tau = \lfloor (\binom{N}{2} + 2)/4 \rfloor + 1$. The above function can be evaluated using the algorithm proposed by Croux and Rousseeuw (1992), which is computationally efficient.

Based on $Q_N(\cdot)$, Ma and Genton (2000) proposed a highly robust estimator for the autocovariance function:

$$\widehat{\gamma}_{Q_N}(h) = \frac{1}{4}\left[Q^2_{N-h}(\mathbf{u}+\mathbf{v}) - Q^2_{N-h}(\mathbf{u}-\mathbf{v})\right], \tag{26}$$

where $\mathbf{u}$ and $\mathbf{v}$ are vectors containing the initial $N-h$ and the final $N-h$ observations of $x_1, ..., x_N$, respectively. The robust estimator for the autocorrelation function is

$$\widehat{\rho}_{Q_N}(h) = \frac{Q^2_{N-h}(\mathbf{u}+\mathbf{v}) - Q^2_{N-h}(\mathbf{u}-\mathbf{v})}{Q^2_{N-h}(\mathbf{u}+\mathbf{v}) + Q^2_{N-h}(\mathbf{u}-\mathbf{v})}. \tag{27}$$

It can be shown that $|\widehat{\rho}_{Q_N}(h)| \leq 1$ for all h.

Now, returning to (24), the robust Q_N-periodogram for a sample $\{Y_1, ..., Y_N\}$ is defined by

$$I_N^{Q_N}(\lambda_j) = \frac{1}{2\pi}\sum_{h=-(N-1)}^{N-1} \widehat{\gamma}_{Q_N}(h)\cos(h\lambda_j), \tag{28}$$

where $\lambda_j = 2\pi j/N, j = 1, \ldots, [N/2]$.

The theoretical properties of $I_N^{Q_N}$ are still under study. Therefore, in the sequel, the asymptotic properties of $\hat{\gamma}_{Q_N}$ are summarized for short and long memory processes. These are well detailed in Lévy-Leduc et al. (2011).

2.2.1 Main asymptotic results for short memory process In the short-memory scenario, the process under study $(Y_i)_{i\geq 1}$ satisfies the following assumption (see, also, Lévy-Leduc et al. (2011)):

(**A8**) $(Y_i)_{i\geq 1}$ is a stationary zero-mean Gaussian process with autocovariance sequence $\gamma(h) = \mathbb{E}(Y_1 Y_{h+1})$ satisfying:

$$\sum_{h\geq 1} |\gamma(h)| < \infty\ .$$

Theorem 3. *Assume that (A8) holds and let h be a non negative integer. Then, the autocovariance estimator $\widehat{\gamma}_{Q_N}(h)$ satisfies the following Central Limit Theorem:*

$$\sqrt{N}\left(\widehat{\gamma}_{Q_N}(h) - \gamma(h)\right) \xrightarrow{d} \mathcal{N}(0, \check{\sigma}^2_h), N \to \infty,$$

where

$$\check{\sigma}^2(h) = \mathbb{E}[\zeta^2(Y_1, Y_{1+h})] + 2\sum_{k\geq 1}\mathbb{E}[\zeta(Y_1, X_{1+h})\zeta(Y_{k+1}, Y_{k+1+h})]\ , \tag{29}$$

and the function ζ is defined by

$$\zeta : (x,y) \mapsto$$
$$\left\{(\gamma(0)+\gamma(h))\ \mathrm{IF}\left(\frac{x+y}{\sqrt{2(\gamma(0)+\gamma(h))}}, Q, \Phi\right) - (\gamma(0)-\gamma(h))\ \mathrm{IF}\left(\frac{x-y}{\sqrt{2(\gamma(0)-\gamma(h))}}, Q, \Phi\right)\right\}\ . \tag{30}$$

where IF *is defined by*

$$\mathrm{IF}(x, Q, \Phi) = \kappa\left(\frac{1/4 - \Phi(x+1/\kappa) + \Phi(x-1/\kappa)}{\int_{\mathbb{R}}\phi(y)\phi(y+1/\kappa)\mathrm{d}y}\right)\ , \tag{31}$$

where Φ and ϕ denote the c.d.f. and p.d.f. of a standard Gaussian random variable, respectively with κ defined in (25).

Theorem 3 is proved in Lévy-Leduc et al. (2011).

2.2.1 Main asymptotic results for long-memory process The following results concern the robust autocovariance function for long-memory process see, also, Lévy-Leduc et al. (2011).

(**A9**) $(Y_i)_{i\geq 1}$ is a stationary zero-mean Gaussian process with autocovariance $\gamma(h) = \mathbb{E}(Y_1 Y_{h+1})$ satisfying:

$$\gamma(h) = h^{-D} L(h),\ 0 < D < 1\ ,$$

where L is slowly varying at infinity and is positive for large h. Note that, as previously stated, $D = 1 - 2d$.

Theorem 4. *Assume that (A9) holds and that L has three continuous derivatives. Assume also that $L_i(x) = x^i L^{(i)}(x)$ satisfy: $L_i(x)/x^\epsilon = O(1)$, for some ϵ in $(0, D)$, as x tends to infinity, for all $i = 0, 1, 2, 3$, where $L^{(i)}$ denotes the ith derivative of L. Let h be a non negative integer. Then, $\widehat{\gamma}_{Q_N}(h)$ satisfies the following limit theorems as N tends to infinity.*

(i) If $D > 1/2$,

$$\sqrt{N}\left(\widehat{\gamma}_{Q_N}(h) - \gamma(h)\right) \xrightarrow{d} \mathcal{N}(0, \check{\sigma}^2(h))\ ,$$

where

$$\check{\sigma}^2(h) = \mathbb{E}[\zeta^2(Y_1, Y_{1+h})] + 2\sum_{k\geq 1} \mathbb{E}[\zeta(Y_1, Y_{1+h})\zeta(Y_{k+1}, Y_{k+1+h})]\ ,$$

ζ being defined in (30).

