UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
15-02
ISSN 2387-0303
January 2015

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Madrid)
Fax (34) 91 624 98 48

Small versus big-data factor extraction in Dynamic Factor Models: An empirical assessment

Pilar Poncela[a] and Esther Ruiz[b, c]

Abstract

In the context of Dynamic Factor Models (DFM), we compare point and interval estimates of the underlying unobserved factors extracted using small and big-data procedures. Our paper differs from previous works in the related literature in several ways. First, we focus on factor extraction rather than on prediction of a given variable in the system. Second, the comparisons are carried out by implementing the procedures considered to the same data. Third, we are interested not only on point estimates but also on confidence intervals for the factors. Based on a simulated system and the macroeconomic data set popularized by Stock and Watson (2012), we show that, for a given procedure, factor estimates based on different cross-sectional dimensions are highly correlated. On the other hand, given the cross-sectional dimension, the Maximum Likelihood Kalman filter and smoother (KFS) factor estimates are highly correlated with those obtained using hybrid Principal Components (PC) and KFS procedures. The PC estimates are somehow less correlated. Finally, the PC intervals based on asymptotic approximations are unrealistically tiny.

Keywords: Confidence intervals, Kalman filter, Principal Components, Quasi-Maximum Likelihood, Sectorial Factors

[a] Corresponding author. Departamento de Análisis Económico: Economía Cuantitativa, Universidad Autónoma de Madrid, C/ Tomás y Valiente, 5, Madrid 28049, Spain. Tel: 34 914975521. Fax: 34 914972991. E-mail: pilar.poncela@uam.es

[b] Departamento de Estadística and Instituto Flores de Lemus, Universidad Carlos III de Madrid.

[c] Financial support from the Spanish Government projects ECO2012-32854 and ECO2012-32401 is acknowledged by the first and second authors respectively. We are very grateful to comments received during the 16th Advances in Econometrics conference on Dynamic Factor Models held in CREATES, Aarhus University in November 2014

Small versus big-data factor extraction in Dynamic Factor Models: An empirical assessment

Pilar Poncela*and Esther Ruiz‡

January 2015

Abstract

In the context of Dynamic Factor Models (DFM), we compare point and interval estimates of the underlying unobserved factors extracted using small and big-data procedures. Our paper differs from previous works in the related literature in several ways. First, we focus on factor extraction rather than on prediction of a given variable in the system. Second, the comparisons are carried out by implementing the procedures considered to the same data. Third, we are interested not only on point estimates but also on confidence intervals for the factors. Based on a simulated system and the macroeconomic data set popularized by Stock and Watson (2012), we show that, for a given procedure, factor estimates based on different cross-sectional dimensions are highly correlated. On the other hand, given the cross-sectional dimension, the Maximum Likelihood Kalman filter and smoother (KFS) factor estimates are highly correlated with those obtained using hybrid Principal Components (PC) and KFS procedures. The PC estimates are somehow less correlated. Finally, the PC intervals based on asymptotic approximations are unrealistically tiny.

KEY WORDS: Confidence intervals, Kalman filter, Principal Components, Quasi-Maximum Likelihood, Sectorial factors.

1 Introduction

It is often argued that macroeconomic and financial variables are governed by a few underlying unobserved factors. Extracting these factors is becoming a central issue that

*Corresponding author. Departamento de Análisis Económico: Economía Cuantitativa, Universidad Autónoma de Madrid, C/Tomás y Valiente, 5, Madrid 28049, Spain. Tel: 34 914975521. Fax: 34 914972991. E-mail: pilar.poncela@uam.es

†Dpt. Estadística and Instituto Flores de Lemus, Universidad Carlos III de Madrid.

‡Financial support from the Spanish Government projects ECO2012-32854 and ECO2012-32401 is acknowledged by the first and second authors respectively. We are very grateful to comments received during the 16th Advances in Econometrics conference on Dynamic Factor Models held in CREATES, Aarhus university in November 2014.

interests econometricians, practitioners and policy decision makers[1]. In this context, dynamic factor models (DFMs), originally introduced by Geweke (1977) and Sargent and Sims (1977), are a very popular instrument to deal with multivariate systems of macroeconomic and financial variables.

