Spanish Journal of Statistics

VOLUME 3, NUMBER 1, 2021

INE

NIPO: 096-22-006-0
ISSN: 2695 - 9070
DOI: https://doi.org//10.37830/SJS.2021.1

Spanish Journal of Statistics

Published by the Spanish National Statistical Institute

SPANISH JOURNAL OF STATISTICS

VOLUME 3, NUMBER 1, 2021

Contents

Editorials **2**

Presentation of Volume 3, 1, 2021
J.M. Sarabia **5**

Research papers **6**

Some recent methods for analyzing high dimensional time series **7**
D. Peña

A note on explicit expressions for moments of order statistics **37**
D. Moriña, A. Fernández-Fontelo, A. Cabaña, A. Arratia, P. Puig

A note on explicit expressions for moments of order statistics **45**
F. Castellares, A.J. Lemonte and M.A.C. Santos

Acknowledgement to Reviewers **54**

SPANISH JOURNAL OF STATISTICS
Vol. **3** No. 1, 2021, Pages 5 – 6
doi: https://doi.org/10.37830/SJS.2021.1.01

EDITORIAL

Presentation of Volume 3, 1, 2021

José María Sarabia

Editor-in-Chief Spanish Journal of Statistics

Dear readers and dear members of the statistical community:

It is a great pleasure for me to present Volume 3, 1, corresponding to the year 2021. This volume is made up of three articles: one invited article, and two articles within the general statistics section.

The invited article is entitled "Some recent methods for analyzing high dimensional time series", whose author is Daniel Peña, winner of the first National Statistics Prize. The National Statistics Institute of Spain (INE) has awarded the National Statistics Prize to Daniel Peña Sánchez de Rivera, Emeritus Professor of Statistics and former Rector of the Carlos III University of Madrid, in recognition of his outstanding contributions, scientific work and eminent contribution to the progress of Statistics.

The article presents six recent advances in the analysis of high-dimensional time series. The first two procedures have the aim of understanding the structure of the set of series: dynamic quantiles for data visualization and clustering by dependency to split the series into homogeneous groups. The other four methods are oriented to modeling and forecasting large sets of time series by dynamic factor models (DFM). This work introduces procedures for determining the number of factors, for estimating DFM with cluster structure, for forecasting generalized dynamic factor models and for modeling matrices of time series. Some additional comments about the future evolution of the field of dependent high-dimensional data are included. It is a pleasure to have this guest article by Prof. Peña.

The next two papers are presented in the general section. The second article has the title "Misreported longitudinal data in epidemiology: Review of mixture-based advances and current challenges", whose authors are David Moriña, Amanda Fernández-Fontelo, Alejandra Cabaña, Argimiro Arratia and Pedro Puig. The problem of dealing with misinformed data is very common in a wide range of contexts. In this article, the authors consider a comprehensive review of recently proposed methods based on mixed models for longitudinal data, both correlated and uncorrelated. Several applied examples are discussed, including approximations to the burden of cases of Covid-19 infection in Spain. In the same vein, different approaches are studied to deal with unreported records of human papillomavirus infections and genital warts in Catalonia.

The third paper is titled: "A note on explicit expressions for moments of order statistics", by Fredy Castellares, Artur J. Lemonte and Marcos A.C. Santos. In this article, the authors provide an alternative closed-form expression for the moments of order statistics. These formulas are applied

 Published by the Spanish National Statistical Institute

to important probability distributions. The authors also consider numerical studies to demonstrate that these formulas lead to satisfactory results.

Finally, I would like to thank all the authors of this volume for choosing our journal as a means of disseminating their research. I appreciate the work of the editors and reviewers of the papers, who contribute to maintaining a high standard of scientific quality.

SPANISH JOURNAL OF STATISTICS
Vol. **3** No. 1 2021, Pages 7–35
doi:https://doi.org/10.37830/SJS.2021.1.02

INVITED ARTICLE

Some recent methods for analyzing high dimensional time series

Daniel Peña

Universidad Carlos III de Madrid, e-mail:daniel.pena@uc3m.es

Received: May 15, 2021. Accepted: June 28, 2021.

Abstract: This article analyzes six recent advances in the analyses of high dimensional time series. The first two procedures have the objective of understanding the structure of the set of series: dynamic quantiles for data visualization and clustering by dependency to split the series into homogeneous groups. The other four methods are oriented to modeling and forecasting large sets of time series by dynamic factor models (DFM): procedures for determining the number of factors, for estimating DFM with cluster structure, for forecasting generalized dynamic factor models and for modeling matrices of time series are described. Some comments about the future evolution of the field of dependent high dimensional data are included in the conclusions.

Keywords: Correlation Matrices, clustering, dynamic factor models, forecasting; tensor data, graphical methods

MSC: 62A01, 62H25, 62M10, 62M20

1 Introduction

Statistical methods have been changing depending on the available data and the computing possibilities. Nowadays, data is generated continuously when working, exercising or resting, by automatic devices, as sensors in mobile phones, computers surfing Internet or social networks web pages. As a result of all these activities, large sets of data are stored, including location and time, forming huge spatio-temporal data bases.

Almost a century ago, Fisher (1925) proposed the first general statistical approach to obtain information from the data. At that time, data were a very scarce resource and a crucial problem was to draw all the information from small random samples. Now, we are in a different environment and new methods are required to deal with our increasing data availability. These new approaches are not only developed in statistics, but, also, in computer science, machine learning, artificial intelligence, operations research and applied mathematics. A new field of Data Science is emerging that will integrate the different methodologies for data analysis. Several works have analyzed the changes in Statistics due to the big data revolution: see, for instance, Breiman (2001), Bühlmann and

Van De Geer (2011), Fan et al. (2014), Efron and Hastie (2016), Donoho (2017), Cao (2017), Bühlmann and van de Geer (2018), Torrecilla and Romo (2018), Galeano and Peña (2019), Peña and Tsay (2021) and Peña et al. (2021), among others.

In this work we present some advances made in the last few years in the analysis of large sets of time series in which the author of this article and his coauthors have been involved. Thus, this article does not pretend to be a comprehensive survey on recent advances in the analysis of high dimensional time series that can be found in the books by Tsay (2014) and Peña and Tsay (2021). His objective is to summarize a few practical procedures that, according to the author's experience, have proved to be useful for generating knowledge with this type of data.

The article is organized as follows. Section 2 introduces plots to reveal the dynamic evolution of the set of time series, including the empirical dynamic quantiles (Peña et al., 2019) and other plots recommended in Peña and Tsay (2021) for the visualization of large sets of time series. Section 3 explains a procedure for clustering time series by their linear dependency, developed by Alonso and Peña (2019) . Section 4 introduces Dynamic Factor models and a new proposal for finding the number of factors due to Caro and Peña (2021) . Section 5 describes a procedure for building Dynamic Factor models when the series have cluster structure, proposed by Alonso et al. (2020) . Section 6 introduces Generalized Dynamic Factor models and the estimation method developed by Peña et al. (2019) by using Dynamic principal components. Section 7 presents the generalization of DFM for matrix and tensor data and Section 8 concludes with some final remarks on the evolution of the field of high dimensional dependent data. The presentation of the methods is oriented to their practical applications and does not include technical details that can be found in the original references. The plots and the computations presented in this article have been made with the R package SLBDD, available from CRAN, that has been developed as supplementary material for the book by Peña and Tsay (2021).

2 Visualization of many time series: Empirical Dynamics Quantiles and other plots

The visualization of a large set of time series is important for having a first understanding of the data. Useful plots may suggest heterogeneity, as measurement errors at some time points in one or more series, series that seems to be outliers from the others, or cluster structure. Given a set of m time series with length T observations we can represent it as a matrix where each column contains one of the time series and each row the values of all the series at a time point. Peña and Tsay (2021) proposed two types of plots to reveal the structure of the data. The first type is called *dynamic*, because shows the evolution over time of selected summaries of all series. The second type is called *static*, because shows some summaries of the behavior of each series over the period of observation.

Quantiles have shown to be useful to summarize the distribution of a set of independent data. Therefore, we may try to reveal the dynamics of a set of time series by plotting over time some selected quantiles computed at each time point. Given a random sample $\{x_i\}_{i=1}^m$ of a scalar random variable X with empirical cumulative distribution function (CDF) $\widehat{F}_m(x)$ the empirical p-th quantile $q^{*(p)}$ is defined as

$$q^{*(p)} = \inf_{x \in R} \left\{ x | \widehat{F}_m(x) \geq p \right\}, \tag{1}$$

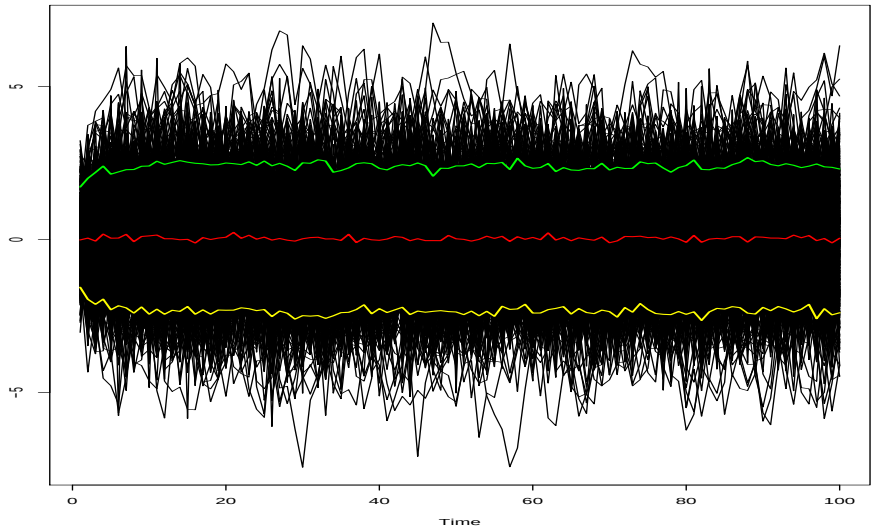

Figure 1: Time-wise quantiles for p =0.95 (green), 0.5 (red) and 0.05 (yellow) of a 1000-dimensional time series generated from a VAR(1) model

and it is well-known, see for instance Ferguson (1967), that this quantile can be computed by

$$q^{*(p)} = \arg\min_{y \in R} \left[p \sum_{x_i \geq y} |x_i - y| + (1-p) \sum_{x_i < y} |x_i - y| \right]. \tag{2}$$

Quantiles for a set of m independent time series, $\{z_{i,t}\}_{t=1}^{T}$ with the same marginal distribution at every time can be defined in a similar way by using the common marginal distribution $F_t(z)$ of z_t. Assuming that $F_t(z)$ is time-varying, the theoretical p-th quantile at time t, $Q_t^{(p)}$, is defined as in (1) by $\inf_{x \in R} \{z | F_t(z) \geq p\}$. Applying this procedure to all time points, $1 \leq t \leq T$, the resulting values at each time t are called empirical time-wise quantiles (ETQ). This definition can be extended to a vector of m dependent time series considering the distribution of the m values at a given time point and obtaining a time series of time-wise quantiles $\{q_t^{*(p)}\}$ over time. However, these quantiles time series only convey information about the marginal behavior of the data and ignore the dynamics of the underlying stochastic process. For instance, if the vector process is strictly stationary the empirical quantiles $q_t^{*(p)}$ for a given p will roughly be a constant straight line. As an illustration, Figure 1 shows the 0.95, 0.5 and 0.05 EDQ for 1000 time series generated by a stationary VAR(1) process. As expected, they do not give much information about the dynamics of the series.

A more informative measure of the dynamics of the series can be obtained in high dimensional data when the quantiles are defined as one of the observed time series. Let $\mathbb{C}_m$ be the set of observed time series. Peña et al. (2019) define the p-th empirical dynamic quantile (EDQ) as the series $\{q_t^{(p)}\}$ in $\mathbb{C}_m$ that satisfies the optimization problem:

$$\{q_t^{(p)}\} = \arg\min_{\{y_t\} \in \mathbb{C}_m} \left[p \sum_{t=1}^{T} \sum_{z_{it} \geq y_t} |z_{it} - y_t| + (1-p) \sum_{t=1}^{T} \sum_{z_{it} \leq y_t} |z_{it} - y_t| \right] \tag{3}$$

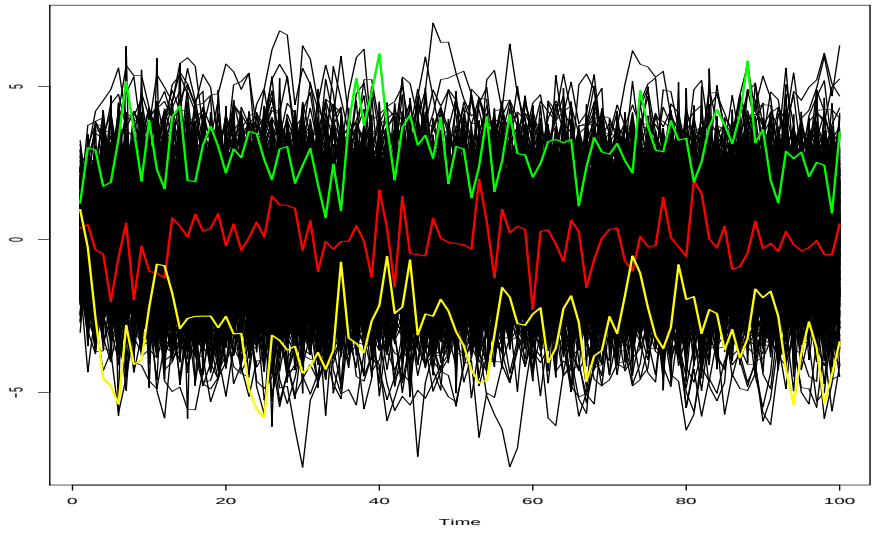

Figure 2: Empirical Dynamic Quantiles for p =0.95 (green), 0.5 (red) and 0.05 (yellow) for 1000 time series generated from a VAR(1) model

Note that the ETQ satisfies equation 3 without the restriction that $\{y_t\} \in \mathbb{C}_m$. Thus, the EDQ are always observed time series and, therefore, they show an observed dynamic evolution of the set of time series.

In order to compute the EDQ we need to try the m time series in equation (3). A fast algorithm to speed up the computations and a proof of the consistency of the proposed method under general assumptions are presented in Peña et al. (2019)

We will compare the EDQ to the ETQ first with simulated data and, in the next subsection, with a real data set. Figure 2 shows the 0.95, 0.5 and 0.05 EDQ for 1000 time series generated by the same stationary VAR(1) process used to compute Figure 1. As expected, the EDQs in Figure 2 exhibit clearly the dynamic dependence of the observed time series, that cannot be seen by examining the ETQ in Figure 1. Also, since the EDQs are always one of the observed time series they can be interpreted: we can obtain the median series and use it as a representative value of the whole set. This average value together with the extreme .01 and .99 EDQ series give some preliminary conclusion about the dynamic variability of the data.

In addition to these dynamic plots, static plots, that summarize the behavior of each series over the whole observation period, are also useful. The first static plot we propose is a boxplot of selected quantiles of each series. That is, calling p_i to the $p-th$ quantile of the values $\{z_{it}\}_{t=1}^{T}$ of the ith series we can compute this quantile for the m series in the set and summarize the distribution of these values by a boxplot. For instance, if we compute the median of each series this boxplot will describe how the median value of each series varies over the set of series. We will call *quantile-boxplot* to a figure that includes five boxplots: those of the maximum and minimum values and the .25, .50 and .75 quantiles of each of the series in the set. When the order of the series is informative, for instance, they are ordered by different latitude locations, the quantiles can be represented with respect to the order of the series. The resulting plot is called a *ordered-quantile plot*.

Two other useful static plots are the scatter plot of the variability/location of the series and the autocorrelation plot of the two first autocorrelations. The variability/location plot shows how a mea-

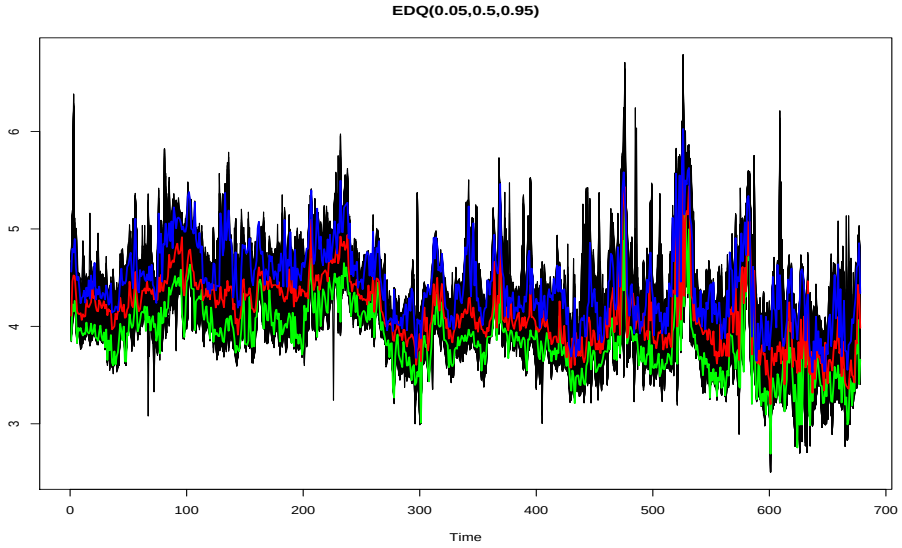

Figure 3: Time plots of three EDQ with probabilities p = 0.05, 0.50, and 0.95 of weekly electricity price in New England.

sure of variability (as the standard deviation or some robust scale, as the median absolute deviation) changes with the location (as the mean, or some robust location measure, as the median) in the series. This figure can identify series that have higher variability, or unusual sample means. The autocorrelation plot shows the scatter plot of lag-1 sample autocorrelation coefficients versus lag-2 sample autocorrelation coefficients for all the series and provides information concerning the dynamic dependence of individual series. If the series share a similar dynamic dependence, then the scatter plot should show a cluster of points along a straight line, whereas with different types of dependency clusters of points will be found. These plots will be illustrated in the next section.

2.1 An Example of Visualizing Electricity Data

We study the hourly day-ahead electricity prices in the New England electric market, also considered by Peña et al. (2019). There are $m = 1344$ ($24 \times 8 \times 7$) time series of hourly price of electricity each day of the week. The series are ordered, so that the first one is the price in the first region at 1 a.m., of January 1st, 2004. The second one is the price, the same day and hour, in the second region and the 9th series the price at 2 a.m. in the same day in the first region, and so on. Thus, the first $8 \times 24 = 196$ series are the prices of the 24 hours of the first day in the sample, that is Thursday January 1st, 2004, in the eight regions of New England (USA). The series are weekly, with an observation every week from January 2004 to December 2016, and a length of $T = 678$ weeks. We analyze the series in logs.

Figure 3 includes all the series in black and three EDQ in color (.05 green, .5 red, and .95 blue) that indicate the dynamic evolution of the set of time series. The variability of the EDQ (.95) is higher than the others and the data shows some large positive outliers. There seems to be two periods in the prices with a drop about $t = 250$ and a decreasing trend in the second period.

