National Accounting Review

NAR, 4(2): 167–190.
DOI: 10.3934/NAR.2022010
Received: 17 May 2022
Revised: 07 June 2022
Accepted: 21 June 2022
Published: 24 June 2022

http://www.aimspress.com/journal/NAR

Research article

Forecasting Economic Indicators with Robust Factor Models

Fausto Corradin, Monica Billio* and Roberto Casarin

Department of Economics, Ca' Foscari University of Venice, Italy

* **Correspondence:** Email: billio@unive.it.

Abstract: Outliers can cause significant errors in forecasting, and it is essential to reduce their impact without losing the information they store. Information loss naturally arises if observations are dropped from the dataset. Thus, two alternative procedures are considered here: the Fast Minimum Covariance Determinant and the Iteratively Reweighted Least Squares. The procedures are used to estimate factor models robust to outliers, and a comparison of the forecast abilities of the robust approaches is carried out on a large dataset widely used in economics. The dataset includes observations relative to the 2009 crisis and the COVID-19 pandemic, some of which can be considered outliers. The comparison is carried out at different sampling frequencies and horizons, in-sample and out-of-sample, on relevant variables such as GDP, Unemployment Rate, and Prices for both the US and the EU.

Keywords: factors models; forecasting; outliers; robust estimation

JEL Codes: C13, C53, C55

1. Introduction

The increased availability of a large amount of data allows researchers to model and forecast more accurately in many fields (e.g., see Choi and Varian, 2012; Varian, 2014; Varian and Scott, 2014; Einav and Levin, 2014). However, the main issues when dealing with high-dimensional models for large datasets are over-parametrization, over-fitting, and high out-of-sample forecasting errors (Granger, 1998). Various solutions have been proposed, such as regularization (Zou and Hastie, 2005), stochastic search variable selection (George et al., 2008), graphical models (Ahelgebey et al., 2016a, 2016b), and random projections (Koop et al., 2017; Casarin and Veggente, 2021). This paper considers factor models (Stock and Watson, 2002, 2004, 2005, 2012, 2014; Banbura et al., 2010; Casarin et al, 2020;

Billio et al., 2022). Relevant information is summarized through a limited number of factors, describing the overall economic conditions and providing accurate forecasts of the variables of interest.

It has been proved, that factor model estimates can be heavily affected by outliers: data points that differ significantly from other observations in the sample. An outlier may be due to variability in the measurement or significant experimental errors; the latter are sometimes excluded from the data set. After the 2009 crisis and the COVID-19 pandemic event, the treatment of outliers attracted the attention of both researchers and the institutes of official statistics, which provided some guidelines on monitoring the effects of outliers when using their data (e.g., see Eurostat, 2020). In this paper we follow Artis et al. (2005), Croux et al. (2003), Bai et al. (2022), Fan et al. (2021) and apply robust estimation methods to factor models to limit the effects of the outliers. We contribute to the robust factor literature by comparing alternative robust factor models in terms of forecasting performances on a set of variables which are central to the economic analysis. Our database includes the 2009 crisis and the beginning of COVID-19 pandemic in March 2020 and consider as a last data January 2021; the pandemic is potentially the most important source of outliers, and its effects on the economic systems have been extensively investigated in some recent studies (Fabeil et al., 2020; Fernandes, 2020; McKibbin and Vines, 2020; McKibbin and Roshen, 2021; Liu, 2021). We shall notice that the amount of sample information is not large enough to estimate forecasting models with structural breaks since adopting them implies that the current model is estimated only using data observed since the most recent break. Similarly, it is not possible to test for a break and compare the two models for the period before and after the pandemic since the spread of contagion and its effects did not yet come to an end. This paper provides an alternative solution and shows that samples from the pandemic period have some information content which can still be used to estimate models without breaks provided a proper inference technique, such as robust inference for outliers, is applied.

The structure of the paper is as follows. Section 2 presents some background on robust inference for outliers. Section 3 introduces standard factors model and the two methodologies used to treat the outliers. Section 4 provides a data description and the empirical results obtained with robust inference methods for factor models. Section 5 concludes the chapter.

2. Background on robust estimation

The true nature of outliers can be very elusive and dealing with data affected by outliers poses some challenges. There is no unanimous definition for what an outlier is. Outliers could be atypical samples that have an unusually large influence on the estimated model parameters. Outliers could also be perfectly valid samples from the same distribution as the rest of the data that happen to be small-probability instances. Alternatively, outliers could be samples drawn from a different model, and therefore they will likely not be consistent with the model derived from the rest of the data. There is no way to tell which is the case for a particular "outlying" sample point, nevertheless some techniques can be applied to detect outliers. A standard procedure makes use of the linear projection of the dependent variable into the linear space of covariates, the hat matrix of the data. The diagonal of the hat matrix is used to detect outlying observations that may have an impact on the inference. Usually, outliers are excluded from the dataset when estimating the model (data-trimming). See, for example Davidson and McKinnon (2004). In this paper, we compare trimming with two alternative approaches.

The first approach is based on Mahalanobis distances and can applied for detection and robust estimation. We consider robust estimators of multivariate location and scatter computed from the

explanatory variables. Many methods for estimating multivariate location and scatter break down in the presence of $T/(n+1)$ outliers, where T is the number of observations and n is the number of variables, as was pointed out by Donoho (1982). For the breakdown value of the multivariate F-estimators of Maronna (1976), see Hampel et al. (1986). In the meantime, several positive breakdown estimators of multivariate location and scatter have been proposed. The Minimum Covariance Determinant (MCD), a highly robust estimator of multivariate location and scatter (Rousseeuw, 1984) which uses only the observations whose covariance matrix has the lowest determinant, was proposed by Rousseeuw and Leroy (1987). Consistency and asymptotic normality of the MCD estimator has been shown by Butler et al. (1993) and Cator and Lopuhaa (2010), whereas has been demonstrated that MVE (Minimum Volume Ellipsoid) has a lower convergence rate (Davies, 1992). The MCD has a bounded influence function (Croux and Haesbroeck, 1999) and it has the highest possible breakdown value (i.e., 50%) when the number of observations used is $\lfloor (T+n+1)/2 \rfloor$ (Lopuha and Rousseeuw, 1991). In addition to being highly resistant to outliers, the MCD is affine equivariant, i.e., the estimates behave properly under affine transformations of the data. Although the MCD was already introduced in 1984, its practical use only became feasible since the introduction of the computationally efficient Fast MCD (FMCD) algorithm of Rousseeuw and Van Driessen (1999), and some extensions have been determined (Hubert et al., 2017); in this paper we follow FMCD technique. MCD have been successfully applied in many fields such as finance and econometrics (Gambacciani & Paolella, 2017; Orhan et al., 2001), quality control (Jensen et al., 2007), geophysics (Neykov, et al., 2007), geochemistry (Filzmoser et al., 2005), image analysis (Vogler et al., 2007). MCD has been used for robust factor model estimation by Croux et al. (2003) and Filzmoser et al. (2003).

The second approach considered, is the Iteratively Reweighted Least Squares (IRLS) proposed in (De la Torre and Black, 2004), which relies on the residuals of the linear projection of the dependent variable on a space generated by a set of factors. The outliers are detected as those that have a large residual with respect to the identified subspace. A new subspace is estimated with the outliers downweighted, and this process is then repeated until the estimated model stabilizes. With this algorithm for every multivariate sample a weight is determined iteratively, reducing the weights related to the outliers until the procedure converge. This technique has been used for outliers' reduction, (Bergstrom and Edlund, 2014), outliers afflicted observations (Kargoll et al., 2018) and in forecasting (Mbamalu et al., 1993). Other applications are statistical estimation (Green, 1984), matrix rank minimization (Mohan and Fazel, 2012), and sparse matrix (Daubechies et al., 2009).

3. Factor models

In the following, we introduce Factor Models (FM), data trimming and three approaches to outlier handling: i) standard FM (FM Std) where all data are included without any transformation; ii) Fast Minimum Covariance Determinant methodology combined with FM (FM FMCD) and iii) Iterated Reweighted Least Squares combined with Factors Model (FM IRLS).

In the empirical analysis, a Vector Autoregressive (VAR) model is used for predicting the factors, and according to Lütkepohl (2005), series without unit roots should be used when forecasting with VARs. To meet this requirement for the factors, we perform a unit root ADF test on all variables included in the factor analysis. If necessary, variables have been differentiated to obtain a stationary time series; after this step, we normalize the series and extract the factors. Thus, in the following we assume our $T \times n$ data matrix X is covariance stationary with null mean and unitary standard deviation.

3.1. A standard factor model

3.1.1. Subheading

In this paper we use factor models, (see, e.g., Stock and Watson, 2002, 2004, 2005, 2012, 2014; Banbura et al., 2010; Banbura et al., 2014; Artis et al., 2005), with reduced number of factors (Bai and Ng, 2002). See Diebold (2003) and Stock and Watson (2009) for review of factor models.