(ii) If $D < 1/2$,

$$\beta(D)\frac{N^D}{\widetilde{L}(N)}\left(\widehat{\gamma}_{Q_N}(h) - \gamma(h)\right) \xrightarrow{d} \frac{\gamma(0) + \gamma(h)}{2}(Z_{2,D}(1) - Z_{1,D}(1)^2)$$

where $\beta(D) = \mathrm{B}((1 - D)/2, D)$, B denotes the Beta function, the processes $Z_{1,D}(\cdot)$ and $Z_{2,D}(\cdot)$ are defined as follows:

$$Z_{1,D}(t) = \int_{\mathbb{R}} \left[\int_0^t (u - x)_+^{-(D+1)/2} \mathrm{d}u\right] \mathrm{d}B(x), \quad 0 < D < 1\ , \tag{32}$$

$$Z_{2,D}(t) = \int_{\mathbb{R}^2}{}' \left[\int_0^t (u - x)_+^{-(D+1)/2}(u - y)_+^{-(D+1)/2} du\right] \mathrm{d}B(x)\mathrm{d}B(y),\ 0 < D < 1/2\ , \tag{33}$$

and

$$\widetilde{L}(N) = 2L(N) + L(N + h)(1 + h/N)^{-D} + L(N - h)(1 - h/N)^{-D}\ , \tag{34}$$

where B is the standard Brownian motion. The symbol $\int'$ means that the domain of integration excludes the diagonal.

Theorem 4 is proved in Lévy-Leduc et al. (2011).

3 Monte Carlo simulation

In this section, small sample size experiments are conducted with the aim to clarify the empirical performance of the spectral estimates discussed previously in a different context such as time series with additive outliers. Based on this, some standard questions, such as (1) what is the best method to be used in a real application? (2) which method (if any) should be considered when dealing with outliers? (3) Does the large observation (if any) make similar outlier´s effect on the statistical time series modelling functions, that is, on the ACF and periodogram functions? among others, are expected to be answered or, at least, clarified.

Let $\{X_t\}_{t=1,...,N}$ be a sample from a Gaussian second order stationary process and let $\{Y_t\}_{t=1,...,N}$ be a sample of the process defined by

$$Y_t = X_t + \omega W_t \tag{35}$$

where the parameter ω represents the magnitude of the outlier, and W_t is a random variable with probability distribution

$$\mathbb{P}(W_t = -1) = \mathbb{P}(W_t = 1) = \delta/2 \text{ and } \mathbb{P}(W_t = 0) = 1 - \delta\,,$$

where $\mathbb{E}[W_t] = 0$ and $\mathbb{E}[W_t^2] = \mathrm{Var}(W_t) = \delta$. Note that (35) is based on the parametric models proposed by Fox (1972). W_t is the product of $Bernoulli(\delta)$ and *Rademacher* random variables; the latter equals 1 or -1, both with probability 1/2. X_t and W_t are independent random variables. Note that, if $\omega = 0.0$ $\{Y_t\}$ is an outlier free time series.

In order to compare the performance of M- and Q_N-periodogram, a Monte Carlo investigation was carried out under different contamination scenarios. For the simulations, the number of replications was 5000, the samples $\{X_t\}$ of size $N = 500$ were generated according to a model autocorrelation structure, which is given in what follows, and the contaminated data Y_t were generated from (35) with $\delta = 0.01$ for magnitudes $\omega = 0$ (no outliers) and 10.

The comparison between the methods is performed by estimating α in the linear regression $\log(I(\lambda_j)) \simeq const + \alpha \log(\lambda_j) + E_j$, $j = 1, \ldots, N^{0.7}$, where $I(.)$ is either $I_N(.)$, $I_N^M(.)$ or $I_N^{Q_N}(.)$. The data were generated based on

$$X_t = (1 - B)^{-d} Z_t = \sum_{j \geq 0} \frac{\Gamma(j + d)}{\Gamma(j + 1)\Gamma(d)} \epsilon_{t-j}\,, \tag{36}$$

where ϵ_t is an AR(1) model, that is, $\epsilon_t = \phi\epsilon_{t-1} + \eta_t$, where η_t, $t = 1, ..., N$, are i.i.d. standard Gaussian random variables.

In the finite sample size investigation, the model correlation structures are divided in two cases:

1. An AR(1) model with $\phi = 0.6$ and $d = 0$.
2. An ARFIMA$(0, d, 0)$ model with $d = 0.3$.

Figure 1 displays the plots of the empirical densities of $\hat{\alpha}_{I_N}$, $\hat{\alpha}_{I_N^M}$ and $\hat{\alpha}_{I_N^{Q_N}}$ for the case of AR(1) models without contamination ($\omega = 0$). Although, $\hat{\alpha}_{I_N^M}$ has a slight better performance than $\hat{\alpha}_{I_N^{Q_N}}$, that is, the first method and the classical periodogram presented very close densities, all the methods provided similar results showing that, even for small sample sizes, the empirical density is very close which corroborate the theoretical results discussed previously. Based on the asymptotic theory and the empirical results all three methods can be used to estimate the spectral density of a time series when there is no contamination of additive outliers. This opens an important contribution in the context that alternative spectral estimators such as I_N^M and $I_N^{Q_N}$ can be used instead of the classical periodogram I_N in the step procedure for modelling time series data. For example, these estimators can be an alternative tools to be used in the Whittle function to obtain the parameter estimates. This will be also discussed in what follows. Note that, the disadvantage of $I_N^{Q_N}$ over I_N^M and I_N is that the ACF using $Q_N(.)$ does not have the positive definite property.

When the data is contaminated with additive outliers the scenario changes significantly. As well known, the periodogram, which depends on the classical autocovariance, is corrupted by the outliers. Therefore, the alternative methods are almost unaffected. This is displayed in Figure 2 in which $\omega = 10$ and $\delta = 0.01$. The empirical density of $\hat{\alpha}_{I_N}$ is shifted to the right side which is an expected result since the variance increases with outliers. The empirical densities of $\hat{\alpha}_{I_N^M}$ and $\hat{\alpha}_{I_N^{Q_N}}$ remain almost unchangeable.