In order to compare the properties of each series over the period observed, Figure 4 shows two static plots of this data. The first one, in the left hand side, is a *quantile Boxplot* of the series, with boxplots of the $.25$, $.50$ and $.75$ quantiles as well as the extreme values of the series. It shows large

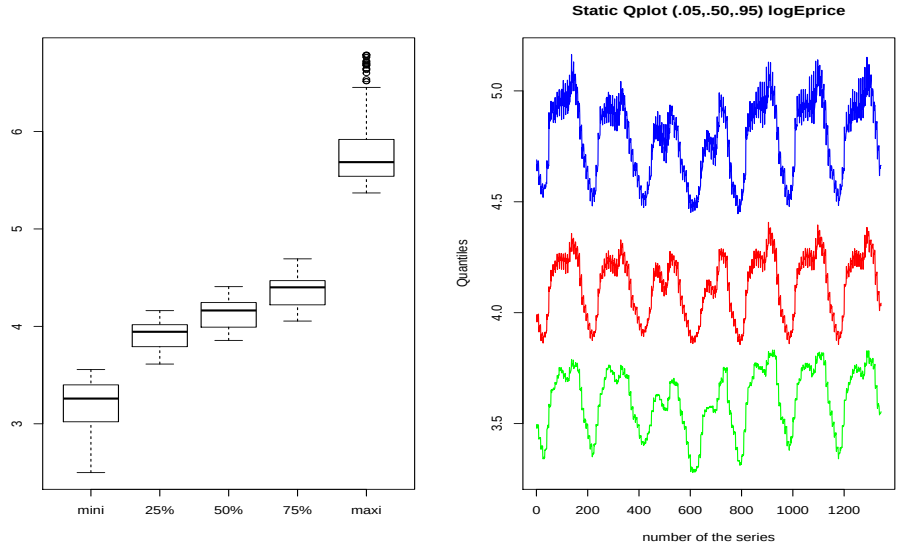

Figure 4: Static plots of quantiles for log electricity price in New England. The left plot shows the distribution of the extremes and the three quantiles (.25,.50,.75) and the right plot the values of the quantiles (.05,.50,.95) with respect to the order of the series.

outliers in the maximum of the series, that were also shown in the dynamic plot of the EDQ in Figure 3. As in the data the order of the series is informative, they are ordered forming blocks of 196 consecutive series corresponding to one day of the week, the three quantiles $p = 0.05, 0.50, 0.95$ have been represented in the right hand side as a function of the order of the series, forming an *ordered-quantile plot*. This plot reveals clearly the pattern of every day, starting on Thursday, that is similar in the eight regions. Inside a given day each sequence of eight observations corresponds to the eight regions. The structure of the seven days of the weeks are clearly shown: larger variability in the four full working days (Thu., Mon., Tue., and Wen.) and smaller in Sat, and Sun., with Fri., as an intermediate day.

To see the evolution inside a given day with more detail Figure 5 shows the same plots but only for the first 196 consecutive series that correspond to Thursday, including the 24 hours and the 8 regions. The pattern of the hours of the day with a maximum price around $19h$ in all the regions is clearly shown.

Finally, Figure 6 gives the scatter plots of the mean and standard deviation of each series and the two first autocorrelation coefficients. The first plot shows large variability with the mean level of the series; the second one a similar relationship between the dynamic indicated by the two first autocorrelation coefficients with no clear signs of heterogeneity in the data.

3 Clustering Time series by Dependency

An important tool in multivariate analysis is clustering the data in homogeneous groups. Clustering time series is usually carried out by choosing a similarity measure between two time series that takes into account their univariate features, as autocorrelation coefficients, periodogram ordinates or coefficients of an autorregresive representation. See Peña and Tsay (2021) for a revision and comparison

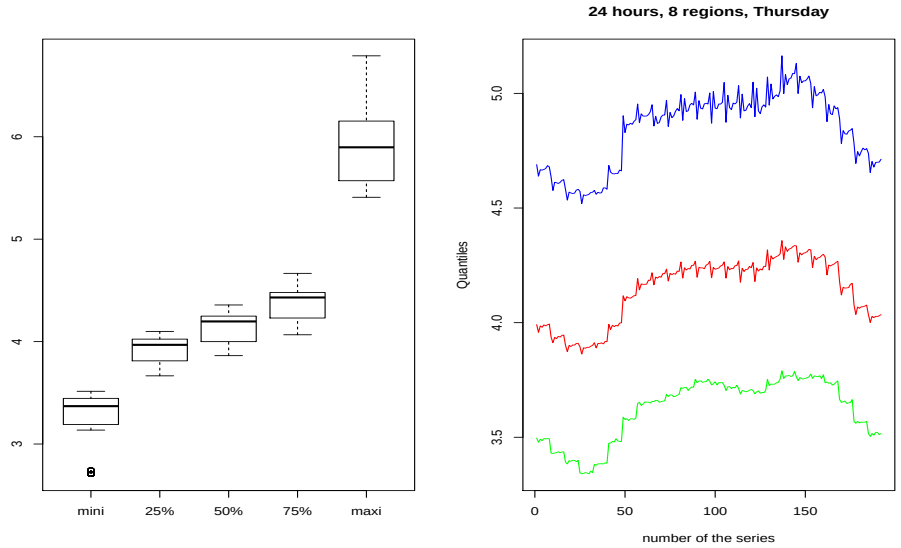

Figure 5: Static plots of quantiles log electricity price in New England for Thursday. The left plot shows the distribution of the extremes and the quantiles .25,.50,.75, and the right plot the values of three quantiles (.05,.50,.95) with respect to the order of the series.

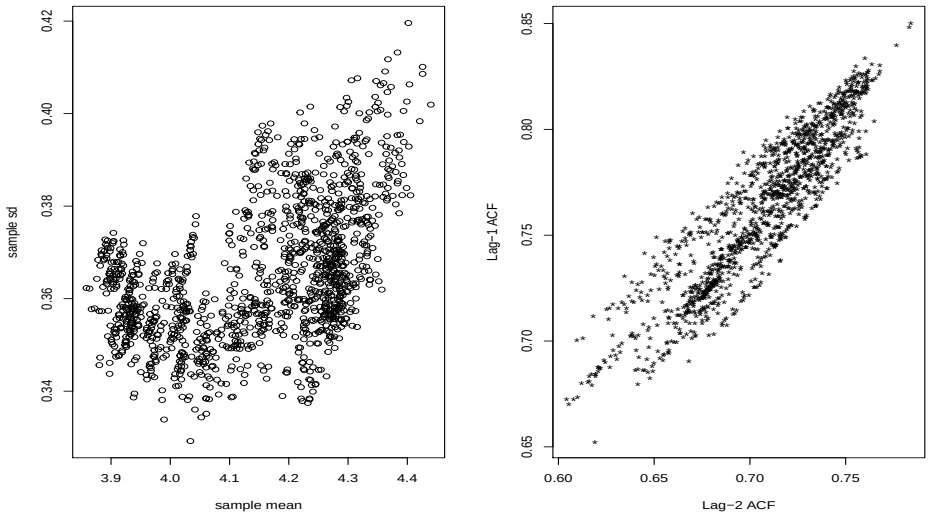

Figure 6: Static plots for the log electricity price in New England. The left plot shows the relationship between mean and standard deviation in each series and the right plot that of the two first autocorrelation coefficients of the series.

of these measures. This approach is useful when we have independent time series and the objective is to cluster them by similarity of their univariate properties. A different approach, proposed by Alonso and Peña (2019), is to cluster the time series by their dependency or association. These authors showed that clustering time series by their univariate properties may classify two series strongly related in different groups, and put together independent series that follow similar models, so that these methods will not be useful if we want to find clusters of related series, as it is usually the objective in many applications.

In order to cluster series for their association a general measure of linear dependency between two stationary time series has to be defined. This can be made as follows. Suppose that, without loss of generality, the series x_t and y_t are standardized to zero mean and unit variance. We can summarize the linear dependency between these series for lags between 0 and k by means of the symmetric non negative definite matrix $\mathbf{R}_{yx,k}$ that corresponds to the covariance matrix of the vector stationary process $(y_t, y_{t-1}, \ldots, y_{t-k}, x_t, x_{t-1}, \ldots, x_{t-k})'$, as

$$\mathbf{R}_{yx,k} = \begin{pmatrix} 1 & \rho_y(1) & \ldots & \rho_y(k) & \rho_{xy}(0) & \rho_{xy}(1) & \ldots & \rho_{xy}(k) \\ \rho_y(1) & 1 & \ldots & \rho_y(k-1) & \rho_{xy}(-1) & \rho_{xy}(0) & \ldots & \rho_{xy}(k-1) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_y(k) & \rho_y(k-1) & \ldots & 1 & \rho_{xy}(-k) & \rho_{xy}(-k+1) & \ldots & \rho_{xy}(0) \\ \rho_{xy}(0) & \rho_{xy}(-1) & \ldots & \rho_{xy}(-k) & 1 & \rho_x(1) & \ldots & \rho_x(k) \\ \rho_{xy}(1) & \rho_{xy}(0) & \ldots & \rho_{xy}(-k+1) & \rho_x(1) & 1 & \ldots & \rho_x(k-1) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho_{xy}(k) & \rho_{xy}(k-1) & \ldots & \rho_{xy}(0) & \rho_x(k) & \rho_x(k-1) & \ldots & 1 \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} \mathbf{R}_{yy,k} & \mathbf{C}_{xy,k}^T \\ \mathbf{C}_{xy,k} & \mathbf{R}_{xx,k} \end{pmatrix}, \tag{5}$$

where $\mathbf{R}_{xx,k}$ is the $(k+1)$ squared and positive definite covariance matrix of the standardized vector of series $X_{t,k} = (x_t, x_{t-1}, \ldots, x_{t-k})'$, $\mathbf{R}_{yy,k}$ corresponds to $Y_{t,k} = (y_t, y_{t-1}, \ldots, y_{t-k})'$ and $\mathbf{C}_{xy,k}$ include the cross correlations between both vectors of series. A global measure of the size of a matrix is its determinant, because, $0 \leq |\mathbf{R}_{yx,k}| \leq 1$, with equality to one holding when $\mathbf{R}_{yx,k}$ is diagonal and the two series are both serially uncorrelated and not linearly related; and $|\mathbf{R}_{yx,k}| = 0$ when there exists a linear combination $a'Z_t = 0$ so that the series are exactly linearly related.

Thus, it may seem that $1 - |\mathbf{R}_{yx,k}|$ could be a global measure of the dependency, with a value of 1 indicating perfect correlation and a value of 0 uncorrelated series. However, it is not so because the determinant depends on both the cross correlations and the autocorrelations of both series. As

$$|\mathbf{R}_{yx,k}| = |\mathbf{R}_{xx,k}| \left| \mathbf{R}_{yy,k} - \mathbf{C}_{xy,k}\mathbf{R}_{xx,k}^{-1}\mathbf{C}_{xy,k}^T \right| \tag{6}$$

if $\mathbf{C}_{xy,k} = 0$, and the series are uncorrelated, $|\mathbf{R}_{yx,k}| = |\mathbf{R}_{xx,k}|\,|\mathbf{R}_{yy,k}|$ and $|\mathbf{R}_{yx,k}|$ can be very small when each of the two determinants, or both, are small because the series have strong autocorrelations and $1 - |\mathbf{R}_{yx,k}|$ will be close to one although the series are uncorrelated. For instance, $|\mathbf{R}_{xx,1}| = 1 - \rho_x^2$ will be very small if the first autocorrelation coefficient is close to one. Note that an exact relationship between the two series implies $|\mathbf{R}_{yx,k}| = 0$, but this determinant can also be very small for uncorrelated series that have strong autocorrelations, by (6).

These properties of $|\mathbf{R}_{yx,k}|$ suggests the following alternative similarity measure proposed by Alonso and Peña (2019)

$$GCC(x_t, y_t) = 1 - \left(\frac{|\mathbf{R}_{yx,k}|}{|\mathbf{R}_{xx,k}|\,|\mathbf{R}_{yy,k}|} \right)^{1/(k+1)} = 1 - \frac{\left| \mathbf{R}_{yy,k} - \mathbf{C}_{xy,k}\mathbf{R}_{xx,k}^{-1}\mathbf{C}_{xy,k}^T \right|^{1/(k+1)}}{|\mathbf{R}_{yy,k}|^{1/(k+1)}}, \tag{7}$$

that was named *generalized cross correlation measure*, $GCC(x_t, y_t)$ between two time series. It can be shown that the GCC verifies: (1) $GCC(x_t, y_t) = GCC(y_t, x_t)$ and the measure is symmetric; (2) $0 \leq GCC(x_t, y_t) \leq 1$; (3) $GCC(x_t, y_t) = 1$ if and only if there is a perfect linear dependency among the series and (4) $GCC(x_t, y_t) = 0$ if and only if all the cross correlation coefficients are zero.

Notice that for $k = 0$ the $GCC(x_t, y_t)$ is just the squared correlation coefficient between the two variables. Also, for any k, when both series are white noise and $\rho_{xy}(h) \neq 0$ for some $h \neq 0$, $k > h$. and $\rho_{xy}(j) = 0$ for all $j \neq h$, then $GCC(x_t, y_t) = \rho^2_{xy}(h)$. In general, for $k > 0$, it can be shown (see Alonso and Peña, 2019) that the $GCC(x_t, y_t)$ represents the increase in accuracy in prediction of the bivariate model with respect to the univariate models and it can be interpreted as an average squared correlation coefficient when we explain the residuals of an autoregressive fitting of one variable by the values of the other.

Alonso and Peña (2019) proposed a procedure to decide about the value of k, the number of lags to be used; estimate the GCC from the data and build a dissimilarity matrix of the series and hierarchical clustering to find the groups.

3.1 Example: Clusters in Electricity Prices

As an example, we consider again the set of time series of hourly day-ahead prices for the New England electric market presented in section 2.1. A hierarchical clustering procedure is applied to the GCC measures for the 1344 regular differentiated series, obtaining the dendrogram presented in Figure 7. This figure shows first seven groups associated by the same weekday. This is to be expected, since the 24 hourly prices of a given day are simultaneously fixed in the daily market, producing a high cross-dependency. The Silhouette statistics provide fourteen groups, that in addition to the weekday take into account two groups of hours in each day: (i) sleeping hours, 01th-06th (or 01th-07th on weekend), and (ii) the awake hours, 07th-24th (08th to 24th on weekend). There is a cluster conformed by a single series (Monday, 10am, Maine) due to the presence of a few very large outliers.

The univariate clustering procedures fail to detect this structure and make two or three groups with no clear interpretation and not related to the weekdays.

4 Dynamic Factor Models

Dynamic Factor Models (DFMs) are a useful approach to model and forecast large sets of big dependent data. Note that the number of parameters in a VARMA model grows with the square of the number of series and, therefore, fitting these models is not feasible when m is large, and even larger than T. In this case, sparse VAR can be fitted by regularization, but DFM provide a good alternative easier to work with. A DFM assumes that the dynamics of the m-dimensional vector of time series variables, $\mathbf{z}_t$, for $t = 1, ..., T$, are explained by the sum of two orthogonal components: the *common* component, that is a low dimensional process with dimension much smaller than m, $r << m$, which is responsible for most of the variability and the *idiosyncratic* component, which includes the weak specific dynamics of each series. A simple representation of these models is:

$$\mathbf{z}_t = \mathbf{P}\mathbf{f}_t + \mathbf{e}_t \tag{8}$$

where, r is the number of latent factors, $\mathbf{P}$ is a $(m \times r)$ matrix of factor loadings, $\mathbf{f}_t$ is a $(r \times 1)$ vector of factors, and $\mathbf{e}_t$ is a $(m \times 1)$ vector of idiosyncratic disturbances or errors. We assume for model identification that $\mathbf{P}'\mathbf{P} = \mathbf{I}$. Also, we consider factors that are not all of them white noise and assume that the lag k covariance matrices of the factors $\mathbf{\Gamma}_f(k) \neq \mathbf{0}$, for some $k > 0$, so that at least one of the

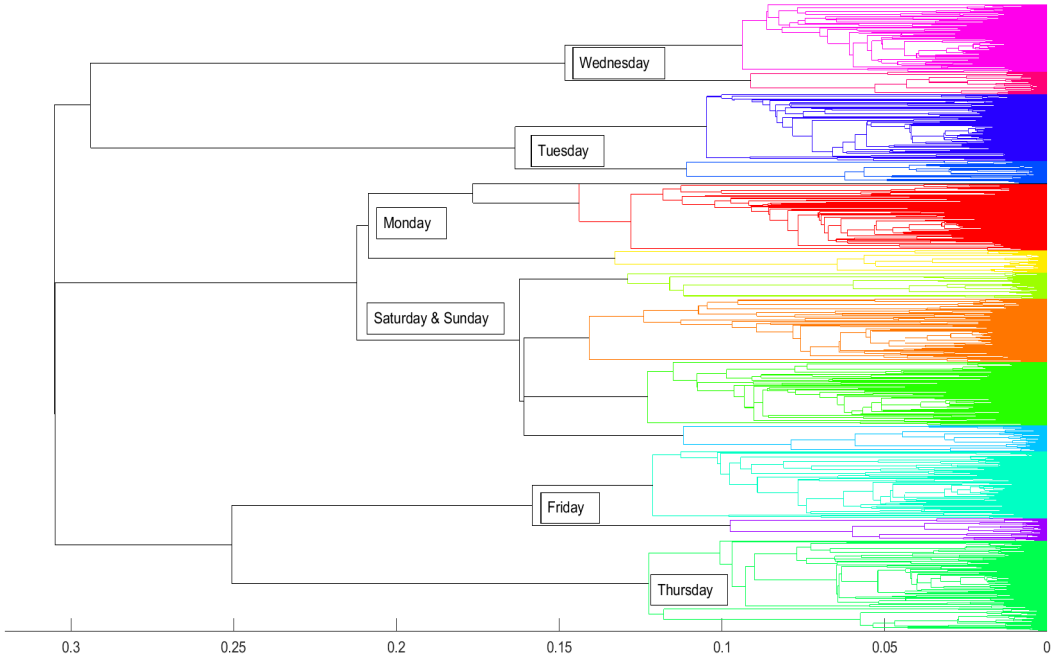

Figure 7: Dendrogram obtained with the regular differentiated series of the eight load zones.

factors presents serial correlation. With the additional assumption that $\mathbf{e}_t$ is a white noise process the model in (8) is identified in finite samples and is called the Exact DFM or EDFM and was studied by Peña and Box (1987). However, the hypothesis of white noise for the idiosyncratic part is often unrealistic in practice. Thus, we can allow weak autocorrelation and cross-section correlation in the idiosyncratic term under assumptions that imply that the noise dynamic vanishes asymptotically, whereas the factors' dynamics remains. See Stock and Watson (2002) and Bai and Ng (2002). The approximate DFM (ADFM) allows the errors, $\mathbf{e}_t$, to be autocorrelated and heteroscedastic, and, also, to present some weak cross-section correlation, but all this dynamic must disappear asymptotically.