Let X_t, *t>0*, be a random process with $X_t = (x_{1t}, \dots, x_{nt})'$ a $(n \times 1)$ random vector. The time index t represents months or quarters, and we assume the process is covariance stationary with null mean and a standard deviation equal to one. Latent factors extraction relies on the following decomposition:

$$E[X_t X_t']a_i = \Gamma_X a_i = \lambda_i a_i, \tag{1}$$

where a_i and λ_i, $i = 1, \dots, n$, are the n-dimensional eigenvectors and the eigenvalues in decreasing order, respectively. Let A be an $(n \times n)$ orthonormal matrix with the normalized eigenvectors in the columns, also called factor loading matrix, then:

$$\Gamma_X A = A\Lambda, \tag{2}$$

where Λ is a diagonal matrix with elements λ_i, $i = 1, \dots, n$, on the main diagonal. The vector of n factors $F_{n,t} = (f_{1,t}, \dots, f_{n,t})'$ is given by the linear transformation:

$$F_{n,t} = A' X_t, \tag{3}$$

And $f_{k,t}$, $t > 0$ is the k-th factor. Let us denote with Γ_n the expectation of the external product of the factor, $E\left[F_{n,t}F_{n,t}'\right]$; then one obtains the following relationship between Γ_n and the eigenvector matrix Λ:

$$\Gamma_n = E[A' X_t X_t' A] = A' E[X_t X_t'] A = A' \Gamma_X A = \Lambda. \tag{4}$$

Let $F_{k,t} = (f_{1,t}, \dots, f_{k,t})'$ be the collection of the first *k* factors at time *t*, with $k < n$, then:

$$F_{k,t} = A_k' X_t, \tag{5}$$

where A_k is the matrix containing the first *k* columns of A. Since the columns of A are orthogonal, then $A_k' A_k = I_k$. The first *k* factors capture the following proportion of the total variance:

$$V_k = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i \,. \tag{6}$$

The collection of factors $F_{k,t}$ is customarily called standard FM (FM Std).

3.2. A robust factor model: the fast minimum covariance determinant estimator

In n-variate data, *n*>2, it is difficult to detect outliers because one can no longer rely on visual inspection, nevertheless a set of summary statistics can be used. One of the statistics used in the literature is the Mahalanobis distance:

$$D\left(x_t, \hat{\mu}, \hat{\Sigma}\right) = D_t = \sqrt{(x_t - \hat{\mu})' \hat{\Sigma}^{-1} (x_t - \hat{\mu})}, \tag{7}$$

where x_t is the t-th row of the data matrix X, $\hat{\mu}$ is the estimator of the location, and $\hat{\Sigma}$ is the covariance matrix estimator. Using this distance, one obtains the classical tolerance ellipse defined as the set of n-dimensional points $x_t, t = 1, \dots, T$. Detecting outliers by means of the Mahalanobis distance no longer suffices for multiple outliers because of the masking effect, by which multiple outliers do not necessarily have large Mahalanobis distances (Hubert ET AL., 2017). We consider a robust estimator of multivariate location and scatter base on the notion of Minimum Covariance Determinant (MCD) (Rousseeuw, 1984; Rousseeuw and Leroy, 1987; Hubert et al., 2017). In the MCD, only the r observations, $\lfloor (T + n + 1)/2 \rfloor \leq r \leq T$, whose classical covariance matrix has the lowest determinant are considered in the computation of the Mahalanobis distance:

$$RD\left(x_t, \bar{\mu}_{MCD}, \hat{\Sigma}_{MCD}\right) = \sqrt{(x_t - \hat{\mu}_{MCD})' \hat{\Sigma}_{MCD}^{-1} (x_t - \hat{\mu}_{MCD})} \tag{8}$$

where $\hat{\mu}_{MCD}$ and $\hat{\Sigma}_{MCD}$ are the MCD estimator of the mean and the covariance matrix respectively defined as follow:

$$\hat{\mu}_{MCD} = \frac{\sum_{t=1}^{T} W(d_t^2) x_t}{\sum_{t=1}^{T} W(d_t^2)}; \hat{\Sigma}_{MCD} = c_1 \frac{1}{T} \sum_{t=1}^{T} W(d_t^2)(x_t - \hat{\mu}_{MCD})\,(x_t - \hat{\mu}_{MCD})' \tag{9}$$

Where $W(d_t^2)$ is an appropriate weight function and c_1 is a consistency factor (e.g., see Lopulhaa and Rousseuw, 1991). Note that the MCD estimator can only be computed when $r>n$, otherwise the covariance matrix of any r-subset has determinant 0, so we need at least $T>2n$. To avoid excessive noise, it is recommended that $T>5n$, so that we have at least five observations per dimension.

The MCD estimator is computationally expensive as it requires the evaluation of $\binom{T}{r}$ subsets of size r and for this reason we use the Fast Minimum Covariance Determinant estimator (FMCD) of Rousseeuw and Van Driessen (1999). A major component of the FMCD algorithm is the concentration step, C-step, which works as follows. Given the initial estimates $\hat{\mu}_{old}$ and $\hat{\Sigma}_{old}$:

- Compute the distances $d_{old}(t) = D\left(x_t, \hat{\mu}_{old}, \hat{\Sigma}_{old}\right),\ t = 1, \dots, T.$
- Sort these distances and yield a permutation τ such that:

$d_{old}\left(\tau(1)\right) \leq d_{old}\left(\tau(2)\right) \leq \cdots \leq d_{old}\left(\tau(n)\right)$ and set $H = \{\tau(1), \tau(2), \dots, \tau(r)\}$.

- Compute location and scale estimators:

$$\hat{\mu}_{new} = \frac{1}{r} \sum_{t \in H} x_t, \quad \hat{\Sigma}_{new} = \frac{1}{r-1} \sum_{t \in H} (x_t - \hat{\mu}_{new})(x_t - \hat{\mu}_{new})' \tag{10}$$

In Theorem 1 of Rousseuw and Van Driessen (1999) it is proved that $det\left(\hat{\Sigma}_{new}\right) \leq det\left(\hat{\Sigma}_{old}\right)$, with equality only if $\hat{\Sigma}_{new} = \hat{\Sigma}_{old}$. Thus, if we iterate the C-step, the sequence of determinants obtained in this way converges in a finite number of iterations. The FMCD algorithm supplies a sequence of weights, one or zero (zero for the outliers), that has length T, and we repeat this sequence for n column to obtain the matrix H_{MCD} with dimension $T \times n$. We multiply the data matrix X by H_{MCD}:

$$H_{MCD} \odot X = X_{MCD} \tag{11}$$

where $\odot$ denotes the Hadamard's product. We use X_{MCD} to extract the factors (Croux et al., 2003; Pison et al., 2003) $F_{k,t}$, as described in Section 3.1 and obtain the model FM FMCD. In this paper we set $r = 0.95T$.

3.3. A robust factor model: the iterated reweighted least squares estimator

The Maximum-Likelihood-Type Estimator (M-Estimator) is another popular robust method for estimating the location and scale of a set of points, and its application leads to the Iterated Reweighted Least Squares (IRLS) (Bergstrom et al., 2014; Daubechies et al., 2009). Define the residual as follows:

$$\varepsilon_t = \left\| x_t - A_k F_{k,t} \right\|_2 \tag{12}$$

where A_k and $F_{k,t}$ have been defined in Section 3.1, and $\|\cdot\|_2$ is the Euclidean norm for vectors. IRLS assumes continuous weights as a function of the residual:

$$w_t = \frac{\rho(\varepsilon_t)}{\varepsilon_t^2} \tag{13}$$

For some given robust loss function $\rho(\cdot)$ from the set of the real to the positive reals. The objective function then becomes:

$$\sum_{t=1}^{T} \rho(\varepsilon_t) \tag{14}$$

Many loss functions have been proposed in the statistics literature (Huber 1981; Barnett and Lewis 1984). When $\rho(\varepsilon_t) = \varepsilon_t^2$, all weights are equal to 1, and we obtain the standard least-squares solution, which is not robust. Other robust loss functions are described in Vidal et al. (2016), in this work we use a Geman-McClure loss (Geman and McClure, 1987):

$$\rho(\varepsilon_t) = \frac{\varepsilon_t^2}{\varepsilon_0^2 + \varepsilon_t^2} \tag{15}$$

where ε_0^2 is a parameter that we consider equal to the square root of mean of ε_t^2. Following De la Torre and Blank (2004), we use a Geman-McClure loss scaled by ε_0^2 which yields the following procedure. Given an initial parameter ε_0^2 and factor loadings and factors, A_k and $F_{k,t}$, respectively, obtained from the FM Std of Section 3.1, iterate until convergence the following steps:

1. Compute the residuals $\varepsilon_t = \left\| x_t - A_k F_{k,t} \right\|_2$.
2. Compute the weights $w_t = \frac{\varepsilon_0^2}{\varepsilon_0^2 + \varepsilon_t^2}$.
3. Estimate the covariance $\Sigma \leftarrow \frac{\sum_{t=1}^{T} w_t x_t x_t'}{\sum_{t=1}^{T} w_t}$.
4. Extract the first k largest eigenvalues of Σ and collect the corresponding eigenvectors in A_k.

The factor matrix $F_{k,t}$ obtained as above is called FM IRLS.