In the case of long-memory process, the empirical density plots are given in Figures 3 and 4 for non-contaminated and contaminated time series, respectively. Similar conclusions of the AR case

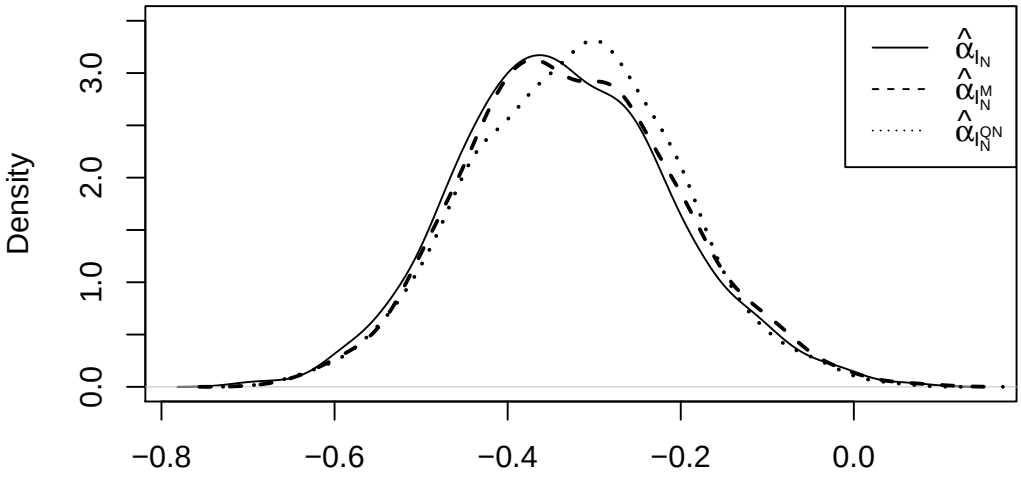


Fig. 1. Densities of $\hat{\alpha}_{I_N}$, $\hat{\alpha}_{I_N^M}$ and $\hat{\alpha}_{I_N^{QN}}$ for AR(1) models with $\phi = 0.6$ and $\omega = 0$.

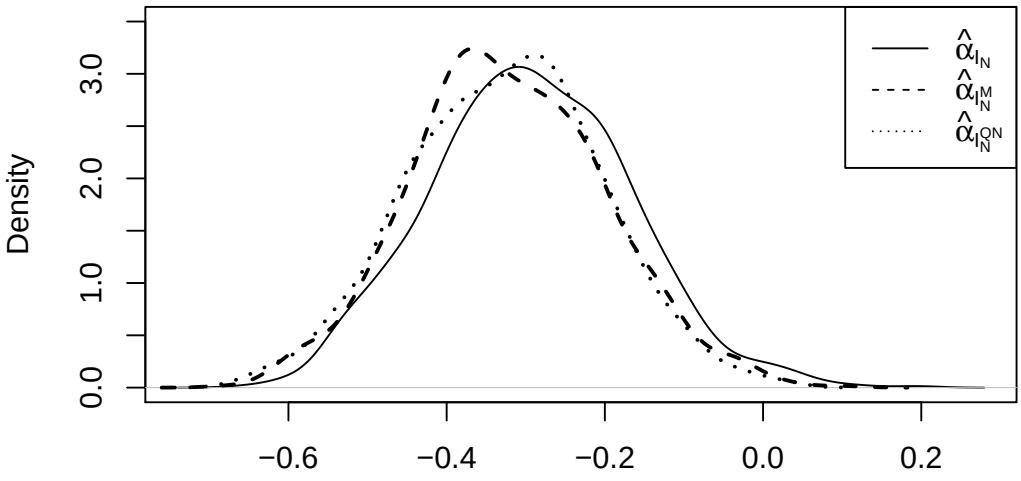


Fig. 2. Densities of $\hat{\alpha}_{I_N}$, $\hat{\alpha}_{I_N^M}$ and $\hat{\alpha}_{I_N^{QN}}$ for AR(1) models with $\phi = 0.6$, $\delta = 0.01$ and $\omega = 10$.

are drawn. That is, in the uncontaminated scenarios, all three methods displayed similar densities although the method M and the classical one (periodogram) are very close. In the contaminated case, the classical one is totally affected by the additive outliers. Reinforcing that the ACF using Q_N does not have the positiveness property.

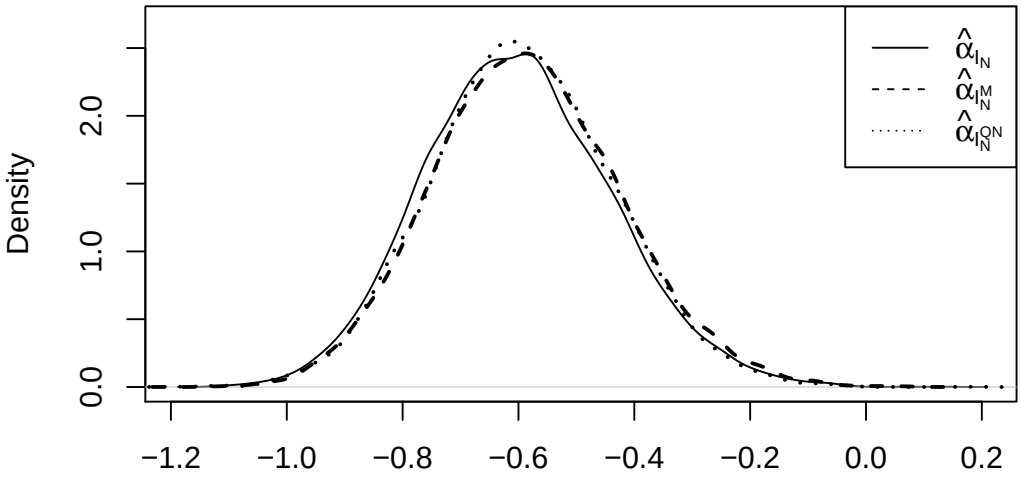


Fig. 3. Densities of $\hat{\alpha}_{I_N}$, $\hat{\alpha}_{I_N^M}$ and $\hat{\alpha}_{I_N^{Q_N}}$ when $d = 0.3$, $N = 500$ and $\omega = 0$.