The eigenstructure of the covariance matrices can be used to find the number of factors. Note that from (8) and assuming independence between the factors and the errors,

$$\mathbf{\Gamma}_z(k) = \mathbf{P}\mathbf{\Gamma}_f(k)\mathbf{P}' + \mathbf{\Gamma}_e(k), \tag{9}$$

and the properties of the covariance matrices of the data $\mathbf{\Gamma}_z(k)$ depend on the hypothesis made about the covariance matrices of the noise, $\mathbf{\Gamma}_e(k)$. Under the EDFM with white noise, $\mathbf{\Gamma}_e(k) = 0$ for $k \neq 0$, and if $\mathbf{\Gamma}_e(0) = \sigma^2\mathbf{I}$, the homoscedastic case, the matrix $\mathbf{\Gamma}_z(0)$ has r large eigenvalues related to the variance of the factors, and $m - r$ small eigenvalues σ^2. The matrices $\mathbf{\Gamma}_z(k)$, for $k > 0$, will have rank at most r with eigenvalues equal to the covariance of lag k of the factors. If $\mathbf{\Gamma}_e(0) = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix, then the number of large eigenvalues in $\mathbf{\Gamma}_z(0)$ depends on the relative size of the minimum variance of the factors and the maximum variance of the noises. If there is autocorrelation and $\mathbf{\Gamma}_e(k) \neq 0$ for $k \neq 0$, this will not affect the eigenvalues of $\mathbf{\Gamma}_z(0)$ but will affect those of $\mathbf{\Gamma}_z(k)$. Finally, in the more general case, with heteroscedasticity and cross-sectional and serial correlations in the errors, the eigenvalues of all the covariance matrices depend on the assumed structure.

Model (8) can be generalized in several ways. First, for non stationary factors, see Bai and Ng (2004) and Peña and Poncela (2006). Second, for seasonal time series, see Nieto et al. (2016). Third, by allowing lags in the relation between the series and the factors, as proposed by Geweke (1977) and Forni et al. (2000). These last authors proposed the *generalized* dynamic factor model (GDFM), given by

$$\mathbf{z}_t = \sum_{j=0}^{\infty} \mathbf{P}_j \mathbf{f}_{t-j} + \mathbf{e}_t, \tag{10}$$

where, for identification, the factors are assumed to be white noise. Other types of generalizations of the DFM have been proposed. For instance, Correal and Peña (2008) proposed to consider threshold effects, so that the model for the factors changes with a threshold variable and in each threshold regime the series follow a different factor model. These models have been further studied by Liu and Chen (2020) assuming changes in the loading matrices. Some authors have proposed other formulations of the DFM assuming, for instance, that the factors are linear combinations of the data. See Gao and Tsay (2019, 2021a) for a recent proposal in this direction. Finally, the DFM can be applied to the quantiles of the distribution of the data, see Chen et al. (2021). In this article we will concentrate in the basic DFM with stationary time series. The estimation of model (8) will be discussed in the next section, whereas model (10) will be considered in section 6.

4.1 Testing the number of factors in DFM

The estimation of model (8) can be carried out by principal components (PC) or by using the eigenvectors of a combination of lagged covariance matrices (see Peña and Tsay (2021)). An important problem is to determine the number of factors. The three more often used approaches to find them are: (1) Canonical correlation analyses, see Bolívar et al. (2021) for a recent proposal in this direction; (2) Information criteria (see Bai and Ng, 2002); or (3) Eigenvalues, or ratios of consecutive eigenvalues, of an appropriate matrix. For instance, Lam and Yao (2012) proposed to compute the ordered estimated eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_m$ of the pooled covariance matrix

$$\mathbf{M}_{1,k_0} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Gamma}}_z(k) \widehat{\mathbf{\Gamma}}_z(k)', \tag{11}$$

where k_0 is a pre-specified positive integer, and select r as

$$\widehat{r} = \arg \min_{1 \leq i \leq r^*} \frac{\widehat{\lambda}_{i+1}}{\widehat{\lambda}_i},$$

for some $r^* = \alpha m$, where m is the number of series and $0 < \alpha < 1$ (an often used value is $\alpha = 0.2$). Suppose that the first r eigenvalues are large and the remaining eigenvalues are small. Then, the ratios $\lambda_{i+1}/\lambda_i \leq 1$ would have a big decrease for $i = r$.

A similar test has been proposed by Ahn and Horenstein (2013) by using the ordered estimated eigenvalues $\widehat{\nu}_1 \geq \widehat{\nu}_2 \geq ... \geq \widehat{\nu}_m$ of the covariance matrix $\widehat{\mathbf{\Gamma}}_z(0)$. The criterion is

$$\widehat{r} = \arg \max_{1 \leq i \leq r^*} \frac{\widehat{\nu}_i}{\widehat{\nu}_{i+1}}.$$

A problem of these two tests is their lack of robustness to a few atypical series. Suppose that one of the series is affected by some large measurement errors, as often happen in time series automatically

collected by sensor devices. Then: (1) the variance of this atypical series will be much larger than the variance of the other series and, (2) the large outliers due to measurement errors will destroy the cross-section correlation between this atypical series and the others in the set. In the limit, the largest eigenvalue of the covariance matrix of the series will be equal to the variance of the atypical series and the corresponding eigenvector will have small values for the uncontaminated series and a value close to one for this outlying series. This problem will not appear if instead of the autocovariances we work with the autocorrelation matrices. See Fan et al. (2020) for other reasons in this direction.

An additional advantage of using autocorrelations is to avoid the heteroscedasticity of the series. Note that when the idiosyncratic terms are white noise but the series have different variances, the eigenvectors of $\mathbf{\Gamma}_z(k)$ for $k > 1$ are different from those of $\mathbf{\Gamma}_z(0)$. For that reason, Lam and Yao (2012) do not include the covariance matrix in the sum of lagged symmetrized matrices. Caro and Peña (2021) proposed an eigenvalue test based on the weighted combination of the sum of contemporaneous and lagged correlation matrices of the observed data. They define the *combined symmetrized correlation matrix* as

$$\mathbf{R}_{k_0} = \sum_{k=0}^{k_0} w_k \mathbf{R}_z(k)\mathbf{R}_z(k)' \tag{12}$$

where k_0 is a pre-specified positive integer, the coefficients $w_k > 0$ are weights which verify $\sum_{k=0}^{k_0} w_k = 1$, and $\mathbf{R}_z(k)$ is the lag k correlation matrix of the series.

A simple solution to select the weights is to use the asymptotic variance of the autocorrelation and cross correlation coefficients for white noise stationary process, $var(r_{ij}(k)) \approx (T-k)^{-1}$. Then, as in the Box-Ljung portmanteau test of goodness of fit, we can standardize the squared correlations by their variance and define the weights as $(T-k)/c$, where c is chosen so that the weights add up to one by

$$c = \sum_{k=0}^{k_0} (T-k),$$

which implies $c = (k_0+1)(T-k_0/2)$ and $w_k = (T-k)/[(k_0+1)(T-k_0/2)]$. Let $\widehat{\alpha}_1 \geq \widehat{\alpha}_2 \geq ... \geq \widehat{\alpha}_m$ be the ordered estimated eigenvalues of the matrix $\mathbf{R}_{k_0}$. This test selects the number of factors as

$$\widehat{r} = \arg \max_{1 \leq i \leq r^*} \frac{\widehat{\alpha}_i}{\widehat{\alpha}_{i+1}}.$$

Caro and Peña (2021) showed that this test is expected to be more powerful than those based on covarience matrices. The reason is that when one of the series has a variance larger than the others the ratio of consecutive eigenvalues of the correlation matrix is expected to be larger than this ratio in the covariance matrix, and the difference between these two ratios will increase with the heteroscedasticity (the value of σ) and in the larger ratios in the covariance matrix. Thus, standardizing the variables will increase the expected ratio in the correlation matrices when this ratio is already large in the covariance matrices. On the other hand, when these ratios are small in the covariance matrices the expected change will be small in the correlation matrices. This result implies that the standardization of the variables when the series are heteroscedastic is expected to increase the ratio of eigenvalues at the exact number of factors in the correlation matrices with respect to the covariance matrices, increasing, therefore, the power of the ratio of eigenvalues test.

4.2 An Example of Estimating the Business cycle in EMU Countries by DFM

In order to estimate the global business cycle in Europe we analyze the total GDP, the private consumption expenditure (CON) and the gross fixed capital formation (INV) of the 19 countries in the

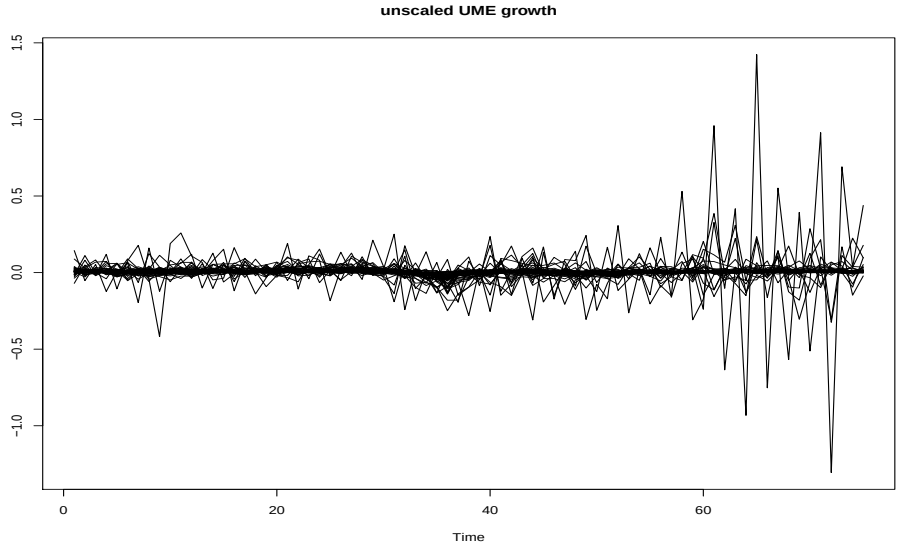

Figure 8: UME growth rate data

euro area, the European Monetary Union (EMU). The data include 57 time series of 76 quarterly observation from 2000 to 2018 and can be downloaded from the SLBDD R package. We analyze here the stationary series, in logs and first difference.

Figure 8 shows a plot of the UME rate of growth data and it is seen that some series have a much large variance of the others. This is corrected in Figure 9, where the series are standardized to zero mean and unit variance. Applying the Caro and Peña (2021) procedure a factor is found and is presented in Figure 10. This factor is able to capture the EMU business cycle taking negative values during the financial crisis in 2008 and during the European Sovereign debt crisis in 2011. It also represents the slow recovery of the EMU economies after 2015. Figure 11 shows that this first factor is an average of all the series with different weights. However, see Caro and Peña (2021), the methods based on ratios of eigenvalues of covariance matrices fail to find a reasonable factor in this case and are dominated by one series of larger variance than the others.

5 Building Factor models with Cluster Structure

Often, there are factors that affect all the series, whereas others are group specific, that is, they affect only to the series included in some clusters but have no effect on the series belonging to other clusters. For instance, the cost of energy in a set of countries may depend on different factors according to groups of countries with similar degree of economic developments, or the evolution of the a pollution indicator measured in different cities may be explained by different factors that depend on the location of the cities. In general, we have a DFM with cluster structure (DFMCS) when the evolution of the time series depends on some general factors, that have influence on all or most of the series, and some specific factors, that are group-dependent. Such models have been studied, among others, by Kose et al. (2003), Wang (2010), Hallin and Liška (2011), Lin and Ng (2012), Bonhomme and Manresa (2015), Ando and Bai (2016, 2017), and Alonso et al. (2020).

We assume that the m-dimensional time series, $\mathbf{z}_t$ can be partitioned as $\mathbf{z}_t = (\mathbf{z}_{1t}', \mathbf{z}_{2t}', \ldots, \mathbf{z}_{ct}')'$, where $\mathbf{z}_{it}$ is a m_i-dimensional time series such that $\sum_{i=1}^{c} m_i = m$. In other

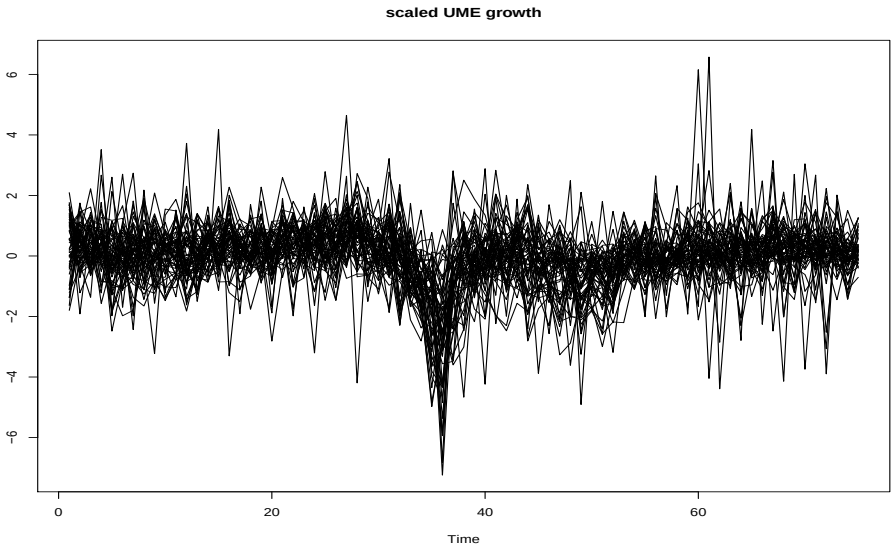

Figure 9: UME data scaled

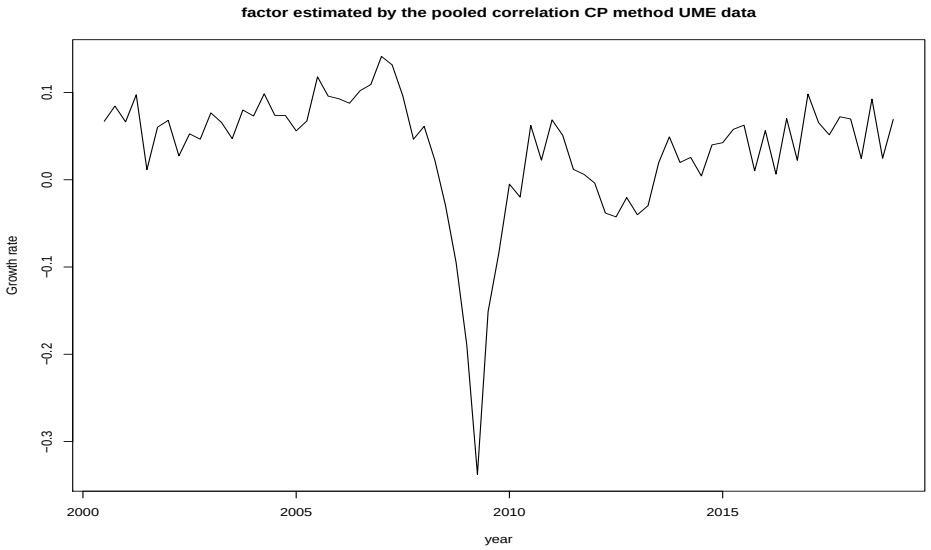

Figure 10: EMU first estimated factor by the CP method

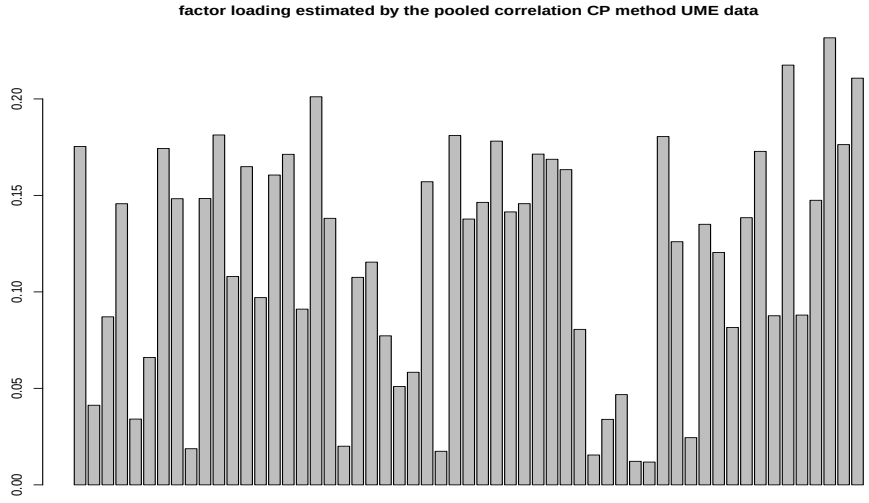

Figure 11: Estimated factor loadings for the first factor for the EMU data by the CP method

words, we assume c clusters of time series where the ith cluster contains m_i time series. The DFMCS can be written as

$$\mathbf{z}_t = \mathbf{P}_0\mathbf{f}_{0t} + \sum_{i=1}^{c} \mathbf{P}_i\mathbf{f}_{it} + \mathbf{n}_t, \tag{13}$$

where $\mathbf{n}_t$ is the vector of noises, $\mathbf{f}_{0t} = (f_{0,1t}, \ldots, f_{0,r_0t})'$ is an r_0-dimensional vector of global factors, $\mathbf{P}_0$ is a $m \times r_0$ global loading matrix, $\mathbf{f}_{it} = (f_{i,1t}, \ldots, f_{i,r_it})'$ is an r_i-dimensional common factor of the ith cluster and $\mathbf{P}_i$ is a $m \times r_i$ loading matrix for the ith cluster such that $\mathbf{P}_i = [\mathbf{0}'_{i-1}, \mathbf{w}'_i, \mathbf{0}'_{c-i}]'$, where $\mathbf{0}_{i-1}$ is a zero matrix of dimension $(\sum_{j=1}^{i-1} m_i) \times r_i$ provided that $i > 1$, $\mathbf{w}_i$ is a $m_i \times r_i$ loading matrix for the ith cluster, and $\mathbf{0}_{c-i}$ is another zero matrix of dimension $(\sum_{j=i+1}^{c} m_i) \times r_i$ provided that $i < c$. The first and the last matrices, $\mathbf{P}_1$ and $\mathbf{P}_c$, have the non zero loadings in the extremes. In this way, $\mathbf{P}_i\mathbf{f}_{it}$ only affects time series in the ith cluster. The total number of factors is $r = r_0 + r_1 + \cdots + r_c$.

The identification conditions of the DFMCS in Equation (13) have been studied by Wang (2010) and are as follows : (1) $\mathbf{P}'_0\mathbf{P}_0 = \mathbf{I}_{r_0}$, where $\mathbf{I}_{r_0}$ is the identity matrix of order r_0; (2) $\mathbf{P}'_i\mathbf{P}_i = \mathbf{w}'_i\mathbf{w}_i = \mathbf{I}_{r_i}$ for $i = 1, \ldots, c$; (3) $\mathbf{P}'_0\mathbf{P}_i = \mathbf{0}_{r_0 \times r_i}$ for $i = 1, ..., c$, and (4) the covariance matrix of the $r = \sum_{j=0}^{c} r_j$ factors is diagonal. Note that, also, by definition, $\mathbf{P}'_i\mathbf{P}_j = \mathbf{0}_{r_i \times r_j}$, for $i \neq j$. We can write this model as a standard factor model. Letting $\mathbf{f}_t = (\mathbf{f}'_{0t}, \mathbf{f}'_{1t}, \ldots, \mathbf{f}'_{ct})'$, and $\mathbf{P} = [\mathbf{P}_0|\mathbf{P}_1|\cdots|\mathbf{P}_c]$, we obtain model (8).