The availability of large (sometimes huge) systems has generated a debate about whether small or big-data DFM should be used to obtain more accurate estimates of the underlying factors. The most popular small-data procedure is based on Kalman Filter and smoothing (KFS) algorithms with the parameters estimated by Maximum Likelihood (ML); see, for example, Engle and Watson (1981) for an early reference. On the other hand, big-data procedures are usually based on Principal Components (PC) techniques. Allowing for weak cross-correlations between the idiosyncratic noises, the factors are given by the first few principal components (ordered by their eigenvalues) of the many variables in the system; see, for example, Stock and Watson (2002) and Forni *et al.* (2005). Finally, Doz *et al.* (2011, 2012) propose hybrid methods that combine the PC and KFS (PC-KFS) procedures taking advantage of the best of each of them in such a way that it is possible to deal with big-data systems having efficiency similar to that of KFS. In particular, Doz *et al.* (2011) propose a two-step Kalman filter (2SKF) procedure which is iterated until convergence in the Quasi-Maximum Likelihood (QML) algorithm of Doz *et al.* (2012). Several works compare small and big-data procedures in the context of forecasting one or several variables of interest; see, for example, Boivin and Ng (2006), Bai and Ng (2008b), Banbura and Runstler (2011), Caggiano *et al.* (2011), Alvarez *et al.* (2012) and Banbura and Modugno (2014). However, few works comparing small and big-data procedures focus on factor estimates on their own; see, for example, Bai and Ng (2006b) for the importance of an adequate estimation of factors. Diebold (2003), in a short note, implements KFS to small-data and PC to big-data to extract the common factor from an empirical system of macroeconomic variables and, after visual inspection of the corresponding plots, concludes that nearly the same factor is extracted by both procedures. Alvarez *et al.* (2012) carry out Monte Carlo experiments to compare point factor estimates obtained using small and big-data procedures. For the big-data case, they implement the QML procedure while for the small-data they extract the factors using KFS and conclude that factors extracted using the small scale model have smaller Mean Squared Errors (MSE) than when they are estimated using the big-data procedure. The differences are more pronounced for high levels of cross-correlation among the idiosyncratic noises and, especially, for high persistence in either the common factors or the idiosyncratic

[1] Stock and Watson (1991), Forni *et al.* (2000, 2005), Aruoba *et al.* (2009), Altissimo *et al.* (2010), Camacho and Perez-Quiros (2010) and Frale *et al.* (2011) extract factors to estimate the business cycle; Chamberlain and Rothschild (1983), Diebold and Nerlove (1989), Harvey *et al.* (1994) and Koopman and van der Wel (2013) deal with financial factors; Bernanke *et al.* (2005), Buch *et al.* (2014), Eickmeier *et al.* (in press) and Han (in press) extract factors to incorporate them in FAVAR models; Banerjee *et al.* (2014) and Bräuning and Koopman (2014) propose incorporating factors in FECM and unobserved component models, respectively.

noises. Finally, Doz *et al.* (2012) also carry out Monte Carlo experiments to compare point estimates obtained using PC and the 2SKF and QML procedures.

In this paper we compare point and interval factor estimates obtained using the four procedures mentioned above. Our contribution is different from other papers in the literature in several aspects. First, as just mentioned, our focus is on estimating the underlying factors implementing the same procedures to the same data sets; see Aruoba *et al.* (2009) who suggests that, in order to make a proper empirical comparison among procedures, small versus big-data approaches should be fitted to the same data set. Furthermore, we compare all the most popular procedures available in the literature, namely, KFS, PC and the two hybrid procedures. Finally, we compare not only point estimates but also interval estimates; see Bai (2003) and Bai and Ng (2006b) for the importance of measuring the uncertainty when estimating the underlying factors. We carry out this comparison using both simulated data and the real data base of Stock and Watson (2012).

We compare the small and big-data procedures for different number of variables in the system. Based on asymptotic arguments, several authors argue that the usual methods for factor extraction turn the curse of dimensionality into a blessing[2]. According to Bai (2003), "economists now have the luxury of working with very large data sets." However, one can expect that, when introducing more variables, it is more likely that the weak cross-correlation assumption fails unless the number of factors increases; see Boivin and Ng (2006). Furthermore, when increasing the number of variables is very likely that additional sectorial factors may appear; see, for example, Kose *et al.* (2003) and Moench *et al.* (2013) for sectorial factors. Also, by having more variables, the uncertainty associated with the parameter estimation is expected to increase; see Poncela and Ruiz (2015). Therefore, if one wants to estimate a particular factor, for example, the business cycle, it is not obvious that having more variables in the system increases the accuracy. Finally, it is important to mention that several authors conclude that the factors are already observable when the number of variables in the system is around 30; see Bai and Ng (2008b) and Poncela and Ruiz (2015) when extracting the factors using PC and KFS procedures respectively. We should point out that, in order to avoid the effect of parameter uncertainty, in this paper we consider large time dimension.