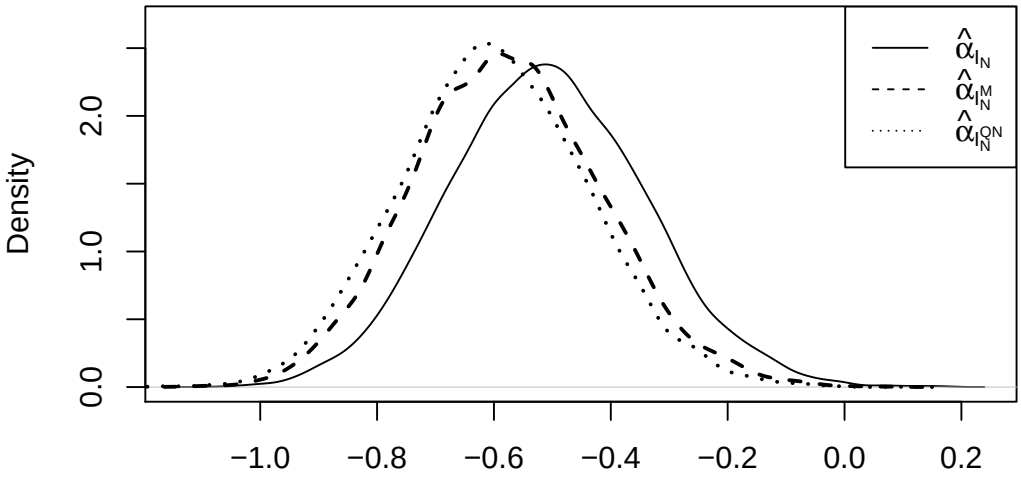


Fig. 4. Densities of $\hat{\alpha}_{I_N}$, $\hat{\alpha}_{I_N^M}$ and $\hat{\alpha}_{I_N^{Q_N}}$ when $d = 0.3$, $N = 500$, $\delta = 0.01$ and $\omega = 10$.

4 Applications of M and Q_N-periodograms

4.1 Robust estimation of the fractional parameter

Based on the theoretical results discussed previously, this section introduces some applications related to the use of $M-$regression and Q_N estimation functions. The application is divided in two cases: (a) Estimation of the fractional parameter d in long-memory processes; (b) Estimation in periodic AR (PAR) processes. Some finite sample size investigation is also addressed in the context of time series with and without outliers.

(a) *Estimation of the fractional parameter in long-memory process*

The estimation methods of the fractional parameter d discussed here are derived from the well-known semi-parametric regression method (GPH) originally proposed by Geweke and Porter-Hudak (1983). The regression estimation methods based on I_N^M and $I_N^{Q_N}$ were previously introduced in Reisen et al. (2017) and Fajardo et al. (2009), respectively, papers where the reader will find more details related to theoretical and empirical results of these estimation methodologies.

(**A10**) $(\varepsilon_i)_{i\geq 1}$ is a stationary mean-zero Gaussian process with spectral density given in Assumption (A7).

For estimating the fractional parameter d of long-memory processes having their spectral density satisfying (14), it is usual to use the standard GPH (Geweke and Porter-Hudak (1983)) estimator defined in the following. This estimator is motivated heuristically by starting from

$$\log(f(\lambda_j)) = -2d\log(|2\sin(\lambda_j/2)|) + \log(f^*(\lambda_j)) = -2dX_j + \log(f^*(\lambda_j)) \\ = \log(f_0^\star) - 2dX_j + \log(f_j^\star/f_0^\star), \quad (37)$$

where $X_j = \log|2\sin(\lambda_j/2)|$ and $f_j^\star = f^\star(\lambda_j)$. If

$$\varepsilon_j^R = \log\left(\frac{I_N(\lambda_j)}{f(\lambda_j)}\right), \tag{38}$$

then

$$\log(I_N(\lambda_j)) = \varepsilon_j^R + \log(f(\lambda_j)),$$

and, by (37),

$$\log(I_N(\lambda_j)) = \log(f_0^\star) - 2dX_j + \log(f_j^\star/f_0^\star) + \varepsilon_j^R. \tag{39}$$

The GPH estimator is given by

$$\hat{d}^{\mathrm{GPH}} = \frac{-0.5\sum_{j=1}^{m_N}(X_j - \bar{X})\log(I_N^{\mathrm{LS}}(\lambda_j))}{\sum_{k=1}^{m_N}(X_k - \bar{X})^2}, \tag{40}$$

where $X_j = \log|2\sin(\lambda_j/2)|$, $\bar{X} = \sum_{j=1}^{m_N} X_j/m_N$, $I_N^{\mathrm{LS}}(\lambda_j)$ is defined in (9) and m_N is a function of N.

Based on the above discussion, one way to define a $M-$regression estimator of d consists in replacing I_N^{LS} in (40) by I_N^M defined in (13):

$$\hat{d}^M = \frac{-0.5\sum_{j=1}^{m_N}(X_j - \bar{X})\log(I_N^M(\lambda_j))}{\sum_{k=1}^{m_N}(X_k - \bar{X})^2}, \tag{41}$$

where $X_j = \log|2\sin(\lambda_j/2)|$, $\bar{X} = \sum_{j=1}^{m_N} X_j/m_N$ and m_N is a function of N which is specified in Theorem 5.

The theoretical properties of $\hat{d}^{\mathrm{M}}$ are established under the following assumptions. The random process (ε_j) is obtained through a moving average process:

$$\varepsilon_j = \sum_{k\le j} a_{j-k}\zeta_k\ , \quad a_j = L(j)j^{-(1+D)/2}\ ,\ j \ge 1\ , \tag{42}$$

for some D in $(0,1)$, where $L(\cdot)$ is a positive slowly varying function at infinity and where the random variables ζ_k are i.i.d. with zero mean and variance 1. It is assumed that the distribution of ζ_0 satisfies

$$\left|\mathbb{E}(\mathrm{e}^{\mathrm{i}u\zeta_0})\right| \le C(1+|u|)^{-\delta}\ ,\ u \in \mathbb{R}\ . \tag{43}$$

where $C < \infty$ and $\delta > 0$ are constants. Note that, Conditions (42) and (43) imply that the cumulative distribution function F_{ε_0} of ε_0 is infinitely boundedly differentiable, see Koul and Surgailis (2000).