The idiosyncratic term, or noise, $\mathbf{n}_t = (n_{1t}, \ldots, n_{mt})'$, is a general sequence of stationary time series with mean $\mathbf{0}_m$ and weak serial dependency as in the ADFM. The global and specific factors are orthogonal to each other and follow a diagonal vector autoregressive moving average, VARMA. Additionally, we assume that both innovation processes appearing in the factor model are uncorrelated for all lags. However, the number of clusters and the allocation of the series to the clusters are unknown.

The estimation of a DFMCS requires obtaining: (1) The number of global factors, r_0, groups, c, and specific factors in each of the c groups, $r_1, \ldots, r_c$; (2) The label variable $g_i \in \{1, \ldots, c\}$, indicating to which group the series belongs, and we call $\mathbf{G}$ the $m \times 1$ vector with components g_i, for $i = 1, \ldots, m$; (3) The loading matrices of the global and specific factors, $\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_c$ and the time series of these

factors, $\mathbf{f}_{0t}, \mathbf{f}_{1t}, \ldots, \mathbf{f}_{ct}$. Given the estimated factors and noises, where $\widehat{\mathbf{n}}_t = \mathbf{z}_t - \widehat{\mathbf{P}}\widehat{\mathbf{f}}_t$, estimators of the parameters of the scalar ARMA models for the factors and noises can be obtained.

Alonso and Peña (2019) proposed a robust procedure for the estimation of this model that can be summarized as follows: (1) Using all the series compute an initial set of factors and their loadings, build the common component of each time series and cluster these common components by using their dependency with the GCC measures; (2) In each of the groups found compute a new set of factors and their loadings, and classify all the factors found in steps (1) and (2) as global or specific using the empirical canonical correlation between them ; (3) Remove from each time series the estimated common component and, in the group-specific residuals, re-estimate a new set of specific factors; (4) Check whether each group has at least one specific factor for possible group recombinations. This four steps are described now.

In the first step the initial estimation of the factors and loadings can be made with PCA applied to the sample autocorrelations matrices of the time series and the number of factors, r_*, can be determined by using the test proposed by Caro and Peña (2021) based on the ratios of consecutive eigenvalues of the matrix $\mathbf{R}_{k_0}$ in (12). Note that in this step we expect to find all the global factors and some (or all) of the group-specific factors, so that the number of factors in this matrix, r_*, is in general larger than the true number of global factors, r_0. Then, the common component is estimated by $\mathbf{c}_t = \widehat{\mathbf{P}}\widehat{\mathbf{f}}_t = \widehat{\mathbf{P}}\widehat{\mathbf{P}}'\mathbf{z}_t$. The groups are now built applying a hierarchical clustering algorithm with single linkage to the dissimilarity matrix of the $\mathbf{c}_t$ series using the GCC measure.

In the second step the series in the groups are used to estimate new sets of factors and their loadings. Let $r_1^s, \ldots, r_c^s$ be the number of factors found in each group. The specific loading matrices $\widehat{\mathbf{P}}_i$ of dimension $m \times r_i^s$ and columns $\widehat{\mathbf{p}}_{i1}, \ldots, \widehat{\mathbf{p}}_{ir_i^s}$ are built by adding to the eigenvectors corresponding to the largest r_i^s eigenvalues in the ith group, a set of zero values for the observations not included in the group. The factors in each group are estimated by $\widehat{\mathbf{f}}_{ij,t}^s = \widehat{\mathbf{p}}_{ij}'\mathbf{z}_t$, with $j = 1, \ldots, r_i^s$. These group factors are expected to include all the specific factors and some (or all) of the global factors.

Next, in order to decide whether a factor is global or specific, we compare the set of r_* factors found in step-1 and the set of $\sum_{i=1}^{c} r_i^s$ factors found in this second step. Note that the factors contained in the first set may be a rotation of the factors contained in the second set and, therefore, it is not evident which ones should be classified as global and which one as specific. Consequently, we first decide if each factor f_t in the first set of r_* factors is global or specific by applying the following three simple rules: (1) If f_t does not belong to any of the second set of factors then it is a global factor; (2) If f_t belongs to only one of the sets of the second set of factors then it is a specific factor in this group; (3) If f_t belongs to more than one of the second set of factors then it is a global factor.

We decide if a factor, f_t, belongs to a set of specific factors by computing the empirical canonical correlation between the factor, f_t, and the ones in the set, $\widehat{\mathbf{f}}_{i1,t}^s, \ldots, \widehat{\mathbf{f}}_{ir_i^s,t}^s$, with $i = 1, \ldots, c$. When the empirical canonical correlation of factor f_t with elements of the set is higher than some threshold value, ρ_0, we say that f_t belongs to this set. The threshold value of $\rho_0 = 0.9$ seems to work well in our Monte Carlo exercise. Afterwards, we check if any of the groups with $r_1^s, \ldots, r_c^s$ factors include any factor that does not belong to the set of factors found in step-2. If this is the case, the factor is classified as specific factor in the corresponding group.

In the third step the residuals $\mathbf{v}_t = \mathbf{z}_t - \widehat{\mathbf{P}}_0\widehat{\mathbf{f}}_{0t}$ are computed, where $\widehat{\mathbf{f}}_{0t}$ is the vector of estimated global factors obtained in step-3 and $\widehat{\mathbf{P}}_0$ is the corresponding loading matrix to these factors, and the specific factors are re-estimated using the series v_{it} corresponding to each group. Then, we verify that the groups obtained are due to different specific factors and not due to differences between factor loadings in a global factor. This is made by checking whether all the groups have at least one specific factor. We may face the following cases: (1) All the c groups found include at least one specific factor,

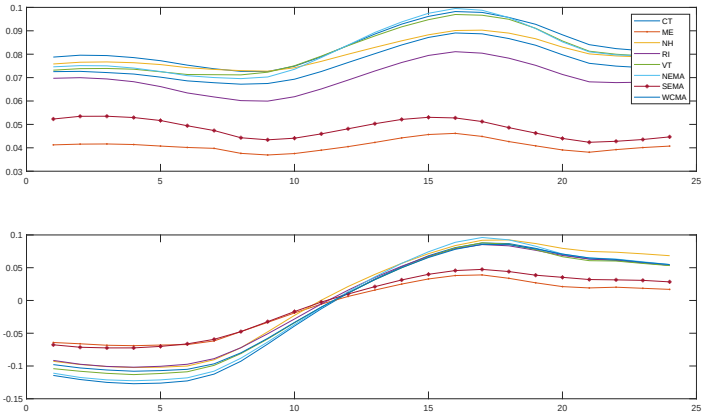

Figure 12: Estimated loadings for the two initial factors for the outlier corrected series. The loadings are shown as functions of the 24 hours of the day and each of the eight curves represents one of the zones.

and we conclude that we have a DFMCS with c groups; (2) c_1 groups, $(1 \leq c_1 < c)$ contain specific factors, and $c_2 = c - c_1$ groups only contain global factors, then we have a DFMCS with $c_1 + 1$ groups; and (3) All the groups only contain global factors, then we have the standard DFM.

Finally, given the estimated factors, loadings, and groups, $\left(\widehat{\mathbf{P}}_0, \widehat{\mathbf{f}}_{0t}, \widehat{\mathbf{P}}_1, \widehat{\mathbf{f}}_{1t}, \ldots, \widehat{\mathbf{P}}_c, \widehat{\mathbf{f}}_{ct}\right)$, estimate AR models for the factors and compute the residuals or idiosyncratic component, $\widehat{\mathbf{n}}_t = \mathbf{z}_t - \widehat{\mathbf{P}}_0\widehat{\mathbf{f}}_{0t} - \sum_{i=1}^{c} \widehat{\mathbf{P}}_i\widehat{\mathbf{f}}_{it}$, and fit AR($p$) models to the idiosyncratic time series.

5.1 An Example of Cluster Specific Factors in Electricity prices

We analyze again the data set of hourly day-ahead demand for the ISO New England electricity market studied in sections 2.1 and 3.1. Here a cleaning procedure for outlier detection is applied to the data as described in Alonso et al. (2020). A proportion of 2.38% of the total number of data points are identified as outliers and they are interpolated to obtain a corrected set of series. With these series two initial factors are found that explain 77.1% and 8.8% of the total variability, respectively. The loadings of these two factors are shown as curves in Figure 12. The first factor is essentially a weighted average of all the series with similar weights (in the range from 0.037 to 0.100) across the 24 hours. Therefore, it reproduces the average global dynamic of the differentiated series. Regarding the effect of each hour, the first estimated factor gives more relative weights to the afternoon (13:00 to 19:00). The second factor gives negative weights to series from 1st to the 11th hours and positive weights to series from the 12nd to 24th hours. Also, it differentiates between the night (1:00 to 7:00) and the rest of the hours, with a peak in 17:00 to 19:00. Note that these factors are not the same across regions, because the loadings for the second (ME) and seventh (SEMA) zones are different from the others. Thus, if we do not consider the presence of clusters in the data, we may conclude that a DFM with two factors seems to be appropriate for the data.

We search for clusters using the GCC measure of the series and two groups are found: the first one broadly includes series in daylight hours and the second one in the night hours. Then, seven

factors are obtained in each of the two groups. These seven factors explain approximately 96.8% and 97.6% of the variability of the series at the first and the second cluster, respectively.

As some of these factors may be global and others are specific, we compare the two initial factors found considering all the series and the fourteen factors found in the two clusters. The conclusion is that the two initial factors are classified as global factors. The first one has 0.984 and 0.967 canonical correlations with the factors in the two clusters, respectively. The second one has weaker correlations, 0.673 and 0.799, respectively, but its canonical correlation with the set of all the specific factors is almost one. This implies that its effect is distributed among several factors found in the groups. Now, we obtain the residuals $R_{it} = z_{it}^c - \widehat{\mathbf{P}}_0\widehat{\mathbf{f}}_{0t}$, where $\widehat{\mathbf{P}}_0$ and $\widehat{\mathbf{f}}_{0t}$ are the estimated loadings and factors, respectively, for the two global factors. With these residuals six and five factors for the first and the second cluster, respectively, are found. These factors are clearly specific and orthogonal to the two global factors.

Figures 13 and 14 show the loadings for these specific factors in the two groups. Note that two extreme zones from the geographical point of view, Maine (ME) and Southeastern Massachusetts (SEMA), have the largest effect in almost all the specific factors in both clusters, whereas for the global factors the situation was just the opposite: these zones have the smallest weights in the two global factors in Figure 12. Regarding the effect of each hour, a richer picture appears in the structure of these group factors with respect to the global ones. In group one, the first three factors give more weights to hours from 11:00 to 18:00 than those from 19:00 to 24:00, and factors four and six account for a peak in electricity demand when most people return home, hours 17:00-18:00. In the second cluster the first two factors have opposite peaks of demand around 1:00-2:00 and 7:00-8:00. The other three factors have small variability in the hours but they differentiate strongly among the eight zones.

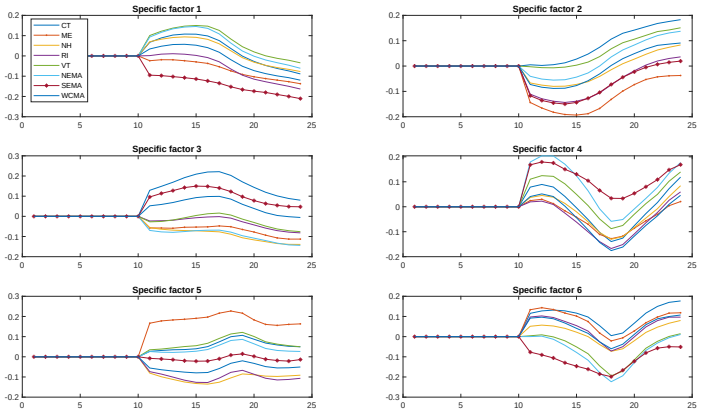

Figure 13: Estimated loadings for the six specific factors in the first group (Hours 11:00 to 24:00) using the outlier corrected series.

6 Generalized Dynamic Factor Models and their Estimation

The DFM assumes a contemporaneous relationship between the series and the common factors. This model can be generalized, see Forni et al. (2000), by assuming that the observed time series are affected by the factors and all their past values. The resulting model is called the Generalized Dynamic Factor Model (GDFM). Assuming a finite number of lags, the representation of the m-dimensional

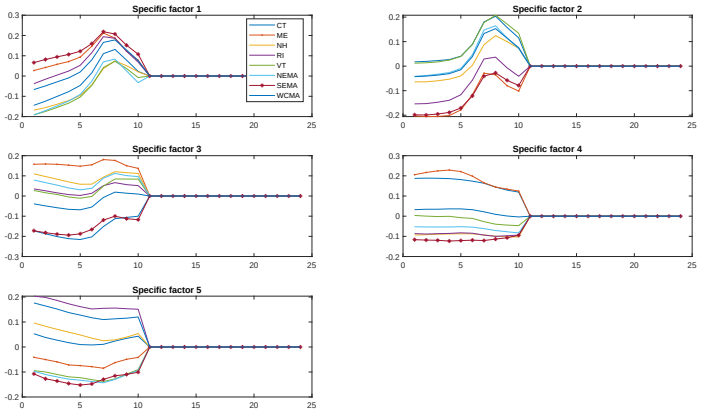

Figure 14: Estimated loadings for the five specific factors of the second group (Hours 1:00 to 10:00) using the outlier corrected series.

vector of time series is:

$$\mathbf{z}_t = \sum_{l=0}^{L-1} \mathbf{P}_l \mathbf{f}_{t-l} + \mathbf{n}_t, \tag{14}$$

where L is a positive integer, $\mathbf{f}_t$ is a q-dimensional stationary vector of uncorrelated common factors, the L matrices $\mathbf{P}_l$ are $m \times q$ factor loading matrices and $\mathbf{n}_t$ is the noise or idiosyncratic part, that is stationary and follows the conditions of the approximate DFM. We denote the number of factors by q in Equation (14) whenever $L > 1$ and return to the notation r if $L = 1$. The reason for this change will be clear shortly. The common component of the series $\mathbf{z}_t$ is $\mathbf{c}_t = \sum_{l=0}^{L-1} \mathbf{P}_l \mathbf{f}_{t-l}$ and is generated by the $q < m$ unobserved factor series. The conditions for identification of the loading matrices and the factors are similar to the DFM studied in section 4 : $\mathbf{P}_l' \mathbf{P}_l = \mathbf{I}_q$ for $l = 0, 1, \ldots, L-1$, and $E(\mathbf{f}_t \mathbf{f}_t')$ is diagonal.

Model (14) can be written with factors without lags, as in (8), by defining a $m \times r$ loading matrix $\mathbf{B} = [\mathbf{P}_0, ..., \mathbf{P}_{L-1}]$, where now $r = qL$ and a $r \times 1$ vector of series $\mathbf{F}_t^D = [\mathbf{f}_t', ..., \mathbf{f}_{t-L+1}']'$.

The estimation of GDFM was initially carried out by Forni et al. (2000; 2005) with dynamic principal component analysis in the frequency domain. More recently Peña and Yohai (2016) and Peña et al. (2019) proposed a new approach to dynamic principal components in the time domain that can be used in the estimation of these models.

6.1 Dynamic Principal Components

The standard principal components provide an optimal reconstruction of the time series data using only contemporaneous information. Brillinger (1981) addressed the reconstruction problem in a more general form, allowing the use of lagged values, and defined the dynamic principal components (DPC) as linear combinations of the time series that provide an optimal reconstruction using all leads and lags of the data. Formally, consider a zero-mean m-dimensional stationary process $\{\mathbf{z}_t | -\infty < t < \infty\}$. Define the first dynamic principal component as a linear combination of all the values of the series

$$f_t = \sum_{h=-\infty}^{\infty} \mathbf{c}_h' \mathbf{z}_{t-h}, \tag{15}$$

where the $\mathbf{c}_h$ are m-dimensional vectors such that f_t provides an optimal reconstruction of the data using all of its lags and leads, that is, the first DPC minimizes:

$$\min E\left[(\mathbf{z}_t - \sum_{j=-\infty}^{\infty} \boldsymbol{\beta}_j f_{t+j})'(\mathbf{z}_t - \sum_{j=-\infty}^{\infty} \boldsymbol{\beta}_j f_{t+j})\right], \tag{16}$$

using some $m \times 1$ vectors $\boldsymbol{\beta}_j, -\infty < j < \infty$. Brillinger elegantly solved this problem in the frequency domain. When this procedure is adapted to finite samples, the number of lags in Equation (15) and in the reconstruction of the series in Equation (16) are to be defined. These DPC have been used by Forni et al. (2000; 2005) for estimation of GDFMs.

These components have been generalized as follows. First, Peña and Yohai (2016) proposed DPC without the assumption that they are computed as linear functions of the data. This generalized DPC were shown to be useful for the estimation of non stationary dynamic factors. Second, Peña et al. (2019) proposed a way to compute the DPC in the time domain using a one side filter, instead of the Brillinger's filter that uses past, present and future values of the data and it is not appropriate for forecasting. They call these new estimates one-sided dynamic principal component (ODPC) and showed that they are very useful for forecasting.

These authors define the first dynamic principal component as,

$$f_t(\widehat{\mathbf{a}}) = \sum_{h=0}^{c_1} \mathbf{z}_{t-h}' \widehat{\mathbf{a}}_h, \quad t = c_1 + 1, \ldots, T, \tag{17}$$

which is a linear combination of the present and lagged values of the series. The $m \times 1$ vectors $\mathbf{a}_h'$ can be aggregated to form a $m(c_1+1)$-dimensional vector $\mathbf{a} = (\mathbf{a}_0', \ldots, \mathbf{a}_{c_1}')'$. The component has the property that the reconstruction of the original data via

$$\mathbf{z}_t^R(\mathbf{a}, \mathbf{B}) = \sum_{h=0}^{c_2} \mathbf{b}_h f_{t-h}(\mathbf{a}), \quad t = c_1 + c_2 + 1, \ldots, T, \tag{18}$$

where $\mathbf{B}$ is a $(c_2+1) \times m$ matrix of the form $\mathbf{B}' = [\mathbf{b}_0, ..., \mathbf{b}_{c_2}]$, where each $\mathbf{b}_i \in R^m$, minimizes the mean squared error in the reconstruction. That is, the vector $\mathbf{a}$ and the matrix $\mathbf{B}$ are such that

$$(\mathbf{a}, \mathbf{B}) = \arg \min_{\|\mathbf{a}\|=1, \mathbf{B}} \frac{1}{T^{*,1}k} \sum_{t=(c_1+c_2)+1}^{T} \left\|\mathbf{z}_t - \mathbf{z}_t^R(\mathbf{a}, \mathbf{B})\right\|^2, \tag{19}$$

where $T^{*,1} = T - (c_1 + c_2)$ is the effective number of observations we can reconstruct with the ODPC. Note that by (17), we can compute the component for $t = c_1 + 1, \ldots, T$, and by (18), the first value we can reconstruct is $t = c_1 + c_2 + 1$. The constraint $\|\mathbf{a}\| = 1$ is included for identification purpose, because if we multiply $\mathbf{a}_h$ in Equation (17) by any positive number, g, and divide $\mathbf{b}_h$ by g, the solution remains the same.