We show that, for a given procedure, factor estimates based on different cross-sectional dimensions are highly correlated. On the other hand, given the cross-sectional dimension, the Maximum Likelihood smoothed Kalman filter factor estimates are highly correlated with those obtained using the hybrid PC-KFS procedures. The Principal Components estimates are somehow less correlated. Finally, the PC intervals based

[2]See Stock and Watson (2002a) and Forni *et al.* (2000, 2005) for PC consistency results and Doz *et al.* (2011, 2012) for results on the 2SKF and QML procedures, in stationary systems with weak cross-correlations of the idiosyncratic noises when both the temporal and cross-sectional dimensions tend to infinity.

on asymptotic approximations are unrealistically tiny. Regardless of the dimension of the system, the two-steps procedures are a compromise between the efficiency of KFS procedures and the inefficient but computationally simple and robust PC procedures.

The rest of this paper is organized as follows. In section 2, we establish notation by briefly describing the DFM and the alternative factor extraction procedures considered which are illustrated using simulated data. Section 3 reports the results of a Monte Carlo experiment to analyze the effect of the number of variables and factors on the properties of the factors extracted using the alternative procedures considered in this paper. Section 4 contains the empirical analysis of the Stock and Watson (2012) data base. Section 5 concludes the paper.

2 Extracting common factors

This section establishes notation and briefly describes the DFM and the factor extraction procedures considered, in particular, the PC, KFS, 2SKF and QML procedures. The procedures are illustrated by implementing them to a simulated system.

2.1 The dynamic factor model

Consider the following DFM in which the factors are given by a VAR(p) model and the idiosyncratic noises are assumed to be a VAR(1) process

$$Y_t = PF_t + \varepsilon_t, \tag{1}$$

$$F_t = \Phi_1 F_{t-1} + ... + \Phi_p F_{t-p} + \eta_t, \tag{2}$$

$$\varepsilon_t = \Gamma \varepsilon_{t-1} + a_t \tag{3}$$

where $Y_t = (y_{t1}, ..., y_{tN})'$ is the $N \times 1$ vector of observed variables at time t, $F_t = (f_{t1}, ..., f_{tr})'$ is the $r \times 1$ vector of underlying factors and ε_t is the $N \times 1$ vector of idiosyncratic noises. The disturbance, η_t, is a Gaussian white noise vector with finite and positive covariance matrix Σ_η. The idiosyncratic noises, ε_t, are independently distributed of $\eta_{t-\tau}$ for all leads and lags. Finally, a_t is a Gaussian white noise vector with finite and positive definite covariance matrix Σ_a. The $r \times r$ autoregressive matrices are such that all the roots of the equation $|I_r - \Phi_1 z - ... - \Phi_p z^p| = 0$ are strictly larger than one. Therefore, the factors are zero mean and stationary[3]. Similarly, the idiosyncratic noises are assumed to be zero mean and stationary. Consequently, in the remainder of this paper, we assume that, prior to the analysis, all the series in Y_t are demeaned and transformed to stationarity. We also assume that all autoregressive matrices, Φ_i, $i = 1, ..., p$, and Γ, are diagonal. In this way, the number of parameters is reduced to a manageable size and we avoid to blur the separate identification of the common and idiosyncratic components; see Jungbaker and Koopman (in press) and Pinheiro *et al.*

[3]The stationarity assumption is made in order to implement procedures based on PC.

(2013). The $N \times r$ factor loading matrix is given by $P = [p_{ij}]$ for $i = 1, ..., N$ and $j = 1, ..., r$; see Bai and Ng (2008a), Breitung and Eickemeir (2006) and Stock and Watson (2006, 2011) for excellent surveys on DFM.

As it stands, the DFM defined in equations (1) to (3) is not identifiable; see Bai and Wang (2014) who point out that "the identification problem is not well understood even for static factor models". The factors and factor loadings are only identified up to a pre-multiplication of an invertible matrix. In classical factor analysis, the covariance matrix of the factors, Σ_F, is assumed to be the identity matrix while $P'P$ is a diagonal matrix; see, for example, Bai (2003). Alternatively, in state space models, it is rather common to assume that $\Sigma_\eta = I_r$ together with $p_{ij} = 0$ for $j > i$; see Harvey (1989). In both cases, the factors are assumed to be contemporaneously independent which is an appealing property. With any of these restrictions, F and P are uniquely fixed up to a column sign change given the product FP'. We identify the sign of the estimated factor by imposing $p_{ii} > 0$. These restrictions are arbitrary in the sense that the factors are fixed up to their multiplication by an invertible matrix. Consequently, the factors obtained may not lead to a particularly useful interpretation. However, once they have been estimated, the factors can be rotated to be appropriately interpreted.