Theorem 5. *Let $Y_i = \varepsilon_i$, for all i in $\{1,\ldots,N\}$, where ε_i satisfy (42) and (A10). Assume that $1/D$ is not an integer and that $\boldsymbol{\beta} = \mathbf{0}$ in (2). Assume moreover that $\mathbb{E}(\zeta_0^{4\vee 2k^\star}) < \infty$, where $k^\star = [1/D]$, ζ_0 is defined in (42) and satisfies (43), $\nu_1 \neq 0$, $\nu_2 = 0$ and $\nu_3 \neq 0$, where the ν_k are defined by*

$$\nu_k = \int_0^\infty \psi(y)\left[1-(-1)^k\right]f^{(k)}(y)\mathrm{d}y, \quad \textit{for all integer } k \ge 0\ , \tag{44}$$

where ψ is the Huber function. Then, if $1/3 < D < 1$,

$$\sqrt{m_N}(\hat{d}^M - d) \xrightarrow{d} \mathcal{N}(0, \pi^2/24), \ \textit{as } N \to \infty, \tag{45}$$

where $\hat{d}^M$ is defined in (41) and $m_N = N^\beta$ with $0 < \beta < (1-D)/3$.

This result is proved in Reisen et al. (2017).

Another way of defining a robust estimator of d is to consider:

$$\hat{d}^{Q_N} = \frac{-0.5\sum_{j=1}^{m_N}(X_j - \bar{X})\log(I_N^{Q_N}(\lambda_j))}{\sum_{k=1}^{m_N}(X_k - \bar{X})^2}, \tag{46}$$

where $X_j = \log|2\sin(\lambda_j/2)|$, $\bar{X} = \sum_{j=1}^{m_N} X_j/m_N$, $I_N^{Q_N}(\lambda_j)$ is defined in (28) and m_N is a function of N. For further information, see Fajardo et al. (2009). The asymptotic property of $\hat{d}^{Q_N}$ is still an open problem, however, the empirical results given in Fajardo et al. (2009) support the use of this method under time series with and without outliers. The performance of fractional estimators $\hat{d}^{GPH}$, $\hat{d}^{M}$ and $\hat{d}^{Q_N}$ is the motivation of the next subsection for long-memory time series with and without additive outliers.

4.1.1- Finite sample size investigation

In this subsection, the numerical experiments were carried out in accordance with the model of Section 3. For the simulations, $N = 500$, $\omega = 10$ and $\delta = 0.01$ for 5000 replications. The results are displayed in Figures 5, 6 and Table 1. Since there is not short-memory component in the model m_N was fixed at $N^{0.7}$ for all tree methods.

Figure 5 presents the boxplots with the results of $\hat{d}_{GPH}$, $\hat{d}_M$ and $\hat{d}_{Q_N}$ estimators for the uncontaminated scenario. $\hat{d}_M$ and $\hat{d}_{Q_N}$ seem to present positive bias and, surprisingly, $\hat{d}_{Q_N}$ displays smaller deviation. However, in general, all methods perform similarly, i.e., all estimation methods leaded to comparable estimates close to the real values of d.

Figure 6 displays the boxplots of $\hat{d}_{GPH}$, $\hat{d}_M$ and $\hat{d}_{Q_N}$ when the series has outliers. As can be perceived from the boxplots, the GPH estimator is clearly affected by additive outliers while the robust ones keep almost the same picture as the one of the non-contaminated scenario, except that the bias of $\hat{d}_{Q_N}$ becomes negative, that is, this estimator tends to overestimate the true parameter.

The empirical mean, bias and mean square root are displayed in Table 1. This numerically corroborates the results discussed based on Figures 5, 6, that is, the estimators have similar performance in the absence of outliers in the data. While the performance of $\hat{d}_{GPH}$ changes dramatically in the presence of outliers, the estimates from $\hat{d}_{Q_N}$ and $\hat{d}_M$ keep almost unchangeable. As a general conclusion, the empirical result suggests that all the methods can be used to estimate the parameter d when there is not a suspicion of additive or abrupt observation. However, in the existence of a single atypical observation, the methods $\hat{d}_{Q_N}$ and $\hat{d}_M$ should be preferred. Similar conclusions are given in Fajardo et al. (2009) and Reisen et al. (2017) for $\hat{d}_{Q_N}$ and $\hat{d}_M$, respectively.

Table 1. Empirical Mean, Bias and RMSE of $\hat{d}_{GPH}$, $\hat{d}_M$ and $\hat{d}_{Q_N}$ when $\omega = 10$ and $\delta = 0, 0.01, 0.05$.

d	δ	MEAN			BIAS			RMSE		
		$\hat{d}_{GPH}$	$\hat{d}_M$	$\hat{d}_{Q_N}$	$\hat{d}_{GPH}$	$\hat{d}_M$	$\hat{d}_{Q_N}$	$\hat{d}_{GPH}$	$\hat{d}_M$	$\hat{d}_{Q_N}$
	0.0	0.3029	0.2950	0.2933	0.0029	-0.0049	-0.0066	0.0601	0.0596	0.0558
0.3	0.01	0.2226	0.2899	0.3052	-0.0773	-0.0101	0.0052	0.0972	0.0581	0.0584
	0.05	0.1225	0.2681	0.3236	-0.1775	-0.0318	0.0236	0.1873	0.0689	0.0682

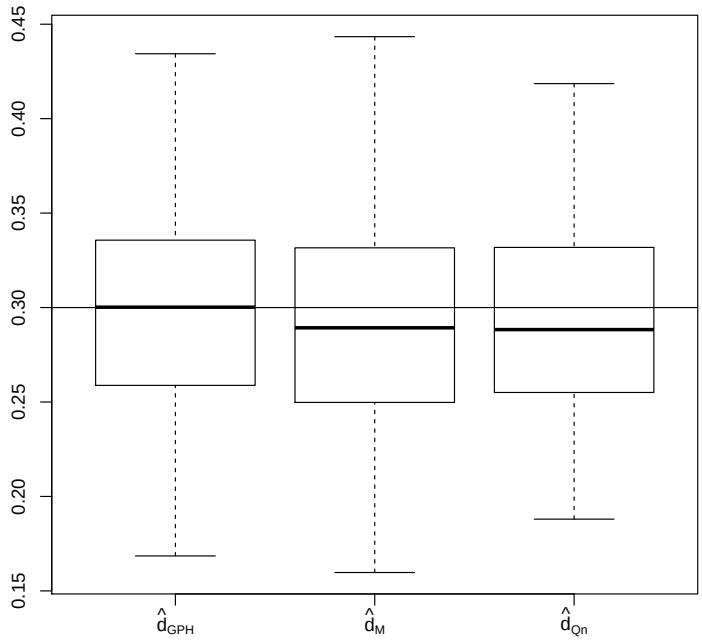


Fig. 5. Boxplots of $\hat{d}_{GPH}$, $\hat{d}_M$ and $\hat{d}_{Q_N}$ when $\delta = 0$.