3.4. A forecasting model

Once factors are extracted a forecasting procedure is needed to predict the variables of interest. We assume the first k latent factors (determined with the three methodologies described above), $F_{k,t} = (f_{1,t}, \dots, f_{k,t})'$, with $k < n$, follow a VAR model. Using only k factors, the reconstruction of the variables derives from the approximated model:

$$X_{k,t} = A_k F_{k,t} \tag{16}$$

With the term $X_{k,t}$ we mean the approximation of the vector X_t obtained using the first k factors $F_{k,t}$. Considering the dynamic part related to the k factors, our model is thus as follows:

$$X_{k,t} = A_k F_{k,t} \tag{17}$$

$$F_{k,t} = c_k + \Phi_k F_{k,t-1} + \varepsilon_{k,t}, \quad \varepsilon_{k,t}\, WN(0, \Sigma_k) \tag{18}$$

where c_k has dimension *k x 1* and Φ_k has dimensions *k x k*. As shown in Billio et al. (2022), under VAR assumption for the factors, the variables of interest $X_{k,t}$ follow a VAR model with restrictions. Thus, the conditional forecasts $X_{k,t+h}$ at the horizon h, $h = 1, \dots, H$ are obtained as follows:

$$X_{k,t+h|t} = A_k \hat{F}_{k,t+h|t} \tag{19}$$

where:

$$\hat{F}_{k,t+h|t} = E\left[F_{k,t+h} | X_1, \dots, X_t\right] = c_k + \Phi_k \hat{F}_{k,t+h-1|t} \tag{20}$$

To summarize, we first estimate the latent factors and then use a VAR model on factors to forecast both the factors and the variables of interest. After then, the variables will be reconverted to their correct values reversing the procedures of normalization and integrated if they have been previously differentiated.

4. Empirical applications

4.1. Data description

We consider a dataset of macroeconomic variables related to the US and the EU economies, provided by Bloomberg. It consists of 42 monthly variables and 2 quarterly variables, sampled from December 2001 to January 2021, and includes some key variables for policy making: core and headline prices, labour market variables, imports, exports, industrial production, consumption, sales, leading indicators of interest rates, and the term structure. See Table 1 for a more detailed description.

Table 1. Macroeconomic variables for two major geographical regions, the US and EU, sampled either at monthly or quarterly frequency from December 2001 to January 2021.

N	C	Definition	L	MU	F
1	US	Export	Ex	m/m	M
2	US	Import	Im	m/m	M
3	US	Unemployment rate	UR	%	M

Continued on next page

N	C	Definition	L	MU	F
4	US	Employment (Agricultural sector)	EA	thousands	M
5	US	Employment (Private sector)	EP	thousands	M
6	US	Average hourly wages	Ahw	m/m	M
7	US	PCE	PCE	y/y	M
8	US	PCE core	PCEc	y/y	M
9	US	PPI	PPI	y/y	M
10	US	Industrial Production	IP	y/y	M
11	US	Industrial Orders	IO	m/m	M
12	US	Durable goods orders	Dgo	m/m	M
13	US	Durable goods orders excluding transport	Dgoet	m/m	M
14	US	Stocks	S	m/m	M
15	US	Use of production capacity	Upc	%	M
16	US	ISM manufacturing	ISMm	level	M
17	US	Start of new construction sites	Snc	m/m	M
18	US	Constructions expenditure	Cse	m/m	M
19	US	Existing homes sale	Ehs	m/m	M
20	US	New homes sale	Nhs	m/m	M
21	US	Expenditure (real)	Er	m/m	M
22	US	Income (real)	Ir	m/m	M
23	US	Retail sales	RS	m/m	M
24	US	Conference Board	CB	level	M
25	US	Michigan Consumer Sentiment Index	MCSI	level	M
26	US	RUS10 (Int.Rate Gov.Bond 10Y US)	RUS10	Yield	M
27	US	DeltaRUS72 (=RUS7-RUS2)	DRUS72	Yield	M
28	EU	Export	Ex	m/m	M
29	EU	Import	Im	m/m	M
30	EU	Unemployment rate	UR	%	M
31	EU	HCPI	HCPI	y/y	M
32	EU	CPI core	CPIc	y/y	M
33	EU	PPI	PPI	y/y	M
34	EU	Industrial Production	IP	y/y	M
35	EU	Constructions expenditure	Cse	m/m	M
36	EU	PMI manufacturing index	PMImI	level	M
37	EU	ESI	ESI	level	M
38	EU	Leading indicator	LeIn	level	M
39	EU	Retail Sales	RS	y/y	M
40	EU	REMU10 (Int.Rate Gov.Bond 10Y EU)	REMU10	Yield	M
41	EU	DeltaREMU72 (=REMU7-REMU2)	DREMU72	Yield	M
42	EU/US	CEUUS	CEUUS	Ratio €/$	M
43	US	Gross Domestic Product	GDPUS	q/q	Q
44	EU	Gross Domestic Product	GDPEU	q/q	Q

Note: In the columns, the series: number (N), country (C), description (Definition), label (L), measure unit (MU), and frequency (F) that is quarterly (Q) or monthly (M).

In our dataset, the 2009 financial crisis and the COVID-19 pandemic generated outliers in many time series. For example, a graphical inspection of the US unemployment rate series reveals the dramatic impact of COVID-19 pandemic after March 2021 (Figure 1). In presence of outliers, the

researcher can choose to trim the data, that is to reduce or eliminate the outliers and to run inference on a linear model overcoming the inference issues (such as bias) that the outliers can generate. Data trimming requires outliers are detected first. To detect the presence of the outliers, a standard procedure consists in fitting the linear regression model:

$$Y = X\beta + \varepsilon \tag{21}$$

By Least Squares and recovering the hat matrix H from the fitted value of Y:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY \tag{22}$$

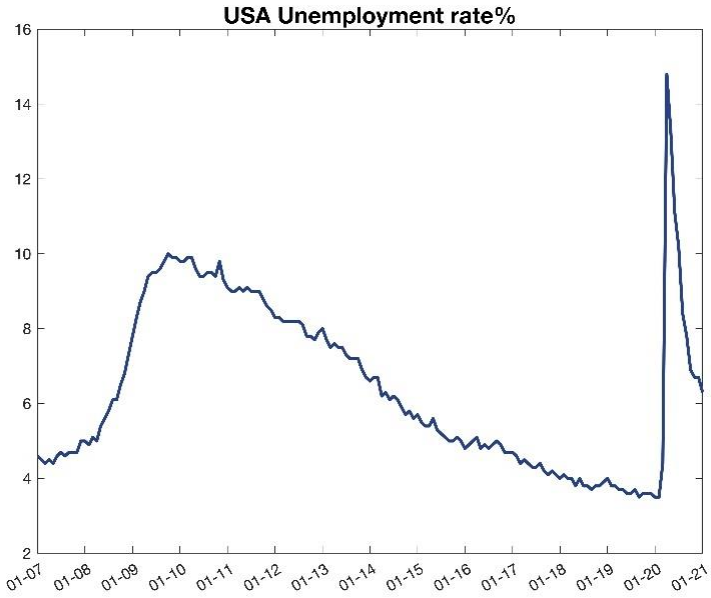

Figure 1. The US Unemployment rate (in percentage) from December 2001 to January 2021.

The hat matrix H is a symmetrical and idempotent *T x T* projection matrix; it has *n* eigenvalues equal to one and *T-n* equal to zero. The diagonal elements $h_{t,t}$ have the following property:

$$0 \leq h_{t,t} \leq 1, t = 1, \ldots, T \tag{23}$$

Points where $h_{t,t}$ have large values are called leverage points, and it can be proved that the presence of leverage points signals that there are observations that might have a decisive influence on the estimation of the regression parameters. We consider the leverage points as a proxy for the quick survey of presence of the outliers. We can see in Figure 2 the greater values of $h_{t,t}$ is detected for the 2009 crisis and COVID-19 pandemic.

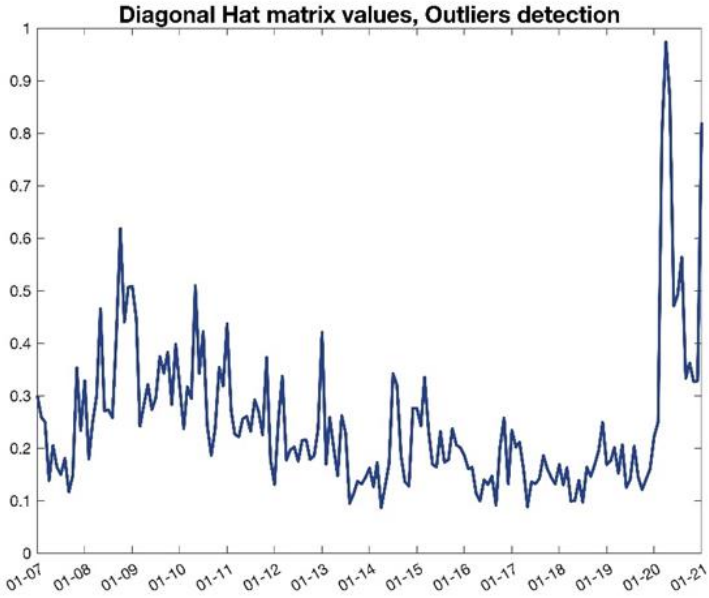

Figure 2. Outliers' detection. The Hat matrix diagonal values for the US Unemployment rate from December 2001 to January 2021.

4.2. Factor Analysis

The factors have been extracted from the monthly variables using three FM methodologies (FM Std, FM FMCD, and FM IRLS), and then they supply the forecast using a VAR model as described in Section 3.4. As regards to the quarterly variables we follow a nowcasting procedure. First, we derive the regression coefficients of the quarterly variables on the nowcasted factors and secondly, we use the coefficients and the forecasted factors to forecast the quarterly variables.

We analyze the stability of the factors and the percentage of explained variance. We follow a rolling window estimation approach and analyze the out-of-sample forecast ability of the FM with a twelve-months horizon. There are 61 overlapping windows of 170 observations each. The first window is from December 2001 to January 2016, the second shift is one month from January 2002 to February 2016, and the 61st is from December 2006 to January 2021. See Figure 3 for a graphical illustration of the procedure (see Figure 3).