The second ODPC is then defined as a linear combination of the present and lagged series that can be used to optimally reconstruct the residuals $\mathbf{r}_t = \mathbf{z}_t - \mathbf{z}_t^R(\mathbf{a}, \mathbf{B})$ from the first component, and the next components are defined in the same way.

An alternating estimation algorithm can be done as follows. For the first component:

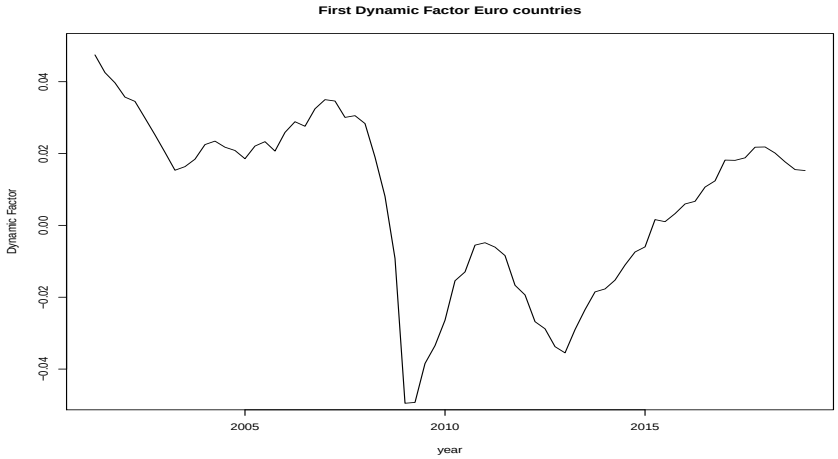

Figure 15: First dynamic factor for the growth rates of GDP series for the 19 countries in the European Monetary Union. The factor was built with 3 lags and entered with 3 lags.

1. Given an estimation of the component $\mathbf{f}^{(i)}$, compute a value for the coefficients of the matrix $\mathbf{B}^{(i)}$ by the multivariate linear regression equation (18).
2. Given the value $\mathbf{B}^{(i)}$ compute $\mathbf{a}^{(i+1)}$ by the equation resulting of taking the derivative with respect to $\mathbf{a}$ in equation (19).

The iteration stops when the relative decrease in the MSE is smaller than δ. Clearly in this algorithm at each step the MSE decreases and, therefore, it converges to a local minimum. To obtain a global minimum, the initial value $\mathbf{f}^{(0)}$ should be sufficiently close to the optimal one. We propose to take $\mathbf{f}^{(0)}$ as the last $T - c_1$ coordinates of the first ordinary principal component of the data.

In practice, the number of components and lags in each component need to be chosen. This is carried out by cross validation and provides a useful way to estimate GDFM, see Peña et al. (2019). With many time series a sparse solution can be computed by regularization of the estimation criterion (Peña et al., 2021).

6.2 Dynamic Factors in EMU countries

Consider, again, the growth rates of standardized GDP series of the 19 Euro countries from 2000.II to 2018.IV. We will illustrate the estimation of one sided dynamic principal components (ODPC) using the R package **odpc** (see Peña et al. (2020)), that is also included in the SLBDD R package. Figure 15 shows the time plot of the first estimated ODPC and Figure 16 the second factor or second ODPC. These results are computed with three lags. Figure 17 plots the coefficients of $\mathbf{a}^1$ of the first dynamic principal component whereas Figure 18 shows the loading in $\mathbf{B}$ also for the 1st dynamic principal component.

7 Tensor Dynamic Factor Models

High dimensional data often appear when instead of measuring the aggregate of a variable we split it into many components. For instance, the total sales in a set of d_1 cities are split into the sales in

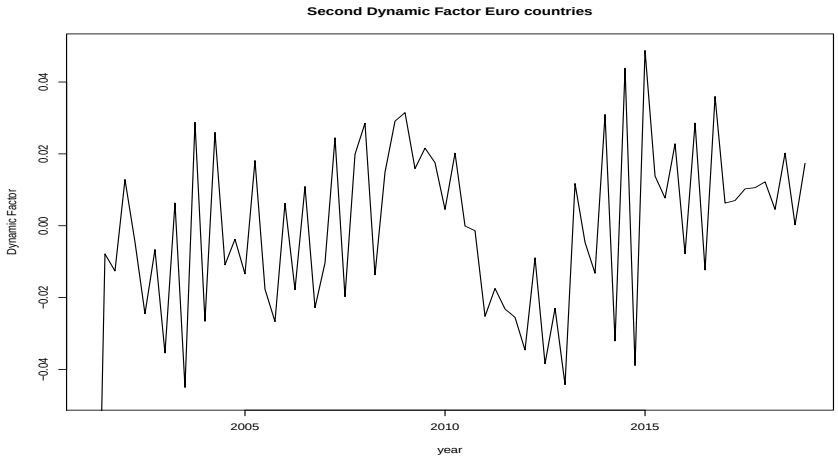

Figure 16: Second dynamic factor for the growth rates of GDP series for the 19 countries in the European Monetary Union. The factor was built with 3 lags and entered with 3 lags.

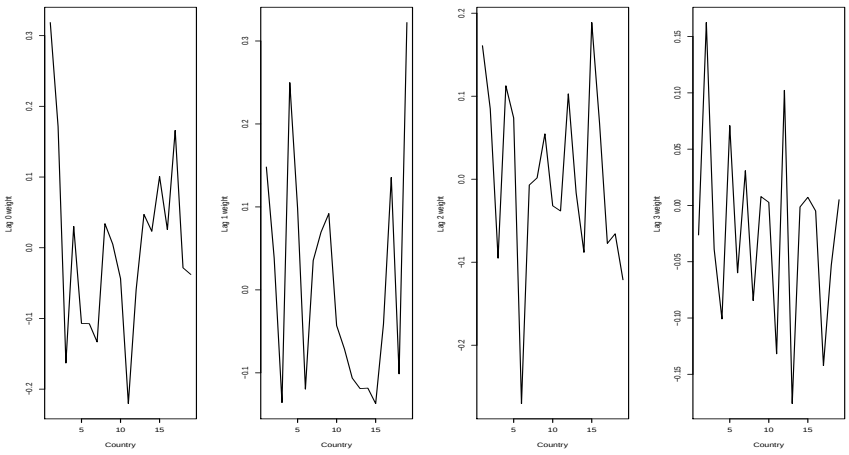

Figure 17: Weights of the lag series, a coefficient, to build the first dynamic factor of the GDP growth rates of the 19 countries of the European Monetary Union. The weights apply to lags from zero to three.

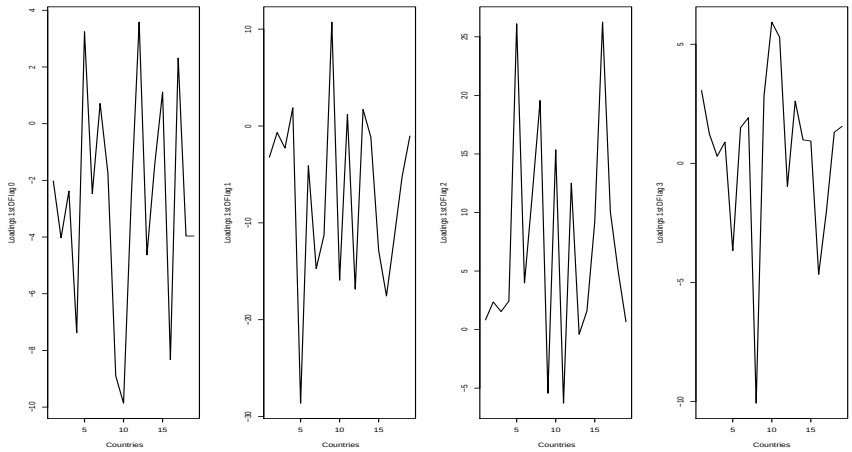

Figure 18: Loadings, (B coefficients) of the lag 1st DF to reconstruct the GDP growth of the European Monetary Union. The loadings apply to lags of the 1st DF from zero to three.

each city of different products, types of shops and sales channel. Then, instead of having a vector of $m = d_1$ components, we have d_2 time series of products, in d_3 type of shops and using d_4 sales channels. The data can be organized as an array, or tensor, of dimension $d_1 \times d_2 \times d_3 \times d_4$ and in each cell of this tensor we have a time series. This kind of situation appears in all fields: pollution time series data can be classified by type of pollution agent, location where the data are measured and method used to collect them. In clinical treatments time series of patients body temperature can be classified by patient age, sex, initial conditions before the illness, family health history and so on. Thus, analyzing these sets of high dimensional time series of several dimensions is becoming and important problem in data analysis.

Wang et al. (2019) proposed a DFM for matrix-valued high-dimensional time series, Gao and Tsay (2021b) a transformed factor model for these data and Chen et al. (2019) allow for constrains in a matrix-variate factor model for time series. These matrix factor models have been generalized to any dimensional array in the tensor DFM proposed by Chen et al. (2020), from a a semiparametric point of view, and Chen et al. (2021), generalizing the approach of Wang et al. (2019). Lam (2021) has studied methods for rank determination in Tensor DFM and a main effect plus interaction integrated proposal for tensor DFM building has been suggested by Peña et al. (2021). Tensor models have also been proposed for vector autoregressive time series modeling (Wang et al., 2021). To simplify the presentation of these models we will concentrate here in the matrix case, the general tensor case can be seen in Chen et al. (2020), Chen et al. (2021) and Peña et al. (2021).

7.1 The Matrix Dynamic Factor Model, MDFM

Suppose that, at each time $1 \leq t \leq T$,we observed a matrix of time series $\mathbf{Y}_t \in \mathbb{R}^{d_1 \times d_2}$, where $y_{t,ij}$ represents the scalar time series at the position (ith, jth) of the data matrix. The simplest way to analyze these data is to assume a factor model for each dimension. Starting with the columns in the data matrix, $\mathbf{y}_{jt} = \mathbf{A}_j \mathbf{f}^c_{jt} + \boldsymbol{\epsilon}_{jt}$, for $j = 1, ..., d_2$, where $\mathbf{A}_j \in \mathbb{R}^{d_1 \times r_j}, \mathbf{f}^c_{jt} \in \mathbb{R}^{r_j \times 1}$ and r_j is the number of factors in the model for jth column. We can write all these models together as a matrix factor model

defining $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{d_2}]$, so that $\mathbf{A} \in \mathbb{R}^{d_1 \times r^c}$ and $r^c = \sum_{j=1}^{d_2} r_j$, and write

$$\mathbf{Y}_t = \mathbf{A}\mathbf{F}_t^c + \mathbf{E}_t, \tag{20}$$

where, $\mathbf{F}_t^c \in \mathbb{R}^{r^c \times d_2}$, is

$$\mathbf{F}_t^c = \begin{bmatrix} \mathbf{f}_{1t}^c & 0, ..., 0 & 0 \\ 0 & 0, .., \mathbf{f}_{jt}^c, ., 0 & 0 \\ 0 & 0, ..., 0 & \mathbf{f}_{d_2 t}^c \end{bmatrix}$$

and $\mathbf{E}_t \in \mathbb{R}^{d_1 \times d_2}$. This model can be written also in the equivalent way, $vec(\mathbf{Y}_t) = (\mathbf{I} \otimes \mathbf{A})vec(\mathbf{F}_t) + vec(\mathbf{E}_t)$, where $\otimes$ is the Kronecker product.

A similar approach can be applied to the rows of the matrices $\mathbf{Y}_t$. Thus, once the column factors have been found we can obtain the residual matrix $\mathbf{E}_t = \mathbf{Y}_t - \mathbf{A}\mathbf{F}_t^c$, in (20), that is expected to be free from column factors, and build the models for the row factors, $\mathbf{y}_{it}' = \mathbf{f}_{it}^{c\prime}\mathbf{B}_i' + \boldsymbol{\upsilon}_{it}'$. We can join all this model together as

$$\mathbf{E}_t = \mathbf{F}_t^r \mathbf{B}' + \mathbf{V}_t,$$

where $\mathbf{F}_t^r \in \mathbb{R}^{d_1 \times s_T}$, s_T is the total number of row factors, and $\mathbf{B}' \in \mathbb{R}^{s_T \times d_2}$ is the row factors matrix loadings. Adding both effects, the complete model will be:

$$\mathbf{Y}_t = \mathbf{A}\mathbf{F}_t^c + \mathbf{F}_t^r \mathbf{B}' + \mathbf{U}_t, \tag{21}$$

where $\mathbf{A} \in \mathbb{R}^{d_1 \times r^c}$ and $r^c = \sum_{j=1}^{d_2} r_j$. Note that each series is affected only by the factors corresponding to the row and column of its location in the data matrix.

As this model may require a large number of parameters in high dimensional settings, an alternative, proposed by Peña et al. (2021), is to assume a two step model in which, first we assume that all the factor models for the rows (columns) are the same and, second, we add some interaction terms to deal with groups of series that do not follow these additive assumptions. Calling $\mathbf{A}_0 \in \mathbb{R}^{d_1 \times m_1}$ and $\mathbf{B}_0 \in \mathbb{R}^{d_2 \times m_2}$ to the common loading matrices for columns and rows with m_1 and m_2 factors, $\mathbf{f}_t^c$, $\mathbf{f}_t^r$ and $\mathbf{1}_{d_2} \in \mathbb{R}^{d_2 \times 1}$ to the vector of ones, we can write the constrained model (21) as

$$\mathbf{Y}_t = \mathbf{1}_{d_2}' \otimes (\mathbf{A}_0 \mathbf{f}_t^c) + \mathbf{1}_{d_1} \otimes (\mathbf{f}_t^r \mathbf{B}_0) + \mathbf{U}_t.$$

We will call this model the constrained main effects model. Calling $a_{i\ell}$ and $b_{j\ell}$ the elements of the ith and jth rows of $\mathbf{A}_0$ and $\mathbf{B}_0$, we have that, in this model, each series is explained by

$$y_{ij,t} = \sum_{\ell=1}^{m_1} a_{i\ell} f_{\ell,t}^c + \sum_{\ell=1}^{m_2} b_{j\ell} f_{\ell,t}^r + u_{ijt}. \tag{22}$$

This constrained main effect model may be too restrictive for some sets of series. Wang et al. (2019) proposed a different formulation of the matrix model assuming that the data are generated by

$$\mathbf{Y}_t = \mathbf{R}\mathbf{F}_t\mathbf{C}' + \mathbf{E}_t, \tag{23}$$

where $\mathbf{R} \in \mathbb{R}^{d_1 \times r_1}$, $\mathbf{F}_t \in \mathbb{R}^{r_1 \times r_2}$, $\mathbf{C} \in \mathbb{R}^{d_2 \times r_2}$, $\mathbf{E}_t \in \mathbb{R}^{d_1 \times d_2}$. Note that this formulation has two important implications. First, all the factors included in the $\mathbf{F}_t$ matrix affect all the series, in all rows and columns. Thus, it is hard to define the factors as associated to columns or rows effects. Second, the loadings in the series y_{ijt} of each factor depend on the position of the series in the data matrix $\mathbf{Y}_t$ and

on the position of the factor in the factor matrix $\mathbf{F}_t$. For instance, if we assume a single factor for all the data and $r_1 = r_2 = 1$ the loading of the factor in the series y_{ijt} is the product of the coefficients R_{i1}, common for all the series in the ith row, and C_{1j}, common for all the series in the jth column. In general, the loadings of the factor f_{kh}, where $k \in (1, ..., r_1)$ indicate the row and $h \in (1, ..., r_2)$ the column of the location of this factor in the $\mathbf{F}_t$ matrix, we have

$$y_{ijt} = \sum_{k=1}^{r_1} \sum_{h=1}^{r_2} R_{ik} C_{hj} f_{kht} + u_{ijt}.$$

I have called interaction effect to this factor model because it is based on the combination of loading defined as product of effects of the common factors affecting all the series.

The factors in this interaction model have only a clear interpretation when they are defined by one of the dimensions. First, assume that $r_1 = d_1$, the model reduces to

$$\mathbf{Y}_t = \mathbf{F}_{1t}\mathbf{C}' + \mathbf{E}_t \tag{24}$$

where that matrix $\mathbf{F}_{1t} \in \mathbb{R}^{d_1 \times r_2}$. In this model each column of $\mathbf{Y}_t$, that we represent by $\mathbf{y}_{.jt}$, for $j = 1, ..., d_2$, is given by

$$\mathbf{y}_{.jt} = c_{j1}\mathbf{f}_{.1t} + ... + c_{jr_J}\mathbf{f}_{.r_jt} + \mathbf{e}_{.jt} \tag{25}$$

and is a linear combination of the columns of $\mathbf{F}_{1t}$ with different loadings, that depend on the column. Note that the columns of $\mathbf{F}_{1t}$ cannot be interpreted as column factors as the r_2 columns column vectors $\mathbf{f}_{.lt}$ are not linked to any of the d_2 columns of the data. However, consider now the ith row of $\mathbf{Y}_t$, $\mathbf{y}'_{i.t}$. Then

$$\mathbf{y}'_{i.t} = \mathbf{f}'_{i.t}\mathbf{C}' + \mathbf{e}'_{i.t} \tag{26}$$

and the ith row follows a standard factor model with loading matrix $\mathbf{C}$ and r_j factors $\mathbf{f}_{i.t}$. Therefore, model (24) can be seen as collecting the factor models for all the rows together and assuming that, although the factors are different for different rows, their number is the same for all rows and the loading matrix of all the rows is also the same.

A similar analysis can be made assuming $r_2 = d_2$, and $\mathbf{C} = \mathbf{I}_{d_2}$ to obtain

$$\mathbf{Y}_t = \mathbf{R}\mathbf{F}_{2t} + \mathbf{E}_t \tag{27}$$

where $\mathbf{F}_{2t} \in \mathbb{R}^{r_I \times d_2}$ and each column $\mathbf{y}_{.jt}$ can be written as

$$\mathbf{y}_{.jt} = \mathbf{R}\mathbf{f}_{jt} + \mathbf{e}_{.jt}$$

where $\mathbf{f}_{jt} \in \mathbb{R}^{r_I \times 1}$ is jth column vector of factors. We see that, again, we assume the same loading matrix for all the columns and the same number of factors, r_I, in all the columns.

The proposal of Peña (2021) is to incorporate the interaction model to the residual of the common main effect model, so that the final model is

$$\mathbf{Y}_t = \mathbf{1}'_{d_2} \otimes (\mathbf{A}_0\mathbf{f}^c_t) + \mathbf{1}_{d_1} \otimes (\mathbf{f}^r_t\mathbf{B}_0) + \mathbf{R}\mathbf{F}_t\mathbf{C}' + \mathbf{U}_t,$$

where each particular time series is explained as

$$y_{ij,t} = \sum_{\ell=1}^{m_1} a_{i\ell} f^c_{\ell,t} + \sum_{\ell=1}^{m_2} b_{j\ell} f^r_{\ell,t} + \sum_{k=1}^{r_1} \sum_{h=1}^{r_1} R_{ik} C_{hj} f_{kht} + u_{ijt}. \tag{28}$$

This model can be considered a main effects plus interaction model and although the number of parameters seems to be larger than (23) in practice, it may not be the case. For instance, if we have series with just main effects model the model (23) may require a larger model to fit the series correctly.