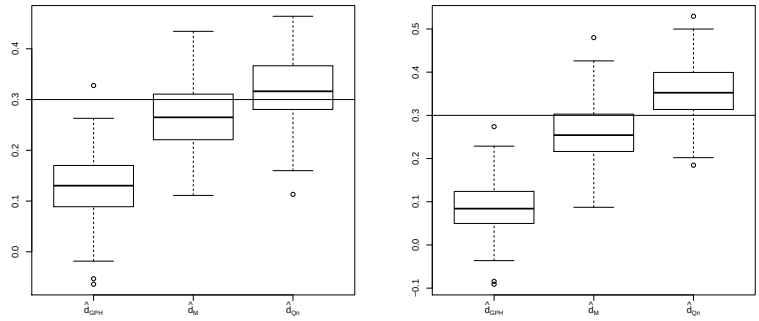


Fig. 6. Boxplots of $\hat{d}_{GPH}$, $\hat{d}_M$ and $\hat{d}_{Q_N}$ when $\delta = 0.05$ and $\delta = 0.1$, respectively.

4.2 Q_n and M-estimators in PARMA models

One of the most popular periodic causal process is the PARMA model which generalizes the ARMA model. $\{Z_t\}_{t\in\mathbb{Z}}$ is said to be a PARMA model if it satisfies the difference equation

$$\sum_{j=0}^{p_\nu} \phi_{\nu,j} Z_{r\mathcal{S}+\nu-j} = \sum_{k=0}^{q_\nu} \theta_{\nu,k} \varepsilon_{r\mathcal{S}+\nu-k}, r \in \mathbb{Z} \tag{47}$$

where for each season ν ($1 \leq \nu \leq \mathcal{S}$) where $\mathcal{S}$ is the period, p_ν and q_ν are the AR and MA orders, respectively, $\phi_{\nu,1}, \ldots, \phi_{\nu,p_\nu}$ and $\theta_{\nu,1}, \ldots, \theta_{\nu,q_\nu}$ are the AR and MA coefficients, respectively, and $\phi_{\nu,0} = \theta_{\nu,0} = 1$. The sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is zero-mean and uncorrelated, and has periodic variances

with period $\mathcal{S}$, i.e. $\mathrm{E}(\varepsilon^2_{r\mathcal{S}+\nu}) = \sigma^2_\nu$ for $\nu = 1, \ldots, \mathcal{S}$. In the following, $p = \max_\nu p_\nu$, $q = \max_\nu q_\nu$, $\phi_{\nu,j} = 0$ for $j > p_\nu$, $\theta_{\nu,k} = 0$ for $k > q_\nu$, and (47) is referred as the PARMA$(p,q)_\mathcal{S}$ model (see, for example, Basawa and Lund (2001) and Sarnaglia et al. (2015)).

To deal with outliers effect in the estimation of PAR model, Sarnaglia et al. (2010) proposed the use of the $Q_N(.)$ function in this model. Following the same lines of the linear time series model described previously, the $Q_N(.)$ function is used to compute an estimator of the periodic autocovariance function $\gamma^{(\nu)}(h)$ at lag h and this sample ACF based on $Q_N(.)$ estimator, denoted here as $\gamma_Q^{(\nu)}(h)$, replaces the classical periodic ACF $\gamma^{(\nu)}(h)$ in the Yule-Walker periodic equations (see, for example, McLeod (1994) and Sarnaglia et al. (2010)) to derive an alternative parameter estimator method for a periodic AR model. The authors derived some asymptotic and empirical properties of the proposed estimator. They showed that the method well accommodate the effect of additive outliers, that is, it presented robustness against these type of observations in the finite sample size series as well as in a real data set.

Let now $Z_1, \ldots Z_N$, where $N = n\mathcal{S}$, be a sample from PAR process which is a particular case of the model definition in (47) with $q_\nu = 0$ and let now $Q_N(.)$ for PAR process be defined as

$$Q_N^{(\nu)}(Z) = Q_N(\{Z_{r\mathcal{S}+\nu}\}_{0\le r\le N}) \,. \tag{48}$$

Based on $Q_N^{(\nu)}(Z)$, the authors derived the sample ACF for periodic stationary processes $\hat{\gamma}_Q^{(\nu)}(h)$. Under some model assumptions, they proved the following main results.

1. For a fixed lag h, $\hat{\gamma}_Q^{(\nu)}(h)$ satisfies the following central limit theorem: As $N \longrightarrow \infty$,

$$\sqrt{N}\left(\hat{\gamma}_Q^{(\nu)}(h) - \gamma^{(\nu)}(h)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \breve{\sigma}_h^2) \,,$$

   where $\gamma^{(\nu)}(h)$ is the periodic ACF function and $\breve{\sigma}_h^2$ is the variance, more details are given in Sarnaglia et al. (2010).
2. The $Q_N^{(\nu)}$ Yule-Walker estimators $(\tilde{\phi}_{\nu,i})_{1\le i\le p_\nu, \nu=1,\ldots,\mathcal{S}}$ satisfy $\tilde{\phi}_{\nu,i} - \phi_{\nu,i} = O_P(N^{-1/2})$ for all $i = 1, \ldots, p_\nu$ and ν in $\{1, \ldots, \mathcal{S}\}$.

Recently, Solci et al. (2018) compared the Yule-Walker estimator (YWE), the robust least squares estimator (Shao (2008)) and the ACF Q_n estimator ($\hat{\gamma}_Q^{(\nu)}(h)$, denoted here RYWE, in the context of estimating the parameters in PAR models with and without outliers. Their main conclusion is similar to the cases discussed previously, that is, for the case of ARFIMA model $\hat{\gamma}_Q^{(\nu)}(h)$ displayed good performance in estimating the parameters in PAR models, periodic samples with and without outliers. As expected, the YWE estimator performed very poorly with the presence of outliers in the data. One of their simulation results is reproduced in the table below (Table 2) in which $n = 100, 400$ (cycles), $\mathcal{S} = 4$, ϵ_t is a Gaussian white noise process and $\delta = 0.01$ (outlier's probability) and magnitude $\omega = 10$. The results correspond to the mean of 5000 replications.