In our empirical applications, the k factors are used to forecast the variables of interest, which are the Unemployment rate and the Harmonized Index of Consumer Prices (HCPI); with nowcasting procedure we produce the forecast also for GDP. Our choice is to explain a given proportion of variance $V_k < 1$ in Eq. (6) with a reduced number of factors in order to limit the dimension of forecast model (i.e., the VAR model). For example, in our application we choose to explain at least 80% of the variance, $V_k = 0.8$, with no more than 9 factors.

Figure 4 reports the values of the leverage point $h_{t,t}$ estimated from the panel series in some relevant windows (see plot labels). The value of $h_{t,t}$ increases slowly in the observation windows where the 2009 crisis is included (e.g., see plots w1, w19, w31, and w43). When the observations associated to the COVID-19 pandemic period are included in the samples, then $h_{t,t}$ reaches much larger values, about 1 (see w51, w52, w53, and w58). A double peak appeared in the window w61 due to the second wave in the CODIV-19 pandemic.

In conclusion, we consider COVID-19 as the greater cause of outliers that the researchers are facing. For this reason, we choose a percentage of $r = 0.95T$ for the values to be saved in FMCD algorithm.

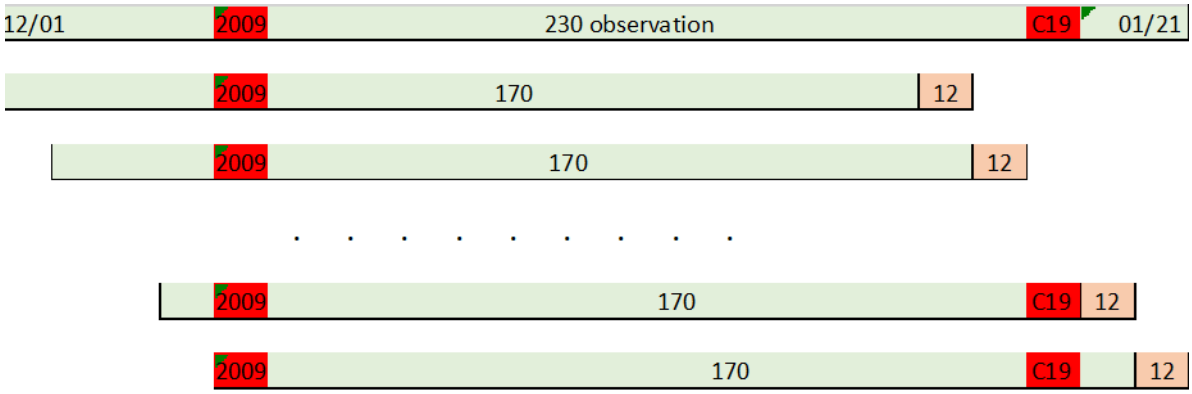

Figure 3. Rolling windows. The window size is 170 observations. We consider 61 overlapping windows. The first window is from December 2001 to January 2016 (second line), the second window is from January 2002 to February 2016 (third line), and the last is from December 2012 to January 2021 (last line). The green segments identify the data used to produce the factors and the forecast. The forecast's horizon is twelve months and is identified by the orange. Segments. The red segments indicate the 2009 crise and the COVID-19 pandemic.

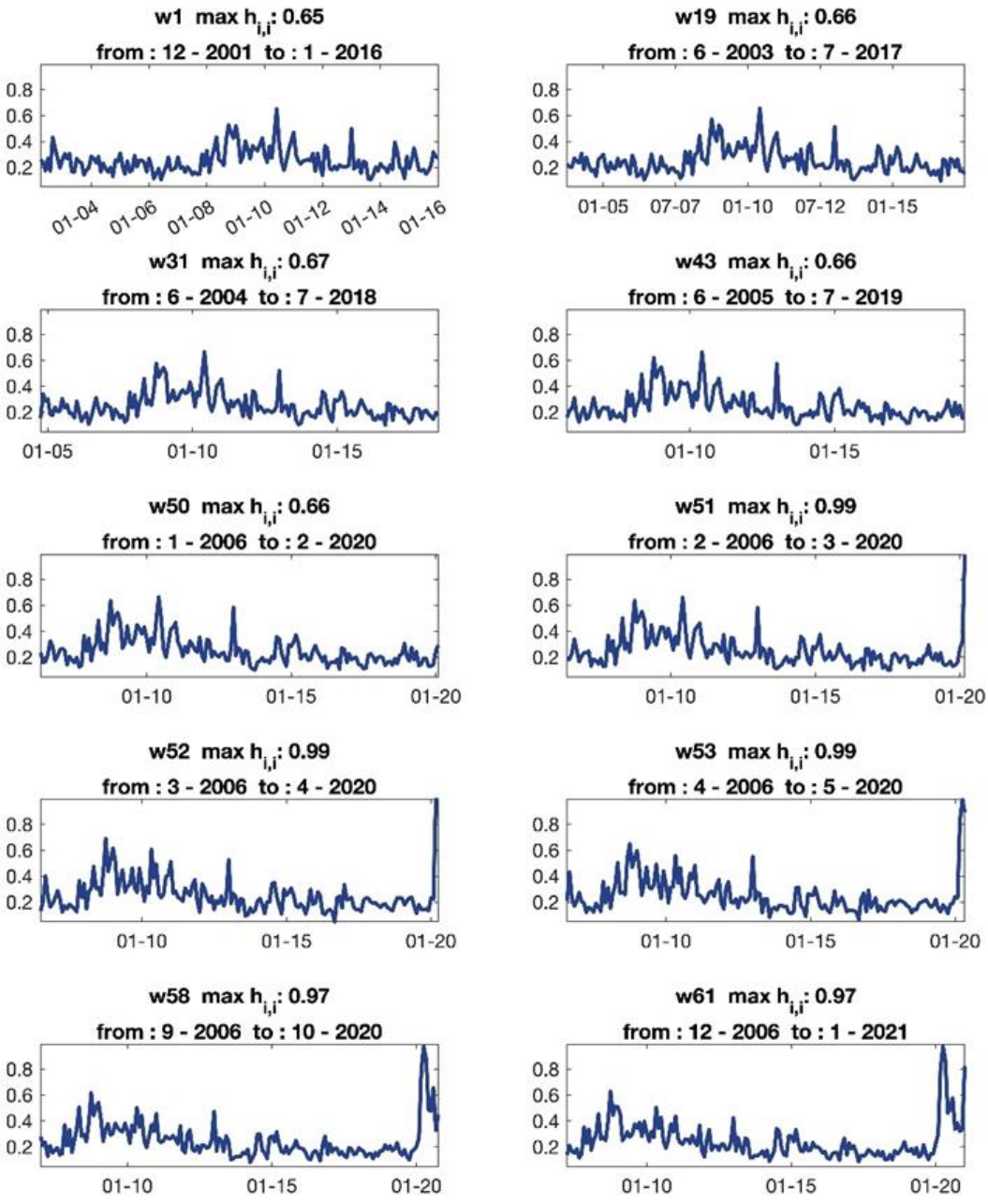

Figure 4. Leverage points $h_{t,t}$ for some relevant windows.

Figure 5 shows the eigenvalues (left column) and the contribution of the first 9 factors (right) to the variance for the three FMs. The graphs refer to the data of the last window in Figure 3. The scale of the eigenvalues differs across models since the weights used in FMCD and IRLS have different size. The decay rate of the spectrum is similar across models, and this indicates small number of factors explain a large proportion of variance. The FM FMCD and FM IRLS models intercept a smaller proportion of variance than in the FM Std case.

The green lines in Figure 6 shows the weights used in the FMCD (left) and IRLS (right). Setting $r = 0.95T$ yields weight equal to one for all windows expect for the COVID-19 pandemic windows where the weight is equal to zero. For the IRLS the weights are strictly positive for all estimation

windows and below one. The two weight sequences have different impact on the extraction of the factors (e.g., see the first factor in the same figure and the three factors in Figure 7).

The FM Std model factors exhibit at least 2 peaks corresponding to the 2009 crisis and COVID-19 pandemics windows. In the robust FM procedures, the weight sequences reduce substantially the effects of the two sources of outliers.