The estimation of the model depends on the hypothesis made for the idiosyncratic components. Assuming that the noises are white noise, as in Wang et al. (2019), the estimation can be carried out by first, finding the number of factors by the rank of some symmetrized sum of lagged covariance matrices and then computing a normalized version of the loading matrices by the eigenvectors of appropriate matrices. If the noises are not white, other approaches, discussed in Peña et al. (2021), can be applied.

8 Conclusions

In the last century most available data were samples of independent observations. Now, with increasing frequency, we observe sequences of dependent data, on time or space. In this article some of the recent procedures created for analyzing these type of data have been reviewed, but many other advances are needed for a better understanding and forecasting of these data. First, we expect new important contributions in the area of spatio-temporal models, with tensor dynamic factor models having and important role for multi-level dependent data. Second, in some fields, as for instance in speech recognition, the data observed is expected to follow some functional form and methods for high dimensional functional data analysis are needed. Also, stronger connections between this literature and the one on time series analysis are required. Third, non linear methods incorporating deep learning and random forests for time series can be very useful for forecasting very disaggregated data, when linearity is not expected. Deep neural networks can be interpreted as continuos non linear version of DFM whereas random forest are very useful with non linear threshold effects, an area in which linear DFM have been already developed. All these advances will increase our accuracy in understanding and forecasting high dimensional dependent data sets.

Acknowledgements

This article responds to the invitation of the Editor of this journal and the President of the Instituto Nacional de Estadística (INE) to write an article as first recipient of the recently created Premio Nacional de Estadística. I am grateful to both of them and, specially, to the President of the INE, Prof. Juan Manuel Rodríguez Poo, for his initiative to create this National Prize to promote the important role of Statistics in solving the new problems that our society is facing today. I am also very grateful to all my coauthors in the research contributions presented here: Andrés Alonso, Stevenson Bolivar, Angela Caro, Pedro Galeano, Fabio Nieto, Pilar Poncela, Esther Ruiz, Dagoberto Saboya, Ezequiel Smucler, Ruey Tsay, Victor Yohai and Ruben Zamar; all of them should be considered as coauthors of this work.

Support for this research from the Agencia Estatal de Investigación of Spain by the project PID2019- 109196GB-I00 is acknowledged.

References

Ahn, Seung C and Alex R Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Alonso, Andrés M, Pedro Galeano, and Daniel Peña (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics 216*(1), 35–52.

Alonso, Andrés M and Daniel Peña (2019). Clustering time series by linear dependency. *Statistics and Computing 29*(4), 655–676.

Ando, Tomohiro and Jushan Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics 31*(1), 163–191.

Ando, Tomohiro and Jushan Bai (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association 112*(519), 1182–1198.

Bai, Jushan and Serena Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, Jushan and Serena Ng (2004). A PANIC attack on unit roots and cointegration. *Econometrica 72*(4), 1127–1177.

Bolívar, Stevenson, Fabio H Nieto, and Daniel Peña (2021). On a new procedure for identifying a dynamic common factor model. *Revista Colombiana de Estadística 44*(1), 1–21.

Bonhomme, Stéphane and Elena Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Breiman, Leo (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science 16*(3), 199–231.

Brillinger, David R (1981). *Time Series Data Analysis and Theory* (Expanded edition ed.). Holden-Day, San Francisco, CA.

Bühlmann, Peter and Sara Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer, Berlin.

Bühlmann, Peter and Sara van de Geer (2018). Statistics for big data: A perspective. *Statistics & Probability Letters 136*, 37–41.

Cao, Ricardo (2017). Ingenuas reflexiones de un estadístico en la era del big data. *Boletín de Estadística e Investigación Operativa 33*(3), 295–321.

Caro, Angela and Daniel Peña (2021). A test for the number of factors in dynamic factor models.

Chen, Elynn Y., Ruey S. Tsay, and Rong Chen (2019). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association 115*, 775–793.

Chen, Elynn Y., Dong Xia, Chencheng Cai, and Jianqing Fan (2020). Semiparametric tensor factor analysis by iteratively projected svd. *arXiv preprint arXiv:2007.02404*.

Chen, Liang, Juan J Dolado, and Jesús Gonzalo (2021). Quantile factor models. *Econometrica 89*(2), 875–910.

Chen, Rong, Dan Yang, and Cun-Hui Zhang (2021). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 1–23.

Correal, María Elsa and Daniel Peña (2008). Threshold dynamic factor model. *Revista Colombiana de Estadística 31*(2), 183–192.

Donoho, David (2017). 50 years of data science. *Journal of Computational and Graphical Statistics 26*(4), 745–766.

Efron, Bradley and Trevor Hastie (2016). *Computer age statistical inference*. Cambridge University Press.

Fan, Jianqing, Jianhua Guo, and Shurong Zheng (2020). Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, In press.

Fan, Jianqing, Fang Han, and Han Liu (2014). Challenges of big data analysis. *National science review 1*(2), 293–314.

Fisher, RA (1925). *Statistical Methods for Research Workers.* Hafner Press, New York.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association 100*(471), 830–840.

Galeano, Pedro and Daniel Peña (2019). Data science, big data and statistics. *TEST 28*(2), 289–329.

Gao, Zhaoxing and Ruey S Tsay (2019). A structural-factor approach to modeling high-dimensional time series and space-time data. *Journal of Time Series Analysis 40*(3), 343–362.

Gao, Zhaoxing and Ruey S Tsay (2021a). Modeling high-dimensional time series: a factor model with dynamically dependent factors and diverging eigenvalues. *Journal of the American Statistical Association*, In press.

Gao, Zhaoxing and Ruey S Tsay (2021b). A two-way transformed factor model for matrix-variate time series. *Econometrics and Statistics*, In press.

Geweke, John (1977). *The dynamic factor analysis of economic time series*. North-Holland.

Hallin, Marc and Roman Liška (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics 163*(1), 29–41.

Kose, M Ayhan, Christopher Otrok, and Charles H Whiteman (2003). International business cycles: World, region, and country-specific factors. *American Economic Review 93*(4), 1216–1239.

Lam, Clifford (2021). Rank determination for time series tensor factor model using correlation thresholding. *Working paper LSE*.

Lam, Clifford and Qiwei Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*, 694–726.

Lin, Chang-Ching and Serena Ng (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods 1*(1), 42–55.

Liu, Xialu and Rong Chen (2020). Threshold factor models for high-dimensional time series. *Journal of Econometrics 216*(1), 53–70.

Nieto, Fabio H, Daniel Peña, and Dagoberto Saboyá (2016). Common seasonality in multivariate time series. *Statistica Sinica 26*, 1389–1410.

Peña, Daniel (2021). On tensor factor models, comment on factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*.

Peña, Daniel and George EP Box (1987). Identifying a simplifying structure in time series. *Journal of the American statistical Association 82*(399), 836–843.

Peña, Daniel and Pilar Poncela (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference 136*(4), 1237–1257.

Peña, Daniel, Pilar Poncela, and Esther Ruiz (Eds.) (2021). *Nuevos Métodos de Predicción Económica con datos masivos*. FUNCAS.

Peña, Daniel, Ezequiel Smucler, and Victor J Yohai (2019). Forecasting multiple time series with one-sided dynamic principal components. *Journal of the American Statistical Association 114*, 1683–1694.

Peña, Daniel, Ezequiel Smucler, and Victor J Yohai (2020). gdpc: an r package for generalized dynamic principal components. *Journal of Statistical Software 92*(1), 1–23.

Peña, Daniel, Ezequiel Smucler, and Victor J Yohai (2021). Sparse estimation of dynamic principal components for forecasting high-dimensional time series. *International Journal of Forecasting 37*(4), 1498–1508.

Peña, Daniel and Ruey S Tsay (2021). *Statistical Learning for Big Dependent Data*. John Wiley & Sons.

Peña, Daniel, Ruey S Tsay, and Ruben Zamar (2019). Empirical dynamic quantiles for visualization of high-dimensional time series. *Technometrics 61*, 429–444.

Peña, Daniel and Victor J Yohai (2016). Generalized dynamic principal components. *Journal of the American Statistical Association 111*(515), 1121–1131.

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

Torrecilla, José L and Juan Romo (2018). Data learning from big data. *Statistics & Probability Letters 136*, 15–19.

Tsay, Ruey S (2014). *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, Hoboken, NJ.

Wang, Dong, Xialu Liu, and Rong Chen (2019). Factor models for matrix-valued high-dimensional time series. *Journal of econometrics 208*(1), 231–248.

Wang, Di, Yao Zheng, Heng Lian, and Guodong Li (2021). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, In press.

Wang, Peng (2010). Large dimensional factor models with a multi-level factor structure: identification, estimation and inference. *Working paper, Department of Economics, HKUST*.

SPANISH JOURNAL OF STATISTICS
Vol. **3** No. 1 2021, Pages 37–44
doi:https://doi.org/10.37830/SJS.2021.1.03

REGULAR ARTICLE

Misreported longitudinal data in epidemiology: Review of mixture-based advances and current challenges

David Moriña[1, 2], Amanda Fernández-Fontelo[3], Alejandra Cabaña[4], Argimiro Arratia[5], Pedro Puig[2, 4]

[1] Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona, dmorina@ub.edu
[2] Centre de Recerca Matemàtica (CRM)
[3] Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin
[4] Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB)
[5] Department of Computer Science, Universitat Politècnica de Catalunya (UPC)

Received: October 1, 2021. Accepted: December 20, 2021.

Abstract: The problem of dealing with misreported data is very common in a wide range of contexts and for different reasons. This has been and still is an important issue for data analysts and statisticians as not accounting for it could led to biased estimates and conclusions, and in many cases that would have implications in a posterior decision making process, as we all have seen in the current worldwide Covid-19 pandemic. In the last few years, many approaches have been proposed in the literature to accomodate data presenting this issue, especially in the fields of epidemiology and public health but also in other areas as social science. In this work, a comprehensive review of the recently proposed methods based on mixture models for longitudinal data (correlated and uncorrelated) is presented and several examples of application are discussed, including several approaches to the burden of Covid-19 infection cases in Spain and different approaches to deal with underreported registries of human papillomavirus infections and genital warts in Catalunya.

Keywords: epidemiology, modelling, misreported data, longitudinal data, public health

MSC: 37M10, 92D30, 46N30, 62-07

1 Introduction

Dealing with misreported data is a very common problem in many fields. The most usual issue is facing data that is only partially registered or underreported, and this type of misreporting has received some attention in the recent years, with alternative approaches being proposed in epidemiology, social and biomedical research, among many other contexts (Bernard et al., 2014; Arendt et al.,

2013; Rosenman et al., 2006; Alfonso et al., 2015; Winkelmann, 1996). The opposite phenomenon, i. e. dealing with overreported data, occurs with less frequency although also arises in some cases (see for instance Mehta (2018)), and has been less studied. The mechanisms leading to misreported data are diverse, and so are the methods proposed in the literature to overcome this issue, ranging from Markov chain Monte-Carlo (Winkelmann, 1996) approaches to time series analysis (Fernández-Fontelo et al., 2016; Fernández-Fontelo et al., 2019) or spatio-temporal modelling (Stoner et al., 2019). From the computational point of view, a new R (R Core Team, 2019) package useful to fit endemic-epidemic models based on approximate maximum likelihood to underreported count data (Bracher, 2019) and the package *MisRepARMA* (Moriña et al., 2021), able to fit misreported time series and to reconstruct the most likely actual series were recently introduced. Additionally a web application allowing its users to fit misreported AutoRegressive Moving Average (ARMA) time series model has been recently published and is accessible at https://dmorina.shinyapps.io/MisRepARMA/.

Misreported data might potentially lead to biased inference, as it may invalidate the assumptions of standard models. For instance, Winkelmann (1996) explores a Markov chain Monte Carlo based methodology to study worker absenteeism where two sources of underreporting are detected: an insufficient surveillance mechanism (if the data are provided by the employer), and a lack of memory if the time series is reconstructed retrospectively by the worker. Also in public health context, it is well known that some diseases related to occupational or food exposures have been traditionally underreported (Alfonso et al., 2015; Rosenman et al., 2006; Arendt et al., 2013). Of course in that case there might be several sources of underreporting, including accuracy of public health registries, political or economical interests, among others.

This review is focused on methods recently proposed to deal with permanent underreporting, assuming that there are more cases in the population than the acknowledged ones. In many applications, however, underreporting occurs by a delay in reporting, but as time passes, the counts became more and more complete, as studied in Höhle and an der Heiden (2014). This is also the case of the public health information systems of many countries regarding the Covid-19 pandemic, as they are being updated when the results of the tests are available, but we have seen that unfortunately, waiting until the information is complete is often too late to make the optimal decisions to handle the pandemic in a proper way.

From the methodological point of view, the first attempts to overcome this problem in the context of longitudinal data were based on count time series, with a growing popularity due to the limited performance of the classical continuous time series approach when the variable of interest consists of low counts (Fernández-Fontelo et al., 2016; Fernández-Fontelo et al., 2019). Similar ideas were used to analyze uncorrelated longitudinal data (Moriña et al., 2021). Another additional issue to deal with inspired by the Covid-19 data is non-stationarity. This issue has been treated in Fernández-Fontelo et al. (2020) by means of discrete time series.

2 Proposed models and examples

Some of the most recently proposed methodologies based on mixture models and able to deal with misreported data are described in this Section, together with some examples of application.

2.1 Longitudinal uncorrelated data

In this context, the simplest case is when the longitudinal data we are dealing with present no temporal correlation. This is the case covered in Moriña et al. (2021), where a model is proposed and used to analyze the underreporting issue in genital warts (GW) record in Catalonia (northeastern

Spain). In the particular case of sexually transmitted diseases, it is very important to have reliable information due to their remarkable morbidity, and therefore, the importance of controlling trends over time and priority setting (see McCormack and Koons (2019) for a comprehensive discussion focused on developing countries).

Several models proposed in the literature able to deal with misreported longitudinal data are based on more sophisticated versions of the following schema, replacing the usual product by other operators in the discrete case (usually by the binomial thinning operator, see Section 2.2).

Consider X_t the series of real GW incidence, where $t = 1, 2, \ldots$ is the time, following a normal distribution with mean μ and variance σ^2. In this setting, the X_t process cannot be directly observed, and all we can see is a part of it, expressed as

$$Y_t = \begin{cases} X_t & \text{with probability } 1 - \omega_t \\ q \cdot X_t & \text{with probability } \omega_t \end{cases} \tag{1}$$

The interpretation of the parameters in Eq. (1) is straightforward: q is the overall intensity of misreporting (if $0 < q < 1$ the observed process Y_t would be underreported while if $q > 1$ the observed process Y_t would be overreported). The parameter ω can be interpreted as the overall frequency of misreporting (proportion of misreported observations).

The series Y_t represents the registered values corresponding to GW incidence. According to Eq. (1), the registered observations series Y_t is a mixture of two normally distributed random variables $Y_t = Y_{1t}$ with probability $(1 - \omega_t)$ and $Y_t = Y_{2t}$ with probability ω_t, where Y_{1t} coincides with the unobserved process X_t and Y_{2t} is a normal random variable with mean $q \cdot \mu$ and variance $q^2 \cdot \sigma^2$. The parameter ω_t is modeled as $logit(\omega_t) = \alpha_0 + \alpha_1 \cdot t$ and can be interpreted as the frequency of underreporting at a time t, while q can be interpreted as the intensity of such underreporting, both taking values between 0 and 1. When $q = 0$ the observed incidence is $Y_t = 0$ and when $q = 1$ there is no underreporting. A value of ω_t equal to 0 indicates that the observed value at time t is not underreported, and a value of ω_t equal to 1 means that underreporting is for sure happening. In order to detect potential differences in GW incidence depending on sex (men and women) and age group (16-29 and 30-94), these covariates were included in the model, so the mean of the observed process Y_{1t} was modeled as $\mu_{1,t} = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s$ (where a is the age, s is the sex and $a * s$ is the interaction between age and sex). The average of the second component Y_{2t} can be recovered as $\mu_{2,t} = q \cdot (\beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot s + \beta_4 \cdot a * s)$.

In the particular case of these data, after diagnostic checks a seasonal cycle of 3 months was observed. This behavior was included in the model through the following trigonometric funtion $f(t) = \beta_5 \cdot sin(\frac{2 \cdot \pi \cdot t}{3}) + \beta_6 \cdot cos(\frac{2 \cdot \pi \cdot t}{3})$ on the terms $\mu_{1,t}$ and $\mu_{2,t}$.

The loglikelihood function corresponding to the final model considered in this paper can be written as

$$l(Y, \theta) = \sum_{t=1}^{n} \log \left((1 - \omega_t) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_t - \mu_{1,t})^2}{2\sigma^2}} + \omega_t \frac{1}{\sqrt{2\pi} q \sigma} e^{\frac{(y_t - \mu_{2,t})^2}{2q^2\sigma^2}} \right), \tag{2}$$

where $Y = y_1, \ldots, y_n$ is the observed series, $\theta = (\alpha_0, \alpha_1, \gamma, \beta_0, \ldots, \beta_6, \sigma)$, $\omega_t = \frac{e^{\alpha_0 + \alpha_1 t}}{1 + e^{\alpha_0 + \alpha_1 t}}$, $q = \frac{e^{\gamma}}{1 + e^{\gamma}}$ and $\mu_{1,t}$ and $\mu_{2,t}$ are as defined before.

The estimates and associated standard errors can be obtained maximizing Eq. 2 in the usual way.

The results of this work show that in relative terms, the underreporting issue has a deeper impact on people over 30 years old (where GW incidence is lower), especially among women. Nonetheless, the relative difference between registered and estimated annual averages range between 13.3% and 24.9%. It is also remarkable that the quality of SIDIAP register regarding GW in Catalunya has been

significantly improving during the study period, as the frequency of underreported observations has been decreasing over time, from around 95% in 2009 to 21% in 2016.

2.2 Discrete stationary time series

As stated before, the underreporting structure defined in Eq. 1 has been adapted to the discrete time series case in Fernández-Fontelo et al. (2016), by means of the *binomial thinning* operator. The proposed model assumes, as before, that the actual values of the series X_t are not fully observed and follow an integer-valued autoregressive model of order 1 (INAR(1)) (Jung and Tremayne, 2006):

$$X_t = \alpha \circ X_{t-1} + W_t(\lambda), \tag{3}$$

where $\alpha \in (0,1)$ and W_t is assumed to follow a Poisson distribution with a fixed mean λ. In addition, X_{t-1} and W_t are assumed to be independent at any time t. The $\circ$ operator in expression (3), called *binomial thinning*, is defined as follows,

$$\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} Y_i, \tag{4}$$

where Y_i are iid Bernoulli random variables with a probability of success equal to α. Therefore, if $X_{t-1} = x_{t-1}$, then $\alpha \circ x_{t-1}$ is binomially distributed with the number of successes equal to x_{t-1}.