Table 2. Bias and RMSE for Model 1 and outliers with probability $\delta = 0.01$.

				YWE		RYWE	
ω	ϵ_t	n	$\phi_{\nu,1}$	Bias	RMSE	Bias	RMSE
0	$\mathcal{N}(0,1)$	100	0.9	-0.007	0.077	-0.003	0.103
			0.8	-0.002	0.065	0.004	0.084
			0.7	0.000	0.063	-0.001	0.083
			0.6	-0.005	0.066	-0.003	0.083
		400	0.9	-0.001	0.037	-0.001	0.047
			0.8	-0.001	0.031	0.000	0.038
			0.7	-0.001	0.032	0.001	0.038
			0.6	0.000	0.032	0.000	0.039
7	$\mathcal{N}(0,1)$	100	0.9	-0.181	0.247	0.014	0.120
			0.8	-0.118	0.176	0.012	0.096
			0.7	-0.105	0.157	0.015	0.091
			0.6	-0.097	0.151	0.012	0.091
		400	0.9	-0.183	0.203	0.017	0.055
			0.8	-0.129	0.144	0.012	0.046
			0.7	-0.108	0.124	0.013	0.044
			0.6	-0.103	0.119	0.014	0.043

As an alternative estimator of $\tilde{\phi}_{\nu,i}$, Sarnaglia et al. (2016) proposed the use of $M-$periodogram function to obtain estimates of the parameters in PARMA models. The estimator is based on the approximated Whittle function suggested in Sarnaglia et al. (2015). Basically, the Whittle $M-$estimator of PARMA parameters is derived by the ordinary Fourier transform with the non-linear M-regression estimator for periodic processes in the harmonic regression equation that leads to the classical periodogram. The empirical simulation investigation in Sarnaglia et al. (2016) considered the scenarios of periodic time series with presence and absence of additive outliers. Their small sample size investigation leaded to a very promising estimation method under the context of modelling periodic time series with additive outliers and heavy-tailed distributions. The theoretical justification of the proposed estimator is still an open problem and it is now a current research theme of the authors.

Table 3 displays results of a simple simulation example to show the empirical performance of the Whittle $M-$estimator with the Huber function $\psi(x)$ (Huber (1964)) compared to the maximum Gaussian and Whittle likelihood estimators to estimate a PAR(2) model with parameters $\phi_{1,1} = -0.2$, $\phi_{2,1} = -0.5$, $\sigma^2_{1,1} = 1.0$ and $\sigma^2_{2,1} = 1.0$. The sample sizes are $N = n\mathcal{S} = 300$, 800 ($n = 150$, 400, respectively) and the Huber function was used with constant equal to 1.345, which ensure that the M-estimator is 95% as efficient as the least squares estimator for univariate multiple linear models with independent and identically distributed Gaussian white noise. The sample root mean square error (RMSE) was computed over 5000 replications. The PAR(2) model with additive outliers was generated with outlier's probability $\delta = 0.01$ and magnitude $\omega = 10$. The values with "$*$" refer to the RMSE for the contaminated series.

Table 3. Empirical RMSE results for estimating an PAR(2) model.

Method	N	$\phi_{1,1}$	$\sigma^2_{1,1}$	$\phi_{2,1}$	$\sigma^2_{2,1}$
MLE	300	0.067; 0.121*	0.117; 1.366*	0.079; 0.252*	0.111; 1.363*
	800	0.048; 0.101*	0.079; 1.122*	0.046; 0.239*	0.074; 1.253*
WLE	300	0.068; 0.121*	0.117; 1.368*	0.079; 0.252*	0.111; 1.364*
	800	0.048; 0.101*	0.079; 1.122*	0.046; 0.239*	0.074; 1.253*
RWLE	300	0.067; 0.067*	0.147; 0.179*	0.083; 0.089*	0.147; 0.189*
	800	0.051; 0.054*	0.118; 0.149*	0.051; 0.058*	0.108; 0.152*

In the absence of outliers, in general, all estimators present similar behaviour. Relating to the estimation of the variance of the innovations, the MLE and WLE seem to be more precise which is an expected result since the data is Gaussian with zero-mean and these two methods are asymptotically equivalents. The RMSE of the estimators decreases as the sample size increases. When the simulated data has outliers, as an expected result the MLE and WLE estimates are totally corrupted by the atypical observations while the RWLE estimator presents generally accurate estimates. This simple example of simulation leads to the same conclusions of the models discussed previously in which $M-$regression method was also considered.

The methods discussed above give strong motivation to use the methodology in practical situations in which periodically correlated time series contain additive outliers. For example, Sarnaglia et al. (2010) applied the robust ACF estimator $\hat{\gamma}_Q^{(\nu)}(h)$ to fit a model for the quarterly Fraser River data. Sarnaglia et al. (2016) and Solci et al. (2018) analysed air pollution variables using the robust methodologies discussed in these papers. In the first paper, the authors considered the daily average SO_2 concentrations and, in the second one, it was analysed the daily average PM_{10} concentrations. Both data set were collected at Automatic Air Quality Monitoring Network (RAMQAr) in the Great Vitória Region GVR-ES, Brazil, which is composed by nine monitoring stations placed in strategic locations and accounts for the measuring of several atmospheric pollutants and meteorological variables in the area. In general, the models well fitted the series and all these applied examples revealed outliers effects on the estimates.