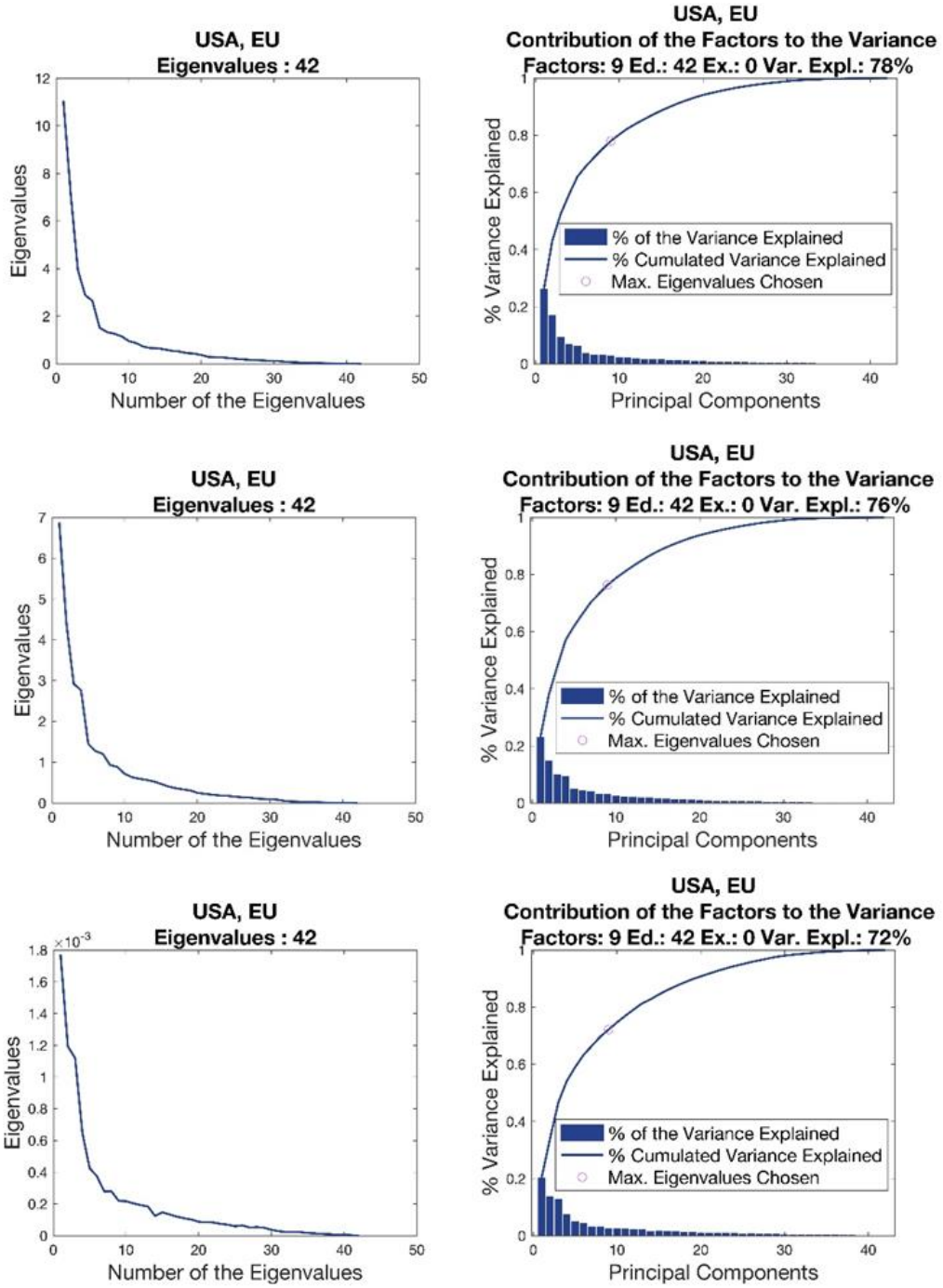

Figure 5. Eigenvalues (left) and contribution of the factors (right) for the FM Std (top), FM FMCD (middle) and FM IRLS (bottom).

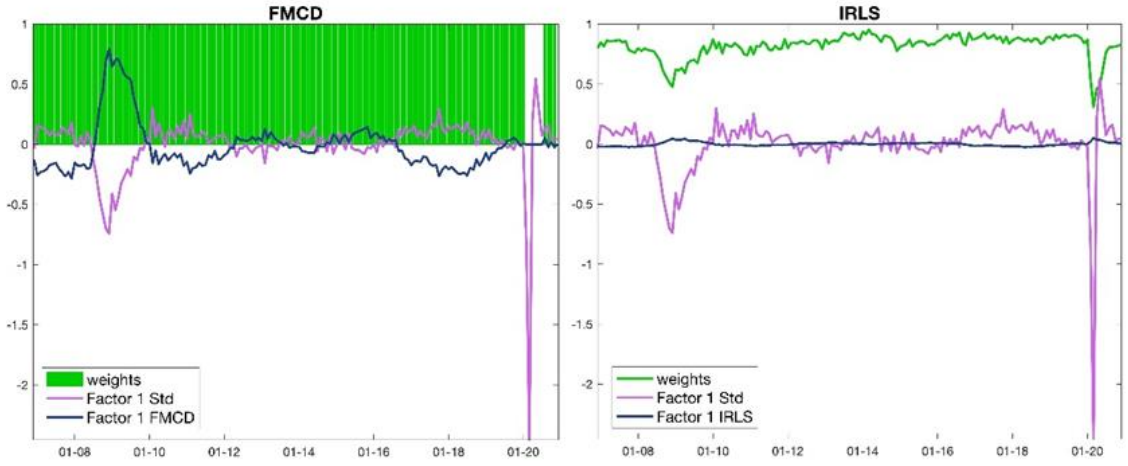

Figure 6. Factor 1 extracted with the FM Std (magenta line), the FM FMCD (left, blue line) and FM IRLS (right, blue line), and the weights used (green line). Due to the configuration of the weights, the amplitude of the peak associated with the pandemics (see top plot in Figure 7) is reduced in the FM FMCD model, whereas both peaks are largely reduced in the IRLS model.

In Table 2, we illustrate the bias issues induced by the presence of outliers, by comparing the correlation between the variables in the dataset (columns) and the first factor of the three models (rows), estimated in the last windows (w61). The first factor in the three models explains the 26%, 23%, and 20% of the variance, respectively. The set of the most correlated variables in the standard FM model differs from the one of the FM FMCD and FM IRLS models, which indicates that the bias in the estimation of the factor can be large in the FM models if the outliers are not treated properly.

Table 2. The 10 most correlated variables with the first factors for the three methodologies in the 61st (December 2006–January 2021) window.

FM Std: Correlations between Factors 1 and Variables.										
Country	USA	USA	EU	USA	EU	USA	USA	EU	USA	EU
Indicator	EP	EA	Ex	Ex	RS	Ahw	Er	Im	IP	LeIn
Measure	thousands	thousands	m/m	m/m	y/y	m/m	m/m	m/m	y/y	level
Factor 1 w61	0.835	0.835	0.798	0.785	0.726	–0,724	0.721	0.708	0.686	0.669
FM FMCD: Correlations between Factors 1 and Variables.										
Country	USA	EU	EU	EU	EU	USA	USA	USA	USA	EU
Indicator	PPI	PMImI	ESI	LeIn	IP	Stocks	PCE	ISMm	Upc	PPI
Measure	y/y	level	level	level	y/y	m/m	y/y	level	%	y/y
Factor 1 w61	–0.847	–0.762	–0.759	–0.730	–0.728	–0.708	–0.707	–0.680	–0.668	–0.600
FM IRLS: Correlations between Factors 1 and Variables.										
Country	EU	USA	EU	USA	EU	EU	USA	USA	USA	EU
Indicator	LeIn	PPI	ESI	Upc	IP	PPI	PCE	Stocks	PCEc	PMImI
Measure	level	y/y	level	%	y/y	y/y	y/y	m/m	y/y	level
Factor 1 w61	–0.877	–0.852	–0.840	–0.796	–0.794	–0.722	–0.708	–0.674	–0.666	–0.623

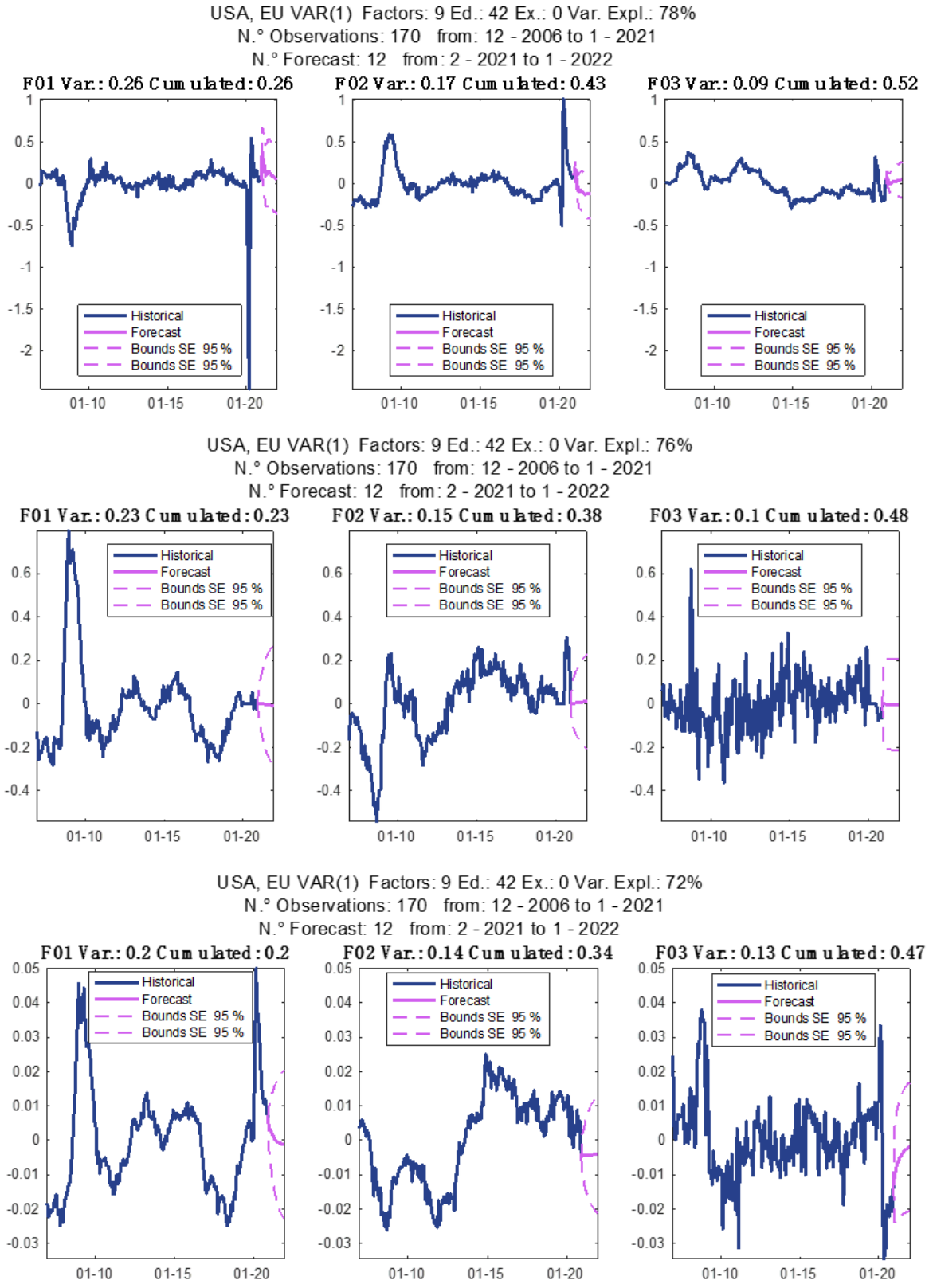

Figure 7. The first three factors (blue line), their out-of-sample forecasts (magenta solid lines) with the confidence bands (magenta dashed lines), for the FM Std (top), FM FMCD (middle) and FM IRLS (bottom).