The observed process Y_t is then defined as

$$Y_t = \begin{cases} X_t & : \text{with probability } 1-\omega \\ q \circ X_t & : \text{with probability } \omega \end{cases} \tag{5}$$

The interpretation of the parameters is very similar to that of the previous models (notice that in this case the frequency of underreported observations ω is not time dependent), and two methods for estimation are provided in Fernández-Fontelo et al. (2016). The first method is based on the fact that the marginal distribution of Y_t is a mixture of two Poisson distributions with parameters $\frac{\lambda}{1-\alpha}$ and $\frac{q\lambda}{1-\alpha}$, with probabilities $1-\omega$ and ω, so an estimation of $\frac{\lambda}{1-\alpha}$, $\frac{q\lambda}{1-\alpha}$ and ω can be found by fitting the whole observed series using a mixture of two Poisson distributions. The parameters α and λ can be estimated using for instance the first autocorrelation coefficients, which are explicitly derived in the paper.

The estimates obtained by using the previously described method could be used as initial values to improve the speed of the maximum likelihood algorithm, based on an adapted version of the forward algorithm (Zucchini and MacDonald, 2009).

One strength of this methodology is that it allows to recover the most likely hidden process X_t using the Viterbi algorithm (Viterbi, 1967; Forney, D. G., 1973), which is implemented in several R packages for situations with a finite number of states, and has to be adapted to cover the present case also.

This model was used in Fernández-Fontelo et al. (2016) to analyze three examples in the field of public health, illustrating the model performance in the case of a stationary series (weekly HPV cases in Girona in the period 2010-2014) but also in the presence of trends (number of annual deaths by mesothelioma in Great Bretain in the period 1968-2013 and number of annual cases of botulism in Canada in the period 1970-2013). According to the reported results, it can be seen that although only an average of 1.27 HPV cases per week were registered, the most likely reconstructed series estimated an average of 3.36 cases per week. Regarding botulism in Canada, an average of 9 cases per

year were registered although a proportion of $\hat{\omega} = 0.671$ underreported observations was estimated by the model. Mesothelioma is a disease related to asbestos exposure and is difficult to diagnose so although an average of 14 annual deaths in Great Britain in the period 1968-2013 it is supposed to be highly underreported. It is confirmed by an estimated proportion of $\hat{\omega} = 0.930$ underreported observations.

A more sophisticated underreporting structure was introduced in Fernández-Fontelo et al. (2019). In this case, the hidden process X_t is again assumed to follow an INAR(1) structure but the observed series Y_t is defined by

$$Y_t = \begin{cases} X_t & : \text{If } I_t = 0, \\ q \circ X_t & : \text{If } I_t = 1, \end{cases} \tag{6}$$

where $\circ$ is again the binomial thinning operator and I_t is a binary discrete time Markov chain indicating whether the observation Y_t is underreported or not. The difference with respect to the previous case is that in this case the states of underreporting I_t are serially dependent. Two methods of parameter estimation are provided in the model, one based on moments method and another based on maximum likelihood, and the most likely hidden process can be recovered by a Viterbi-like algorithm.

This model was used to analyze the underreporting issue on the gender-based violence reports in several Galician districts, and it was shown that as expected, it is an important problem related to these kind of data that should be accounted for, but also that the underreporting is severely non-uniformly distributed across the considered districts, as the frequency of underreported observations ranged from $\hat{\omega} = 0.078$ (very few observations are underreported in this area) to $\hat{\omega} = 0.976$ (almost all observations are underreported).

2.3 Discrete non-stationary time series

The outbreak of the Covid-19 pandemic in 2020 has made very clear that models able to handle non-stationary time series with potential underreporting issues were absolutely necessary to help decision-makers. The previously discussed models for discrete stationary time series can be extended in several ways to be able to deal with non-stationary discrete time series. A possible way is explored in Fernández-Fontelo et al. (2020), where the authors consider that the true (unobserved) counts X_t follow again an INAR(1) structure and the observed process Y_t is, as before, defined by

$$Y_t = \begin{cases} X_t & \text{with probability } 1-\omega \\ q \circ X_t & \text{with probability } \omega. \end{cases} \tag{7}$$

The difference with respect to the previous models for discrete time series is that the mean of the innovations of the hidden process X_t and the underreporting parameter q are modelled as functions of time and thus allowing for non-time-homogeneous processes, and that only maximum likelihood estimation is suitable in the case of non-stationary time series. Both underreporting-related parameters ω and q could be considered time-dependent, but in this work, to avoid potential convergence issues, only time-dependent q is considered. In particular, as a weekly seasonal behavior was observed in the reported data and to ensure that the estimates are within the interval $(0, 1)$, it was adjusted by means of the logistic function

$$q_t = \frac{\exp\left(\gamma_0 + \gamma_1 t + \gamma_2 \sin\left(\frac{2\pi t}{7}\right) + \gamma_3 \cos\left(\frac{2\pi t}{7}\right)\right)}{1 + \exp\left(\gamma_0 + \gamma_1 t + \gamma_2 \sin\left(\frac{2\pi t}{7}\right) + \gamma_3 \cos\left(\frac{2\pi t}{7}\right)\right)}, \tag{8}$$

where γ_1 indicates whether q increases or decreases over time and γ_2 and γ_3 indicate whether the series has a seasonal pattern with period $p = 7$ (weekly). Notice that if $\gamma_1 = \gamma_2 = \gamma_3 = 0$, then the previous logistic function becomes constant and thus $q_t = q$, resulting in the model (9). Hence, considering this function for the intensity of the underreporting, the underreporting process in this model was defined by:

$$Y_t = \begin{cases} X_t & \text{with probability } 1-\omega \\ q_t \circ X_t & \text{with probability } \omega. \end{cases} \tag{9}$$

At this point, however, the model is not able to handle non-stationarity yet. In order to be able to fit non-stationary discrete time series and answer the main question of researchers facing an epidemic, the innovations of the hidden INAR(1) process X_t are linked to a compartimental epidemiological model. In this case, a Susceptible-Infected-Recovered (SIR) model was used with that purpose. This model considers three compartments were individuals are included in at each time t: Those who are healthy but susceptible to get the disease ($S(t)$), those who are infected and thus transmitters of the disease ($I(t)$), and those individuals who have been removed from the system and will not get infected again ($R(t)$) (Anderson and May, 1992; Vynnycky and White, 2010). The parameters of interest are the infection rate β, the removal rate γ, and the total susceptible population N.

It can be seen that the number of individuals affected by the disease at time t can be represented by:

$$A(t) = \frac{M^* A(0) \mathrm{e}^{kt}}{M^* + A(0)(\mathrm{e}^{kt} - 1)}, \tag{10}$$

where $k = \beta - \gamma$ and $M^* = \frac{N(\beta-\gamma)}{\beta-\gamma/2}$.

The information on the spread of the disease can be included in the hidden process X_t by considering that the expectation of the innovations is not constant as in the previous models but time varying as $\lambda_t = \text{new}(t) = A(t) - A(t-1)$, where new$(t)$ are the new affected cases at time t.

Therefore, the unobserved process X_t becomes:

$$X_t = \alpha \circ X_{t-1} + W_t(\lambda_t), \tag{11}$$

As in the previous models described, an appropriate version of the Viterbi algorithm could be used to reconstruct the most likely hidden process, which is specially interesting in this case as it provides a more realistic picture of the pandemic situation in a given moment of time.

Additionally, two forecasting methods are provided; one based on average point predictions given the sample observations and another based on the conditional distribution of Y_{t+k} given the last value of the latent process X_t. Standard errors for the first forecasting method could be obtained using a numerical mechanism as the Delta method. For the second method forecasts, prediction regions of size $1-\alpha$ can be found through the values l (lower limit) and u (upper limit) satisfying $\sum_{j=1}^{l} \mathrm{P}(Y_{t+k} = j | X_t = x_t) \approx \frac{\alpha}{2}$ and $\sum_{j=1}^{u} \mathrm{P}(Y_{t+k} = j | X_t = x_t) \approx 1 - \frac{\alpha}{2}$.

This model was used to analyze the Covid-19 data registered in several regions in Spain, and the results reported in the paper confirm that the underreporting issue is indeed present in Covid-19 data from various regions conditioned to different management, policies, and climate conditions, and that it also varies across geographic areas, with registered coverages ranging from 33.7 % in Ourense (Galicia) to 71.8 % in Málaga (Andalucía) of the estimated cases.

3 Discussion

Facing misreported information from public health registers is very common in many situations, for instance data regarding potentially asymptomatic diseases like HPV infection or Covid-19, or difficult to diagnose as mesothelioma.

One of the lessons that should certainly be learned from the current Covid-19 pandemic is that it is crucial to provide researchers with reliable data under extremely complex circumstances, in order to be able to assure public health decision makers are provided with the most reliable information at any time. When this is by no ways possible, the issue should be at least taken into account by using a model capable of accommodating underreported data like the one used in this study.

Although several efforts have been done in this direction in the last years, several challenges still remain. From the methodological point of view, there are a few works dealing with underreported discrete time series, but they are a bit limited for instance in the allowed structure for the hidden process. These models could be extended in different ways, such as considering more complex correlation structures in the underlying process (for instance INAR(p) or INARMA(p, q) structures), or considering more general thinning operators for representing the observed process.

Also from the methodological point of view, the extension of the model introduced in Moriña et al. (2021) to underreported continuous time series is not available yet, nor for stationary or non-stationary processes. Additionally, the considered methods could certainly be useful in the development of new, more general methodologies able to deal with overreporting as well, as this is another issue that, although with minor frequency, appears in the practice of epidemiology and public health.

The availability of packages for commonly used software as R (R Core Team, 2019) would also help to make these methods reach a wider public of potential users, and some efforts in this sense are currently been done (Bracher (2019) for instance).

From the applied point of view, it would be very interesting to use these kind of models to analyze other issues that might be potentially underreported and to analyze more thoroughly the examples used to illustrate the performance of the discussed models. For instance, the differences across geographic areas observed in Fernández-Fontelo et al. (2019) related to underregistered reports of gender-based violence could be better explained if covariates (as rurality index, socioeconomic variables, spatial correlation...) were included in the model.

References

Alfonso, J. H., E. K. Løvseth, Y. Samant, and J. Holm (2015). Work-related skin diseases in Norway may be underreported: data from 2000 to 2013. *Contact Dermatitis 72*(6), 409–412.

Anderson, R. M. and R. M. May (1992). *Infectious diseases of humans: dynamics and control*. Oxford University Press.

Arendt, S., L. Rajagopal, C. Strohbehn, N. Stokes, J. Meyer, and S. Mandernach (2013). Reporting of foodborne illness by U.S. consumers and healthcare professionals. *International journal of environmental research and public health 10*(8), 3684–3714.

Bernard, H., D. Werber, and M. Höhle (2014). Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104: H4 in 2011 - a time series analysis. *BMC Infectious Diseases 14*(1).

Bracher, J. (2019). hhh4u: Fit an endemic-epidemic model to underreported data.

Fernández-Fontelo, A., A. Cabaña, H. Joe, P. Puig, and D. Moriña (2019). Untangling serially dependent underreported count data for gender-based violence. *Statistics in Medicine 38*(22), 4404–4422.

Fernández-Fontelo, A., A. Cabaña, P. Puig, and D. Moriña (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine 35*(26), 4875–4890.

Fernández-Fontelo, A., D. Moriña, A. Cabaña, A. Arratia, and P. Puig (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE 15*, e0242956.

Forney, D. G. (1973). The viterbi algorithm. *Proceedings of the IEEE 61*(3), 268–278.

Höhle, M. and M. an der Heiden (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics 70*(4), 993–1002.

Jung, R. C. and A. R. Tremayne (2006). Binomial thinning models for integer time series. *Statistical Modelling 6*(2), 81–96.

McCormack, D. and K. Koons (2019). Sexually Transmitted Infections. *Emergency Medicine Clinics of North America 37*(4), 725–738.

Mehta, R. (2018). Allergy and Asthma: Food Allergies. *FP essentials 472*, 16–19.

Moriña, D., A. Fernández-Fontelo, A. Cabaña, and P. Puig (2021). MisRepARMA: Misreported time series analysis.

Moriña, D., A. Fernández-Fontelo, A. Cabaña, P. Puig, L. Monfil, M. Brotons, and M. Diaz (2021). Quantifying the under-reporting of genital warts cases. *BMC Medical Research Methodology 21*(1), 6.

R Core Team (2019). R: A Language and Environment for Statistical Computing.

Rosenman, K. D., A. Kalush, M. J. Reilly, J. C. Gardiner, M. Reeves, and Z. Luo (2006). How much work-related injury and illness is missed by the current national surveillance system? *Journal of Occupational and Environmental Medicine 48*(4), 357–365.

Stoner, O., T. Economou, and G. Drummond Marques da Silva (2019). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association 114*(528), 1481–1492.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*(2), 260–269.

Vynnycky, E. and R. White (2010). *An introduction to infectious disease modelling*. Oxford University Press.

Winkelmann, R. (1996). Markov Chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics 21*(4), 575–587.

Zucchini, W. and I. L. MacDonald (2009). *Hidden Markov Models for Time Series : An Introduction Using R*. Chapman & Hall/CRC.

SPANISH JOURNAL OF STATISTICS
Vol. **3** No. 1 2021, Pages 45–53
doi:https://doi.org/10.37830/SJS.2021.1.04

REGULAR ARTICLE

A note on explicit expressions for moments of order statistics

Fredy Castellares[1], Artur J. Lemonte[2], Marcos A.C. Santos[3]

[1] Universidade Federal de Minas Gerais, Belo Horizonte/MG, Brazil, fwcc29@gmail.com
[2] Universidade Federal do Rio Grande do Norte, Natal/RN, Brazil, arturlemonte@gmail.com
[3] Universidade Federal de Minas Gerais, Belo Horizonte/MG, Brazil, msantos@est.ufmg.br

Received: August 30, 2021. Accepted: December 20, 2021.

Abstract: Closed-form expressions for moments of order statistics from the normal, log-normal, gamma and beta distributions were provided in the statistics literature. In particular, the explicit expressions involve the Lauricella function of type A, and the generalized Kampé de Fériet function. We note that the expressions provided by the author do not appear correct, which implies that the expressions cannot be recommended to users. An alternative closed-form expression for moments of order statistics is then provided. We also consider numerical studies to show that the formulas we provide deliver satisfactory results.

Keywords: Kampé de Fériet function, Lauricella function

MSC: 60E05

1 Moments of order statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from some random variable X with probability density function (PDF) given by f, and cumulative distribution function (CDF) given by F. Let $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$ denote the corresponding order statistics. The PDF of $X_{r:n}$ is given by (see, for example, David and Nagaraja, 2003)

$$f_{r:n}(x) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} F^{m+r-1}(x) f(x), \quad x \in \mathbb{R}, \tag{1}$$

where $r = 1, 2, \ldots, n$. The moments of order statistics are

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} \int_{-\infty}^{\infty} x^k F^{m+r-1}(x) f(x) dx,$$

where $k = 1, 2, \ldots$ and $r = 1, 2, \ldots, n$; that is,

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} I(k, m-r+1),$$

where

$$I(k,s) = \int_{-\infty}^{\infty} x^k F^s(x) f(x) dx, \quad k, s \in \mathbb{N}.$$

In general, the above integral has no closed-form analytical solution and, hence, the use of special functions like Lauricella function of type A and (generalized) Kampé de Fériet function to express $I(k,s)$ in closed-form are welcome. However, one has to use these special functions with some caution, mainly with respect to the convergence radius of these functions, otherwise, they will be divergent and so the use of them becomes inviable.

The Lauricella function of type A Exton (1978) is defined as

$$F_A^{(n)}(a; b_1, \ldots, b_n, c_1, \ldots, c_n; x_1, \ldots, x_n) = \sum_{i_1, \ldots, i_n = 0}^{\infty} \frac{(a)_{i_1 + \cdots + i_n} (b_1)_{i_1} \cdots (b_n)_{i_n}}{(c_1)_{i_1} \cdots (c_n)_{i_n} i_1! \cdots i_n!} x_1^{i_1} \cdots x_n^{i_n}, \tag{2}$$

where $(q)_n$ is the (rising) Pochhammer symbol, i.e. $(q)_0 = 1$, and $(q)_n = q(q+1)\cdots(q+n-1)$ for $n \geq 1$. We have that the convergence radius of the Lauricella function of type A is

$$|x_1| + \cdots + |x_n| < 1.$$

Therefore, if the above condition is not satisfied, the Lauricella function of type A will divergent for any values of a, $b_1, \ldots, b_n$ and $c_1, \ldots, c_n$. For example, if the n arguments of the Lauricella function in expression (2) are all equal to -1, we have $|x_1| + \cdots + |x_n| = n \geq 1$ and, hence, we have that

$$F_A^{(n)}(a; b_1, \ldots, b_n, c_1, \ldots, c_n; -1, -1, \ldots, -1) = +\infty.$$

2 Results from Nadarajah (2008)

In the following, some closed-form expressions for $\mathbb{E}(X_{r:n}^k)$ derived by Nadarajah (2008) are presented.

2.1 Moments of normal order statistics

Nadarajah (2008, eq. (12)) arrived at the following closed-form expression for $\mathbb{E}(X_{r:n}^k)$:

$$\begin{aligned} \mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{l=0}^{n-r} \binom{n-r}{l} \left(-\frac{1}{2}\right)^l \sum_{\substack{p=0 \\ p+k \text{even}}}^{r+l-1} \binom{r+l-1}{p} \\ \times (\pi)^{(1-p)/2} 2^p \Gamma\left(\frac{k+p+1}{2}\right) F_A^{(p)}\left(\frac{k+p+1}{2}; \frac{1}{2}, \ldots, \frac{1}{2}; \frac{3}{2}, \ldots, \frac{3}{2}; -1, \ldots, -1\right), \end{aligned} \tag{3}$$

where $\Gamma(\cdot)$ is the complete gamma function. Note that the (p) arguments of the Lauricella function in expression (3) are all equal to -1. From (3), it is evident that

$$F_A^{(p)}\left(\frac{k+p+1}{2}; \frac{1}{2}, \ldots, \frac{1}{2}; \frac{3}{2}, \ldots, \frac{3}{2}; -1, \ldots, -1\right) = +\infty,$$

which invalidates the formula (3) for $\mathbb{E}(X_{r:n}^k)$ obtained by Nadarajah (2008).

2.2 Moments of lognormal order statistics

Nadarajah (2008, eq. (21)) arrived at the following closed-form expression for $\mathbb{E}(X_{r:n}^k)$:

$$\mathbb{E}(X_{r:n}^k) = \frac{n!2^{1-r}}{(r-1)!(n-r)!}\exp\left(\frac{k^2}{2}\right)\sum_{l=0}^{n-r}\left(-\frac{1}{2}\right)^l\binom{n-r}{l}\sum_{p=0}^{r+l-1}\binom{r+l-1}{p}\left(-\frac{2}{\sqrt{\pi}}\right)^p \times \mathbb{E}\left[\left(\frac{N\sqrt{2}-k}{2\sqrt{2}}\right)^p F_A^{(p)}\left(\frac{1}{2},\ldots,\frac{1}{2};\frac{3}{2},\ldots,\frac{3}{2};-\frac{(N\sqrt{2}-k)^2}{8},\ldots,-\frac{(N\sqrt{2}-k)^2}{8}\right)\right],$$

where N is a standard normal random variable. The above expression for $\mathbb{E}(X_{r:n}^k)$ does not converge. Note that the Lauricella function depends on random arguments, given by the quantity

$$\frac{(N\sqrt{2}-k)^2}{8}.$$

In this case, it is not possible to ensure that

$$\left|\frac{(N\sqrt{2}-k)^2}{8}\right| + \cdots + \left|\frac{(N\sqrt{2}-k)^2}{8}\right| < 1,$$

which is the convergence radius of the Lauricella function.