5 Proof of Theorem 1

By Propositions 1 and 4 and Example 1 of Wu (2007) the assumptions of Theorem 1 of Wu (2007) hold. Thus,

$$\sqrt{\frac{N}{2}}(F(c)-F(-c))\hat{\boldsymbol{\beta}}_N^{\mathrm{M}}(\lambda_j) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Delta}^{(j)}\right),\ N\to\infty\,,$$

with

$$\boldsymbol{\Delta}^{(j)} = \sum_{k\in\mathbb{Z}} \mathbb{E}\{\psi(\varepsilon_0)\psi(\varepsilon_k)\}\boldsymbol{\Delta}_k^{(j)},$$

where

$$\boldsymbol{\Delta}_k^{(j)} = \lim_{N\to\infty}\frac{2}{N}\sum_{\ell=1}^{N-|k|}\begin{pmatrix}\cos(\ell\lambda_j)\\ \sin(\ell\lambda_j)\end{pmatrix}\left(\cos((\ell+k)\lambda_j)\ \sin((\ell+k)\lambda_j)\right).$$

Observe that

$$\begin{aligned}\boldsymbol{\Delta}_k^{(j)} &= \lim_{N\to\infty}\frac{2}{N}\sum_{\ell=1}^{N-|k|}\begin{pmatrix}\frac{\cos(k\lambda_j)+\cos((2\ell+k)\lambda_j)}{2} & \frac{\sin(k\lambda_j)+\sin((2\ell+k)\lambda_j)}{2}\\ \frac{-\sin(k\lambda_j)+\sin((2\ell+k)\lambda_j)}{2} & \frac{\cos(k\lambda_j)-\cos((2\ell+k)\lambda_j)}{2}\end{pmatrix}\\ &= \begin{pmatrix}\cos(k\lambda_j) & \sin(k\lambda_j)\\ -\sin(k\lambda_j) & \cos(k\lambda_j)\end{pmatrix} + \lim_{N\to\infty}\frac{2}{N}\sum_{\ell=1}^{N-|k|}\begin{pmatrix}\frac{\cos((2\ell+k)\lambda_j)}{2} & \frac{\sin((2\ell+k)\lambda_j)}{2}\\ \frac{\sin((2\ell+k)\lambda_j)}{2} & \frac{-\cos((2\ell+k)\lambda_j)}{2}\end{pmatrix}.\end{aligned}$$

By observing that

$$\begin{aligned}&\frac{1}{N}\sum_{\ell=1}^{N-|k|}\cos((2\ell+k)\lambda_j) = \frac{\cos(k\lambda_j)}{N}\sum_{\ell=1}^{N-|k|}\cos(2\ell\lambda_j) + \frac{\sin(k\lambda_j)}{N}\sum_{\ell=1}^{N-|k|}\sin(2\ell\lambda_j)\\ &= \frac{\cos(k\lambda_j)}{N}\cos(\lambda_j(N-|k|-1))\frac{\sin(\lambda_j(N-|k|))}{\sin(\lambda_j)} + \frac{\sin(k\lambda_j)}{N}\sin(\lambda_j(N-|k|-1))\frac{\sin(\lambda_j(N-|k|))}{\sin(\lambda_j)}\end{aligned}$$

tends to zero as N tends to infinity and that the same holds for $N^{-1}\sum_{\ell=1}^{N-|k|}\sin(2\ell+k)$, this concludes the proof.

Acknowledgements

V. A. Reisen gratefully acknowledges partial financial support from FAPES/ES, CAPES/Brazil and CNPq/Brazil and CentraleSupélec. Màrton Ispàny was supported by the EFOP-3.6.1-16-2016-00022 project. The project is cofinanced by the European Union and the European Social Fund. Paulo Roberto Prezotti Filho and Higor Cotta are Ph.D students under supervision of V. A. Reisen and P. Bondon. The authors would like to thank the referee for the valuable suggestions.

Bibliography

Basawa, I., Lund, R., 2001. Large sample properties of parameter estimates for periodic ARMA models. Journal of Time Series Analysis 22 (6), 651–663.

Croux, C., Rousseeuw, P. J., 1992. Time-efficient algorithms for two highly robust estimators of scale. Computational Statistics 1, 1–18.

Fajardo, F., Reisen, V. A., Cribari-Neto, F., 2009. Robust estimation in long-memory processes under additive outliers. Journal of Statistical Planning and Inference 139 (8), 2511–2525.

Fajardo, F. A., Reisen, V. A., Lévy-Leduc, C., Taqqu, M., 2018. M-periodogram for the analysis of long-range-dependent time series. Statistics, 665–683.

Fox, A. J., 1972. Outliers in time series. Journal of the Royal Statistical Society 34 (B), 350–363.

Geweke, J., Porter-Hudak, S., 1983. The estimation and application of long memory time series model. Journal of Time Series Analysis 4 (4), 221–238.

Huber, P., 1964. Robust estimation of a location parameter. The Annals of Mathematical Statistics 35, 73–101.

Koul, H. L., Surgailis, D., 2000. Second order behavior of M-estimators in linear regression with long-memory errors. Journal of Statistical Planning and Inference 91 (2), 399–412.

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M., Reisen, V. A., 2011. Robust estimation of the scale and the autocovariance function of Gaussian short and long-range dependent processes. Journal of Time Series Analysis 32 (2), 135–156.

Ma, Y., Genton, M., 2000. Highly robust estimation of the autocovariance function. Journal of Time Series Analysis 21 (6), 663–684.

McLeod, A. I., 1994. Diagnostic checking periodic autoregression models with application. Journal of Time Series Analysis 15 (2), 221–33.

Reisen, V. A., Lévy-Leduc, C., Taqqu, M. S., 2017. An M-estimator for the long-memory parameter. Journal of Statistical Planning and Inference 187, 44 – 55.

Rousseeuw, P. J., Croux, C., 1993. Alternatives to the median absolute deviation. Journal of the American Statistical Association 88 (424), 1273–1283.

Sarnaglia, A. J. Q., Reisen, V. A., Bondon, P., 2015. Periodic ARMA models: Application to particulate matter concentrations. In: 23rd European Signal Processing Conference. pp. 2181–2185.

Sarnaglia, A. J. Q., Reisen, V. A., Bondon, P., Lévy-Leduc, C., 2016. A robust estimation approach for fitting a PARMA model to real data. In: IEEE Statistical Signal Processing Workshop. pp. 1–5.

Sarnaglia, A. J. Q., Reisen, V. A., Lévy-Leduc, C., 2010. Robust estimation of periodic autoregressive processes in the presence of additive outliers. Journal of Multivariate Analysis 101 (9), 2168–2183.

Shao, Q., 2008. Robust estimation for periodic autoregressive time series. Journal of Time Series Analysis 29 (2), 251–263.

Solci, C. C., Reisen, V. A., Sarnaglia, A. J. Q., Pascal, B., 2018. Empirical study of robust estimation methods for PAR models with application to PM_{10} data. in press, Communication in Statistics 15.

Wu, W. B., 2007. M-estimation of linear models with dependent errors. The Annals of Statistics 35 (2), 495–521.