FM FMCD and FM IRLS share 9 common variables, and the correlation levels are similar in the two models; the result indicates that the choice of the weights can have an impact on the results, but the economic interpretation of the factors is not affected too much.

4.3. Forecast comparison

We use the rolling window analysis introduced in the previous section for comparing the three models: FM Std, FM FMCD and FM IRLS. For each window, the models produce 12 forecasts out of sample (see Figure 3). We measure and compare sequentially the ability of the models to forecast the following variables: GDP, Unemployment rate, as well as PCE for both the EU and the US regions.

For every window, we determine the factors and compute the forecast at the horizon of 12 months for the monthly variables and 4 quarters for the quarterly variables. The rolling window of 170 observations is moved forward by one month, and the forecasts are computed again. We repeat this exercise 61 times until the end date of the observation window coincides with January 2021.

For every series, compute the square of the difference between the forecast and the actual values, sum the squared differences, divide them by the total number of forecast points, and take a square root to obtain the Root Mean Square Error (RMSE). Let s_t be the forecast horizon at time t for monthly data. In our application, it is equal to 12 for all t except when the end of the window is close to January 2021, when the horizon decreases. Moreover, let $TEp = 61$ be the number of forecasts for each one of the s_t months.

At time t, we have the following error for every forecast (we omit here the identification of the variable):

$$e(t) = \sum_{i=1}^{s_t} [f(t+i) - v(t+i)]^2 \tag{24}$$

where $f(t+i)$ indicates the forecast for the variable $v(t+i)$. The forecast is made at time t with forecasting horizon i. The RMSE is thus defined as follows:

$$RMSE = \sqrt{\frac{1}{\sum_{t=1}^{TEp} s_t} \sum_{j=1}^{TEp} e(j)} \tag{25}$$

As a first step, we show the RMSE value for the first three factors, that intercept more than 52.5%, 47.8%, 46.7% of the total variance for FM Std, FM FMCD, FM IRLS respectively.

The left column of Figure 8 shows the actual values of the three factors (solid blue lines, in the rows), the 12-step-ahead FM forecasts (dashed lines), and their envelope (solid red lines), which can be considered an approximation of the forecasting error bands. The forecast comparison includes the COVID-19 pandemic period, but cannot be made for the 2009 crisis one, due to the choice of the rolling window size (see Figure 3). Thus, in the following we focus on the forecast ability during the COVID-19 period.

For the first factor, the actual values belong to the envelope region for all periods except for the pandemic crisis periods, which reveal the difficulties in predicting the effects of the pandemic events. A similar behavior can be detected for the second factor. The middle and right columns in Figure 8 show the first three factors for FMCD and IRLS methodologies respectively; their behavior is

comparable only by graphical point of view, because they have different scale due to the two applied algorithms. Figure 9 shows the forecasts and the RMSEs for our variables of interest and for the three methodologies: FM Std (left column), FM FMCD (middle) and FM IRLS (right).

By using the envelope (solid red lines) as reference lines, it is possible to compare graphically the forecast performance of the models. Since the actual data belong to the area delimited by the envelope of the FM FMCD and FM IRLS models, we conclude that they usually perform better than FM Std. The lower RMSE level of the FM FMCD and FM IRLS model for both the monthly and quarterly variables allows us to confirm this result (see panel (a) in Table 3).

Table 3. Root Mean Square Error (RMSE) for the variables of interest (rows) following different forecasting models (columns).

Panel a. Cross-horizon overall RMSEs.

RMSE	FM Std	FM FMCD	FM IRLS
USA GDP	0.42	0.39	0.36
USA Unemployment rate	40.06	33.62	33.65
USA PCE	0.83	0.62	0.49
EU GDP	0.48	0.41	0.42
EU Unemployment rate	0.80	0.72	0.70
EU HCPI	0.98	0.81	0.80

Panel b. RMSEs at different horizons.

	1 month ahead			5 months ahead			7 months ahead			12 months ahead		
RMSE	FM Std	FM FMCD	FM IRLS	FM Std	FM FMCD	FM IRLS	FM Std	FM FMCD	FM IRLS	FM Std	FM FMCD	FM IRLS
USA Unempl. Rate	3.54	3.11	3.11	33.53	27.06	27.08	56.97	47.31	47.35	4.91	3.61	3.32
USA PCE	0.49	0.46	0.31	0.79	0.60	0.47	0.82	0.69	0.52	1.05	0.67	0.58
EU Unempl. rate	0.16	0.15	0.14	0.66	0.62	0.65	0.80	0.75	0.68	1.29	1.38	0.98
EU HCPI	0.51	0.49	0.45	0.91	0.79	0.74	1.00	0.87	0.82	1.25	1.06	1.05

The effect of outliers on the most impacted variables propagates to the forecast of the other variables through the factors, which can explain the bad performance of the standard FM model. The variables of interest that are the most difficult to predict are the GDPs and the Unemployment rate. Their values have been most affected by the crisis. On the other hand, prices maintain good predictability for both regions because this variable was not heavily penalized by the crisis. The impact of outliers can be reduced in the FM FMCD and FM IRLS models, nevertheless, for the US Unemployment Rate, the effects of COVID-19 on the forecasting performances have been disruptive for all the three methodologies.

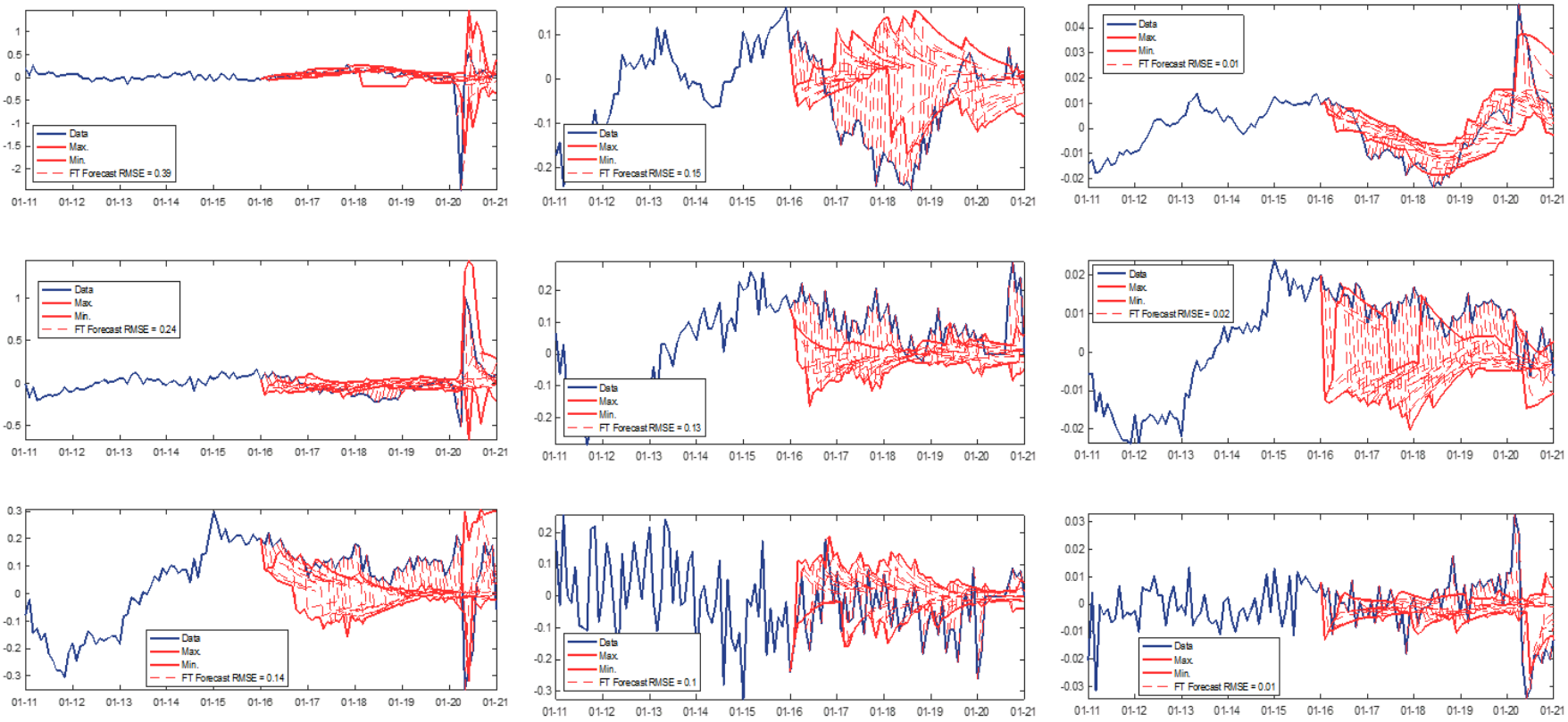

Figure 8. In the rows, the first three factors FM Std (left), FM FMDC (middle) and FM IRLS (right). In each plot, the actual value of the factor (blue solid), the forecasts (red dashed) and the forecast envelop (red solid).