2.3 Moments of gamma order statistics

Nadarajah (2008, eq. (26)) arrived at the following closed-form expression for $\mathbb{E}(X_{r:n}^k)$:

$$\begin{aligned}\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!}\sum_{l=0}^{n-r}\binom{n-r}{l}\Gamma(\alpha)^{-r-l}\alpha^{1-r-l}\Gamma(k+\alpha(r+l)) \\ \times F_A^{(r+l-1)}(k+\alpha(r+l);\alpha,\ldots,\alpha,\alpha+1,\ldots,\alpha+1;-1,\ldots,-1).\end{aligned} \tag{4}$$

Note that the $(r+l-1)$ arguments of the Lauricella function of type A in expression (4) are all equal to -1, which implies that we can have $|x_1|+\cdots+|x_{r+l-1}| = r+l-1 > 1$ for $r = 1,2,\ldots,n$ and $l = 0,1,,\ldots,n-r$. In this case, the Lauricella function of type A is also divergent, which invalidates expression (4) for $\mathbb{E}(X_{r:n}^k)$ derived by Nadarajah (2008).

2.4 Moments of beta order statistics

Nadarajah (2008, eq. (31)) arrived at the following closed-form expression for $\mathbb{E}(X_{r:n}^k)$:

$$\begin{aligned}\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!}\sum_{l=0}^{n-r}(-1)^l\binom{n-r}{l}a^{1-r-l}B(a,b)^{-r-l}B(b,k+a(r+l)) \\ \times F_{1:1}^{1:2}\left(\begin{matrix}(k+a(r+l)):(1-b,a);\cdots:(1-b,a);\\(b+k+a(r+l)):(1+a);\cdots:(1+a);\end{matrix}1,1,\ldots,1\right),\end{aligned} \tag{5}$$

where $B(\cdot,\cdot)$ is the beta function, and $F_{1:1}^{1:2}(\cdot)$ is the Kampé de Fériet function, which is a generalization of the generalized hypergeometric series, introduced by Joseph Kampé de Fériet; see, for

example, Exton (1978, eq. (1.4.2)). The $F_{1:1}^{1:2}(\cdot)$ function is of the form

$$F_{C:D}^{A:B} := F_{1:1}^{1:2}\left(\begin{array}{c}(a)\colon \overbrace{(b_1,b_2);\cdots:(b_1,b_2)}^{n};\\ (c)\colon \underbrace{(d);\cdots:(d)}_{n};\end{array} x_1,x_2,\ldots,x_n\right) = \sum_{i_1=0}^{\infty}\cdots\sum_{i_n=0}^{\infty}\frac{(a)_{i_1+\cdots+i_n}\prod_{j=1}^{n}(b_1)_{i_j}\prod_{j=1}^{n}(b_2)_{i_j}}{(c)_{i_1+\cdots+i_n}\prod_{j=1}^{n}(d)_{i_j}}\frac{x_1^{i_1}\cdots x_n^{i_n}}{i_1!\cdots i_n!}.$$

The convergence radius of the Kampé de Fériet function is as follows (see, for example, Exton, 1978, p. 26): if $A+B=C+D+1$ and $A=C$, the Kampé de Fériet function converges at the region

$$|x_1|<1,\ldots,|x_n|<1.$$

However, we have that all arguments of the Kampé de Fériet function in expression (5) are all equal to 1, which implies that

$$F_{1:1}^{1:2}\left(\begin{array}{l}(k+a(r+l))\colon(1-b,a);\cdots:(1-b,a);\\ (b+k+a(r+l))\colon(1+a);\cdots:(1+a);\end{array} 1,1,\cdots,1\right)=+\infty.$$

Therefore, the expression (5) for $\mathbb{E}(X_{r:n}^k)$ derived by Nadarajah (2008) is not valid.

3 An alternative expression

Here, let $X_1, X_2, \ldots, X_n$ be a random sample of size n identically distributed sampled from a random variable X, with CDF given by F. Let $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$ denote the corresponding order statistics. Let S be the suport of X. Assume that the CDF F can be evaluated using a convergent power series expansion (with convergence radius $\rho > 0$) of the form

$$F(x)=\sum_{j=0}^{\infty}b_j x^j, \quad x\in S \quad \text{and} \quad S\subseteq(-\rho,\rho),$$

where b_j ($j=0,1,\ldots$) are coefficients that can depend on the model parameters. For example, for random variables with support $S=(0,\infty)$ (as the gamma distribution, for example), we have that $\rho=\infty$. It is worth stressing that the result derived in this section cannot be applied without the above assumptions. A simple example where the result derived in this section cannot be applied is the Cauchy distribution. The Cauchy CDF is

$$F(x)=\frac{1}{\pi}\arctan\left(\frac{x-\mu}{\theta}\right)+\frac{1}{2}, \quad x\in(-\infty,\infty),$$

where $-\infty<\mu<\infty$ and $\theta>0$. The above CDF does not admit a convergent power series expansion, and so our result cannot be applied.

Now, consider an equation for a power series raised to a positive integer s (see, for example Gradshteyn and Ryzhik, 2007)

$$\left(\sum_{j=0}^{\infty}b_j x^j\right)^s=\sum_{j=0}^{\infty}b_j^{(s)}x^j, \quad x\in S,$$

where

$$b_j^{(s)} = \sum_{k=0}^{j} b_k^{(s-1)} b_{j-k}, \quad b_j^{(1)} := b_j, \quad b_j^{(0)} = \begin{cases} 1, & j = 0, \\ 0, & j > 0. \end{cases} \tag{6}$$

We can express the PDF of $X_{r:n}$ in (1) as

$$f_{r:n}(x) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} \sum_{j=0}^{\infty} b_j^{(m+r-1)} x^j f(x), \quad x \in S.$$

The moments of order statistics become

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} I(k, m-r+1),$$

where

$$I(k,s) = \sum_{j=0}^{\infty} b_j^{(s)} \int_S x^{k+j} f(x) dx.$$

Note that

$$\int_S x^{k+j} f(x) dx = \mathbb{E}(X^{k+j}),$$

and so it follows that

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} \sum_{j=0}^{\infty} b_j^{(m+r-1)} \mathbb{E}(X^{k+j}). \tag{7}$$

For example, from expression (7), the moments of the maximum value of the sample $(X_{n:n})$ take the form

$$\mathbb{E}(X_{n:n}^k) = n \sum_{j=0}^{\infty} b_j^{(n-1)} \mathbb{E}(X^{j+k}), \quad k = 1, 2, \ldots.$$

Also, for $n = 1$, the above expression reduces exactly to the moments of the random variable X, which is given by $\mathbb{E}(X^k)$, for $k = 1, 2, \ldots$.

4 Examples

4.1 Gamma distribution

Here, we assume that $X_1, X_2, \ldots, X_n$ are n identically distributed gamma random variables, and $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$ are the corresponding order statistics. The lower incomplete gamma function can be evaluated using the convergent power series expansion

$$\gamma(\alpha, x) = \exp(-x) x^{\alpha} \Gamma(\alpha) \sum_{j=0}^{\infty} \frac{x^j}{\Gamma(\alpha + j + 1)}, \quad x > 0,$$

and, hence, the gamma CDF can be expressed as

$$F(x) = \frac{\gamma(\alpha, x)}{\Gamma(\alpha)} = \exp(-x) x^{\alpha} \sum_{j=0}^{\infty} \frac{x^j}{\Gamma(\alpha + j + 1)}, \quad x > 0. \tag{8}$$

Let $b_j = [\Gamma(\alpha + j + 1)]^{-1}$, for $j = 0, 1, \ldots$. We have that

$$F^{m+r-1}(x) = \exp[-(m+r-1)x]x^{(m+r-1)\alpha}\left(\sum_{j=0}^{\infty} b_j x^j\right)^{m+r-1}.$$

Note that $b_0 = [\Gamma(\alpha+1)]^{-1} \neq 0$, and so (8) is convergent for all $x > 0$ and convergence radius $\rho = \infty$. Hence, we can express the PDF of $X_{r:n}$ in (1) as

$$f_{r:n}(x) = \frac{n!}{(r-1)!(n-r)!}\sum_{m=0}^{n-r}(-1)^m\binom{n-r}{m}x^{\alpha(m+r-1)}\mathrm{e}^{-(m+r-1)x}\sum_{j=0}^{\infty} b_j^{(m+r-1)}x^j f(x),$$

where $x > 0$. The moments of gamma order statistics are

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!}\sum_{m=0}^{n-r}(-1)^m\binom{n-r}{m}\sum_{j=0}^{\infty}\frac{b_j^{(m+r-1)}}{\Gamma(\alpha)}\int_0^{\infty} x^{k+\alpha(m+r)+j-1}\mathrm{e}^{-(m+r)x}dx.$$

After some algebra, we arrive at the following closed-form expression:

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!}\sum_{m=0}^{n-r}\frac{(-1)^m}{\Gamma(\alpha)}\binom{n-r}{m}\sum_{j=0}^{\infty} b_j^{(m+r-1)}\frac{\Gamma(\alpha(m+r)+j+k)}{(m+r)^{\alpha(m+r)+j+k}}. \tag{9}$$

For example, the moments of the maximum value of the sample take the form

$$\mathbb{E}(X_{n:n}^k) = \frac{1}{\Gamma(\alpha)}\sum_{j=0}^{\infty} b_j^{(n-1)}\frac{\Gamma(\alpha n + j + k)}{n^{n\alpha+k+j-1}}, \quad k = 1, 2, \ldots.$$

Also, for $n = 1$, the above expression reduces exactly to the moments of a single gamma random variable, which is given by $\Gamma(\alpha + k)/\Gamma(\alpha)$, for $k = 1, 2, \ldots$.

4.2 Beta distribution

Here, we assume that $X_1, X_2, \ldots, X_n$ are n identically distributed beta random variables, and $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$ are the corresponding order statistics. The beta CDF can be expressed as

$$F(x) = \frac{x^a}{B(a,b)}\sum_{j=0}^{\infty}\frac{(1-b)_j}{(a+j)\,j!}x^j, \quad 0 < x < 1, \tag{10}$$

where $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du$ is the beta function, and $a > 0$ and $b > 0$. Define $(\alpha)_n = (\alpha)(\alpha+1)\cdots(\alpha+n-1)$ for $n \in \mathbb{N}$, $\alpha \in \mathbb{R}$, and $(\alpha)_0 = 1$. Let $b_j = (1-b)_j/[(a+j)j!]$, for $j = 0, 1, \ldots$. We have that

$$F^{m+r-1}(x) = \frac{x^{a(m+r-1)}}{B(a,b)^{m+r-1}}\left(\sum_{j=0}^{\infty} b_j x^j\right)^{m+r-1}.$$

Note that $b_0 = a^{-1} \neq 0$, and so (10) is convergent for all $0 < x < 1$ and convergence radius $\rho = 1$. Hence, we can express the PDF of $X_{r:n}$ in (1) as

$$f_{r:n}(x) = \frac{n!}{(r-1)!(n-r)!}\sum_{m=0}^{n-r}(-1)^m\binom{n-r}{m}\frac{x^{\alpha(m+r-1)}}{B(a,b)^{m+r-1}}\sum_{j=0}^{\infty} b_j^{(m+r-1)}x^j f(x),$$

where $0 < x < 1$.

The moments of beta order statistics are

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} \sum_{j=0}^{\infty} \frac{b_j^{(m+r-1)}}{B(a,b)^{m+r}} \int_0^1 x^{k+a(m+r)+j-1}(1-x)^{b-1} dx.$$

After some algebra, we arrive at the following closed-form expression:

$$\mathbb{E}(X_{r:n}^k) = \frac{n!}{(r-1)!(n-r)!} \sum_{m=0}^{n-r} (-1)^m \binom{n-r}{m} \sum_{j=0}^{\infty} b_j^{(m+r-1)} \frac{B(k+a(m+r)+j,b)}{B(a,b)^{m+r}}. \tag{11}$$

For example, the moments of the maximum value of the sample take the form

$$\mathbb{E}(X_{n:n}^k) = n \sum_{j=0}^{\infty} b_j^{(n-1)} \frac{B(k+an+j,b)}{B(a,b)^n}, \quad k = 1, 2, \ldots.$$

Also, for $n = 1$, the above expression reduces exactly to the moments of a single beta random variable, which is given by $B(a+k,b)/B(a,b)$, for $k = 1, 2, \ldots$.

5 Numerical results

In this section, we provide numerical values for some moments of gamma order statistics. The scripts were written in the R program (R Core Team, 2020). We use the proposed closed-form expansion (9) to obtain numerical approximations for $\mathbb{E}(X_{r:n}^k)$. Table 1 lists the values of $b_j = [\Gamma(\alpha+j+1)]^{-1}$ and $b_j^{(s)}$ (obtained using the recurrence relation (6)) for $j = 0, 1, \ldots, 6$ and $s = 0, 1, \ldots, 6$ of a gamma distribution with $\alpha = 1$. Table 2 lists numerical values of $\mathbb{E}(X_{r:6}^k)$ for $k = 1$ and 2 obtained by computational implementation of expansion (9) for a gamma distribution with $\alpha = 6.401$, which is the same value used for numerical calculations in Nadarajah (2008). Table 3 presents numerical values of $\mathbb{E}(X_{r:10})$ for $r = 1, 2, \ldots, 10$, sample size $n = 10$ and $\alpha = 6.401$. In Tables 2 and 3, we also compare the numerical moments obtained from (9) with those calculated from R code by using direct numerical integration (last columns in Tables 2 and 3). Note that the numerical values are near, mainly when $j_{max} = 100$.

Table 1: $b_j^{(s)}$ coefficients for a gamma distribution with $\alpha = 1$.

j	b_j	$b_j^{(0)}$	$b_j^{(1)}$	$b_j^{(2)}$	$b_j^{(3)}$	$b_j^{(4)}$	$b_j^{(5)}$	$b_j^{(6)}$
0	1.000000	1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1	0.500000	0	0.500000	1.000000	1.500000	2.000000	2.500000	3.000000
2	0.166667	0	0.166667	0.583333	1.250000	2.166667	3.333333	4.750000
3	0.041667	0	0.041667	0.250000	0.750000	1.666667	3.125000	5.250000
4	0.008333	0	0.008333	0.086111	0.358333	1.012500	2.298611	4.529167
5	0.001389	0	0.001389	0.025000	0.143750	0.513889	1.406250	3.237500
6	0.000198	0	0.000198	0.006300	0.050017	0.225562	0.741733	1.989616

Table 2: $\mathbb{E}(X^k_{r:6})$ obtained by the proposed expansion with j_{max} terms for a gamma distribution with $\alpha = 6.401$. The last column shows the corresponding values by numerical integration.

k	r	$j_{max} = 60$	$j_{max} = 100$	$\mathbb{E}(X^k_{r:6})$
1	1	3.6038	3.5882	3.5878
1	2	4.5781	4.7153	4.7173
1	3	6.0764	5.6749	5.6708
1	4	6.1215	6.6589	6.6632
1	5	8.2274	7.8864	7.8841
1	6	9.8048	9.8884	9.8889
2	1	14.1859	14.0150	14.0062
2	2	21.3902	23.3735	23.4198
2	3	40.0213	33.5971	33.4992
2	4	36.8939	45.9763	46.0795
2	5	70.5991	64.6280	64.5737
2	6	101.2353	102.7357	102.7471

Table 3: $\mathbb{E}(X_{r:10})$ obtained by the proposed expansion with j_{max} terms for a gamma distribution with $\alpha = 6.401$. The last column shows the corresponding values by numerical integration.

r	$j_{max} = 60$	$j_{max} = 100$	$\mathbb{E}(X_{r:10})$
1	3.0374	3.1384	3.1301
2	4.6470	3.9578	3.9992
3	4.2703	4.6399	4.6560
4	2.7310	5.6396	5.2467
5	7.7786	5.1628	5.8255
6	11.7461	6.2690	6.4285
7	1.7441	8.6948	7.0954
8	3.5427	5.9839	7.8888
9	16.1067	9.9258	8.9495
10	8.4161	10.6081	10.8002

Table 4: $\mathbb{E}(X^k_{r:6})$ for a beta distribution with parameters $a = 2$, $b = 1.2$, obtained by the correspondent expansion with j_{max} terms. The last column shows the values by numerical integration.

k	r	$j_{max} = 60$	$j_{max} = 100$	$\mathbb{E}(X^k_{r:6})$
1	1	0.306093	0.306093	0.306093
1	2	0.465444	0.465444	0.465444
1	3	0.589541	0.589551	0.589552
1	4	0.697362	0.697439	0.697461
1	5	0.796690	0.796966	0.797120
1	6	0.894870	0.894507	0.894329
2	1	0.116362	0.116362	0.116362
2	2	0.238599	0.238599	0.238599
2	3	0.366921	0.366930	0.366931
2	4	0.502333	0.502409	0.502430
2	5	0.646928	0.647201	0.647354
2	6	0.807429	0.807071	0.806894

In a similar way, we use the proposed expansion for the moments of beta order statistics to obtain some numerical values using the coefficients b_j for $j = 0, 1, \ldots, j_{max}$ according to (11). Table 4 lists the values of $\mathbb{E}(X^k_{r:n})$ for some values of k, n and r for a beta distribution with parameters $a = 2$ and $b = 1.2$. We compare the numerical results from (11) with those calculated by using direct numerical integration. In this particular example, we can verify a faster convergence rate for the expression (11).

Acknowledgments

The authors would like to thank the Editor and an anonymous reviewer for their insightful comments and suggestions. Fredy Castellares gratefully acknowledges the financial support from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). Artur Lemonte gratefully acknowledges the financial support of the Conselho Nacional de Desenvolvimento Científico e Tecnológivo (CNPq, grant 304776/2019-0).

References

David, Herbert A and Haikady N Nagaraja (2003). *Order statistics* (3rd Edition ed.). John Wiley & Sons.

Exton, Horwood (1978). *Handbook of hypergeometric integrals: theory, applications, tables, computer programs*. Halsted Press, New York.

Gradshteyn, Izrail Solomonovich and Iosif Moiseevich Ryzhik (2007). *Table of integrals, series, and products*. Academic press, New York.

Nadarajah, Saralees (2008). Explicit expressions for moments of order statistics. *Statistics & Probability Letters 78*(2), 196–205.

R Core Team (2020). R: A language and environment for statistical computing.

4 Acknowledgement to Reviewers

The Editors of Spanish Journal of Statistics gratefully acknowledge the assistance of the following people, who reviewed manuscripts

Enrique Calderín Ojeda, University of Melbourne, Australia
Diego I. Gallardo, University of Atacama, Chile.
Héctor W. Gómez, University of Antofagasta, Chile
Jorge Navarro, Universidad de Murcia, Spain