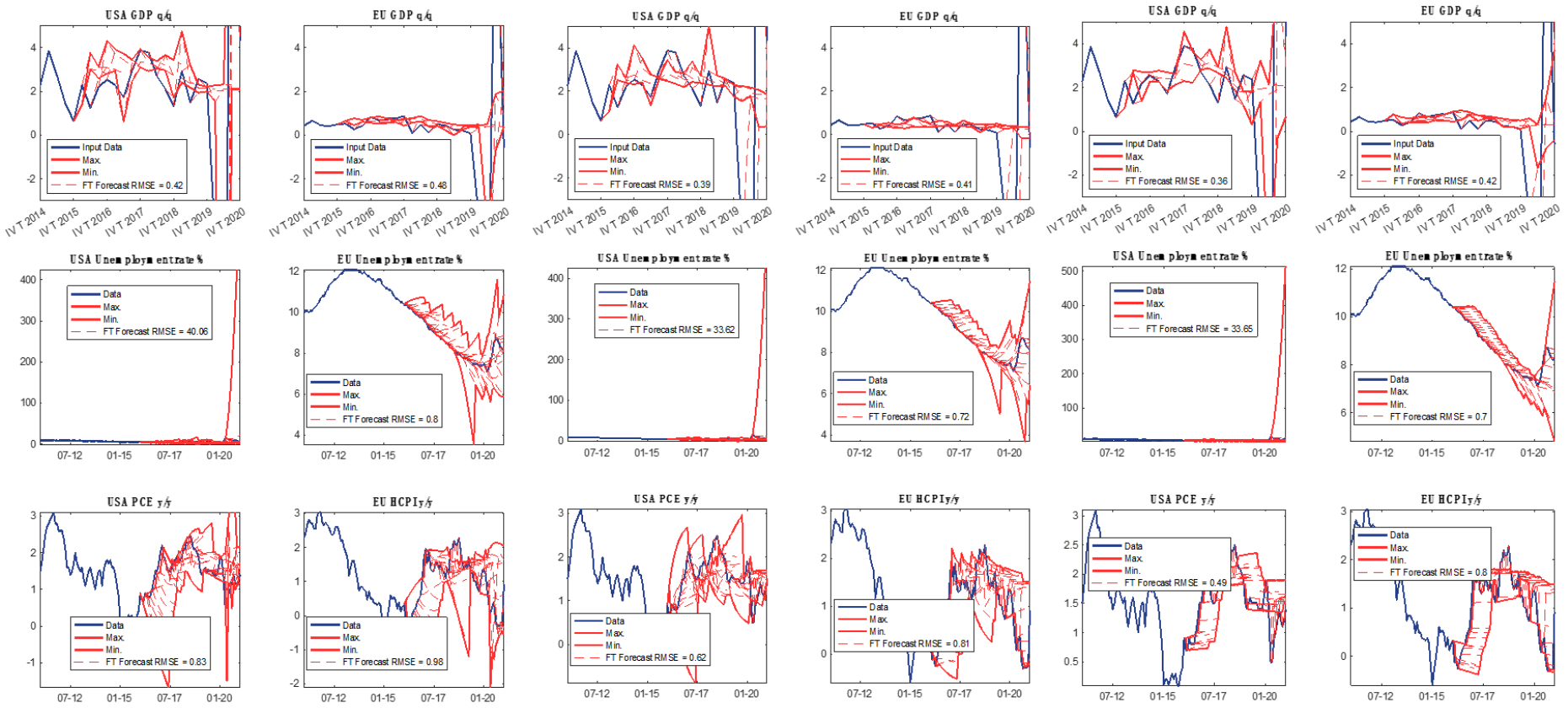

Figure 9. In the rows, the variables forecasted with the FM Std (left), FM FMDC (middle) and FM IRLS (right) model. In each plot, the actual value of the variable (blue solid), the forecasts (red dashed) and the forecast envelop (red solid).

In Table 3(b) we can see that RMSE of the 12-month-ahead forecast for USA Unemployment Rate is smaller than one of the forecasts at 5 and 7 months ahead. This is mainly due to the error magnitude of forecast done in April 2020. This forecast exercise includes in its horizon the first sample impacted by the COVID-19 and has a very large forecast error. Since the dataset ends in January 2021, the forecast horizon s_t in this exercise is 9 months (see Figure 3); which implies that for this forecast it is possible to measure the errors at 1, 5, 7 months but not at 12.

Finally, following the guidelines provided by Eurostat (2020) on modelling outliers due to COVID-19, we monitor sequentially the forecasting errors. The RMSEs for the one-, two-, seven- and twelve-step-ahead forecasts of the three methodologies indicate that the FM FMCD and FM IRLS models have better performances than the FM Std model at all the horizons (see Figures 10, 11 and 12 in Appendix). The numerical results in the panel (b) of Table 3 suggest the FM IRLS model has superior forecasting ability at all horizons.

The bottom line from this section is the following:

- the sample observations during the 2009 crisis and the 2020 COVID-19 pandemic heavily affect factor estimates obtained with the standard procedure;
- consequently, standard factor models can produce significant forecasting errors in the presence of outliers, whereas robust models perform better;
- the variables most impacted by the 2009 crisis and the pandemic (such as GDP and unemployment) exhibit the most significant forecast errors in all estimation procedures;
- the sequential forecasting comparison between MCD and IRLS showed that the latter approach usually leads to superior forecasting performances.

5. Conclusions

Outliers can have disruptive effects on inference, biasing the estimates and the conclusion of the statistical analysis. Through the lens of factor models we provide evidence of the effects of outliers due to the 2009 crisis and the COVID-19 pandemic on the forecast abilities of the models. We applied two techniques for robust factor estimation based on robust covariance matrix estimators. The robust methodologies that we chose have the advantage of avoiding data deletion or manipulation. We compare the standard factor estimation with the robust estimation approaches for an extended period and on a set of relevant variables. The choice to include the COVID-19 pandemic period in the estimation and forecasting exercises has the scope to highlight the relevance of handling outliers in periods of large shocks to the world's economies. We show that robust estimation can reduce outliers' influence and produce good forecasts.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Ahelgebey DF, Billio M, Casarin R (2016a) Bayesian Graphical Models for Structural Vector Autoregressive Processes. *J Appl Economet* 31: 357–386. https://doi.org/10.1002/jae.2443

Ahelgebey DF, Billio M, Casarin R (2016b) Sparse Graphical Vector Autoregression: A Bayesian Approach. *Ann Econ Stat* 123: 333–361. https://doi.org/10.15609/annaeconstat2009.123-124.0333

Artis MJ, Banerjee A, Marcellino M (2005) Factor forecasts for the UK. *J Forecasting* 28. https://doi.org/10.1002/for.957

Bai J, Ng S (2002) Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221. https://doi.org/10.1111/1468-0262.00273

Bai X, Zheng L (2022) Robust factor models for high-dimensional time series and their forecasting. *Commun Stat-Theor M,* 1–14. https://doi.org/10.1080/03610926.2022.2033777

Banbura M, Giannone D, Reichlin L (2010) Large Bayesian vector autoregressions. *J Appl Economet* 25: 71–92. https://doi.org/10.1002/jae.1137

Banbura M, Giannone D, Lenza M (2014) Conditional Forecast and Scenario Analysis with vector autoregressions for large cross-sections. Available from: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1733.pdf.

Barnett V, Lewis T (1994) Outliers in Statistical Data. *Int J Forecasting* 12. https://doi.org/10.1002/bimj.4710370219

Bergstrom P, Edlund O (2014) Robust Registration of point sets using Iteratively Reweighted Least Squares. *Comput Optim Appl* 58: 543–561. https://doi.org/10.1007/s10589-014-9643-2

Billio M, Casarin R, Corradin F (2022) Understanding Economic Instability during the Pandemic: A Factor Model Approach. In Baltagi, B. H., Moscone, F., Tosetti, E., *The Economics of COVID-19*, Emerald Publishing. https://doi.org/10.1108/S0573-855520220000296003

Birch J, Jensen W, Woodall WH (2007) High Breakdown Estimation Methods for Phase I Multivariate Control Charts. *Qual Reliab Eng Int* 23: 615–629. https://doi.org/10.1002/qre.837

Butler RW, Davies PL, Jhun M (1993) Asymptotic for the Minimum Covariance Estimator. *Ann Stat* 21: 1385–1400. https://doi.org/10.1214/aos/1176349264

Casarin R, Corradin F, Ravazzolo F, et al. (2020) A Scoring Rule for Factor and Autoregressive Models Under Misspecification. *Adv Decis Sci* 2: 66–103. https://doi.org/10.47654/v24y2020i2p66-103

Casarin R, Veggente V (2021) Random Projection Methods in Economics and Finance. In Petr, H., Uddin, M.M., Abedin, M. Z., *The Essentials of Machine Learning in Finance and Accounting*, Routledge. https://doi.org/10.4324/9781003037903-6

Cator E, Lopuhaa H (2010) Asymptotic expansion of the minimum covariance determinant estimators, *J Multivariate Anal* 101: 2372–2388. https://doi.org/10.1016/j.jmva.2010.06.009

Choi H, Varian H (2012) Predicting the present with Google trends. *Econ Rec* 88: 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Croux C, Haesbroek G (1999) Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *J Multivariate Anal* 71: 161–190. https://doi.org/10.1006/jmva.1999.1839

Croux C, Filzmoser P, Rousseeuw J, et al. (2003) Robust factor analysis. *J Multivariate Anal* 84: 145–172. https://doi.org/10.1016/S0047-259X(02)00007-6

Davidson R, MacKinnon JG (2004) *Econometric theory and methods.* New York: Oxford University Press.

Davies L (1992) The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. *Ann Stat* 20: 1828–1843. https://doi.org/10.1214/aos/1176348891

Daubechies I, DeVore R, Fornasier M, et al. (2009) Iteratively Reweighted Least Squares minimization for sparse recovery. *Wiley Pure Appl Math* 63: 1–38. https://doi.org/10.1002/cpa.20303

De la Torre F, Black MJ (2004) A framework for robust subspace learning. *Int J Comput Vision* 54: 117–142. https://doi.org/10.1023/A:1023709501986

Diebold FX (2003) "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Lucrezia Reichlin and by Mark W. Watson. In Dewatripont, M, Hansen, L., Turnovsky S., *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Cambridge: Cambridge University Press, 115–122. https://doi.org/10.1017/CBO9780511610264.005

Donoho DL (1982) Breakdown Properties of Multivariate Location Estimators. Qualifying paper, Harward University, Boston.

Einav L, Levin J (2014) Economics in the age of big data. *Science* 346: 715–718. https://doi.org/10.1126/science.1243089

Eurostat (2020) Guidance on Time Series Treatment in the Context of the COVID–19 Crisis. Available from: https://ec.europa.eu/eurostat/documents/10186/10693286/Time_series_treatment_guidance.pdf.

Fabeil NF, Langgat J, Pazim KH (2020) The Impact of COVID–19 Pandemic Crisis on Microenterprises: Entrepreneurs' Perspective on Business Continuity and recovery Strategy. *J Econ Bus* 3: 837–844. https://doi.org/10.31014/aior.1992.03.02.241

Fan J, Wang K, Zhong Y, et al. (2021) Robust High-Dimensional Factor Models with Applications to Statistical Machine Learning. *Stat Sci 3*6: 303–327. https://doi.org/10.1214/20-STS785

Fernandes N (2020) Economic Effects of Coronavirus outbreak (COVID–19) on the world economy. *IESE Business School working paper*. https://doi.org/10.2139/ssrn.3557504

Filzmoser P, van Gaans PFM, van Helvoort PJ (2005) Sequential Factor Analysis as a new approach to multivariate analysis of heterogeneous geochemical datasets: An application to a bulk chemical characterization of fluvial deposits (Rhine-Meuse delta, The Netherlands). *Appl Geochem* 20: 2233–2251. https://doi.org/10.1016/j.apgeochem.2005.08.009

Gambacciani M, Paolella MS (2017) Robust Normal mixtures for financial portfolio allocation. *Econom Stat* 3: 91–111. https://doi.org/10.1016/j.ecosta.2017.02.003

Geman S, McClure D (1987) Statistical methods for tomographic image reconstruction. *Proceedings of the 46th Session of the ISI, Bulletin of the ISI* 52: 5–21.

George EI, Sun D, Ni S (2008) Bayesian stochastic search for VAR model restrictions. *J Econom* 142: 553–580. https://doi.org/10.1016/j.jeconom.2007.08.017

Goldstein S, Pavlovic V, Stolfi J, et al. (2004) Outlier Rejection in Deformable Model Tracking. *2004 Conference on Computer Vision and Pattern Recognition Workshop* 19–19. https://doi.org/10.1109/CVPR.2004.415.

Granger CWJ (1998) Extracting Information from mega–panels and high frequency data. *Stat Neederlanica* 52: 257–272. https://doi.org/10.1111/1467-9574.00084

Green PJ (1984) Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and resistant Alternatives. *J R Stat Soc* 46: 149–170. https://doi.org/10.1111/j.2517-6161.1984.tb01288.x

Hampel FR, Ronchetti EM, Rousseeuw PJ, et al. (1986) Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons.

Hubert M, Debruyne M, Rousseeuw PJ (2017) Minimum covariance determinant and extension. *Wiley Computational Statistics*, 101002. https://doi.org/10.1002/wics.1421

Hubert M (1981) Robust Statistics. *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/0471725250

Kargoll B, Omidalizarandi M, Loth I, et al. (2018) An Iteratively reweighted least squares approach to adaptive robust adjustment of parameters in linear regression models with autoregressive and t-distributed deviations. *J Geodesy* 92: 271–297. https://doi.org/10.1007/s00190-017-1062-6

Koop G, Korobilis D, Pettenuzzo D (2017) Bayesian compressed VARs. *J Economet* 1:1–30. https://doi.org/10.1016/j.jeconom.2018.11.009

Liu K (2021) COVID–19 and the Chinese economy: impacts, policy responses and implications. *Int Rev Appl Econ* 35: 308–330. https://doi.org/10.1080/02692171.2021.1876641

Lopuhaa H, Rousseeuw P (1991) Breackdown points of affine equivalent estimators of multivariate location and covariance matrices. *Ann Stat* 19: 229–248. https://doi.org/10.1214/aos/1176347978

Lütkepohl H (2005) New introduction to multiple time series analysis. *Springer Verlag*. https://doi.org/10.1007/978-3-540-27752-1

Maronna R, Zamar R (2002) Robust Estimates of Location and Dispersion for High–dimensional Datasets. *Technometrics* 44: 307–317. https://doi.org/10.1198/004017002188618509

Mbamalu GAN, Hawary ME (1993) Load forecasting via suboptimal seasonal autoregressive models and Iteratively Reweighted Least Squares. *IEEE T Power Syst* 8: 343–348. https://doi.org/10.1109/59.221222

McKibbin W, Vines D (2020) Global macroeconomic cooperation in response to the COVID-19 pandemic: a roadmap for the G20 and the IMF. *Oxford Rev Econ Pol* 36: S297–S337. https://doi.org/10.1093/oxrep/graa032

McKibbin W, Roshen F (2021) The global macroeconomics impacts of COVID–19: seven scenarios. *Asian Econ Pap* 20: 1–30. https://doi.org/10.1162/asep_a_00796

Mohan K, Fazel M (2012) Iterative Reweighted Algorithms for Matrix Rank Minimization. *J Mach Learn Res* 13: 3441–3473.

Neykov NM, Neytchev PN, Todorov V, et al. (2013) Robust detection of discordant sites in regional frequency analysis. *Water Resour Res* 43: W06417. https://doi.org/10.1029/2006WR005322

Orhan M, Rousseuw PJ, Zaman A (2001) Econometric applications of high- breakdown regression techniques. *Econ Lett* 1: 1–8. https://doi.org/10.1016/S0165-1765(00)00404-3

Rousseuw P (1984) Least Median of Squares Regression. *J Am Stat Assoc* 79: 871–880. https://doi.org/10.1080/01621459.1984.10477105

Rousseeuw P, Leroy AM (1987) Robust Regression and Outliers Detection. *Wiley Series in Probability and Statistics.* https://doi.org/10.1002/0471725382

Rousseeuw P, Van Driessen K (1999) A Fast Algorithm for the minimum Covariance Determinant Estimator. *Technometrics* 41: 212–223. https://doi.org/10.1080/00401706.1999.10485670

Stock JH, Watson WM (2002) Forecasting using principal components from a large number of predictors. *J Am Stat Assoc* 97: 1167–1179. https://doi.org/10.1198/016214502388618960

Stock JH, Watson WM (2004) Combination forecasts of output growth in a seven–country data set. *J f Forecasting* 23: 405–430. https://doi.org/10.1002/for.928

Stock JH, Watson WM (2005) Implications of dynamic factor models for VAR analysis. *Natl Breau Econ Res*. https://doi.org/10.3386/w11467

Stock JH, Watson WM (2009) Forecasting in dynamic factor models subject to structural instability. The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. *Hendry* 173: 205. https://doi.org/10.1093/acprof:oso/9780199237197.001.0001

Stock JH, Watson WM (2012) Disentangling the channels of the 2007–09 recession. *Brookings Pap Eco Ac*, 81–156. https://doi.org/10.1353/eca.2012.0005

Stock JH, Watson WM (2014) Estimating turning points using large data sets. *J Economet* 178: 368–381. https://doi.org/10.1016/j.jeconom.2013.08.034

Varian H (2014) Machine Learning: New tricks for econometrics. *J Econ Perspect* 28: 3–28. https://doi.org/10.1257/jep.28.2.3

Varian H, Scott S (2014) Predicting the present with Bayesian structural time series. *International J Math Model Numer Optim* 5: 4–23. https://doi.org/10.1504/IJMMNO.2014.059942

Vidal R, Ma Y, Sastry SS (2016) *Generalized Principal Component Analysis,* Springer Verlag. https://doi.org/10.1007/978-0-387-87811-9

Zou H, Hastie T (2005) Regularization and variable selection via the elastic-net. *J R Stat Soc B* 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

AIMS Press