There are several particular cases of the DFM in equations (1) to (3) that have attracted a lot of attention in the related literature. When $\Gamma = 0$ and Σ_a is diagonal, the idiosyncratic noises are contemporaneously and serially independent. In this case, the DFM is known as *strict*. When there is serial correlation with Γ being diagonal, the model is known as *exact*. Chamberlain and Rothschild (1983) introduce the term "*approximate factor structure*" in static factor models where the idiosyncratic components do not need to have a diagonal covariance matrix.

Next, we briefly describe each of the four procedures considered in this paper to extract the factors in DFM.

2.2 Principal Components

In the context of big-data, the factors are usually extracted using procedures based on PC which are attractive because they are computationally simple and are nonparametric and, consequently, robust to potential misspecifications of the dynamics of factors and idiosyncratic noises. The price to pay for this robustness is that PC extraction losses efficiency with respect to procedures based on well specified dynamics.

In this section, we describe the PC procedure following Bai (2003). PC procedures allow estimating the space spanned by the factors. Consequently, in order to extract the individual factors one needs to know r, the number of factors in the system. The $T \times r$ matrix of PC factor estimates, $\widehat{F} = \left(\widehat{F}_1', ..., \widehat{F}_T'\right)'$, is given by $\sqrt{T}$ times the r eigenvectors associated with the r largest eigenvalues of YY', where Y is the $T \times N$ matrix given by $Y = (Y_1'...Y_T')'$, arranged in decreasing order. Then, assuming that

$\frac{1}{T}\widehat{F}'\widehat{F} = I_r$, the estimates of the loadings are given by

$$\widehat{P} = \frac{Y'\widehat{F}}{T}.$$

The properties of the PC factors are based on asymptotic arguments when both the cross-sectional, N, and temporal, T, dimensions tend simultaneously to infinity. Stock and Watson (2002) show that, if the cross-correlations of the idiosyncratic noises are weak and the variability of the common factors is not too small, the estimated factors are consistent[4]. Under general conditions that allow for serial and contemporaneous cross-correlation and heteroscedasticity in both dimensions, Bai (2003) shows that the estimated factors can be treated as if they were observed as long as the number of factors is known and fixed as both N and T grow and $\frac{\sqrt{T}}{N} \to 0$ and $N, T \to \infty$. He also derives the following asymptotic distribution

$$\sqrt{N}(\widehat{F}_t - H'F_t) \xrightarrow{d} N(0, V^{-1}Q\Gamma_t Q'V^{-1}), \tag{4}$$

where $H = \widehat{V}^{-1}\left(\widehat{F}'F/T\right)(P'P/N)$ with $\widehat{V}$ being the $r \times r$ diagonal matrix consisting of the first r largest eigenvalues of the matrix $YY'/(TN)$, arranged in decreasing order, and V is its limit in probability, Q being the $r \times r$ limit in probability matrix of $\frac{\widehat{F}'F}{T}$ and the $r \times r$ matrix Γ_t is defined as follows

$$\Gamma_t = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} E(p_{i.}'p_{j.}\varepsilon_{it}\varepsilon_{jt}),$$

with $p_{i.}$ being the $i-th$ row of the factor loading matrix P. Given that the factors are estimated according to the normalization $\frac{\widehat{F}'\widehat{F}}{T} = I_r$, an estimate of Q is just the identity matrix. Therefore, an estimate of the asymptotic variance of $\widehat{F}_t$ would be

$$var(\widehat{F}_t) = \frac{1}{N}\widehat{V}^{-1}\widehat{\Gamma}_t\widehat{V}^{-1}, \tag{5}$$

with $\widehat{\Gamma}_t$ being a consistent estimate of Γ_t (or more precisely of $H^{-1}\Gamma_t H^{-1}$). Bai and Ng (2006a) propose three different estimators of Γ_t depending on the properties of the idiosyncratic errors. Two of them assume cross-sectionally uncorrelated idiosyncratic errors but do not require stationarity while the third is robust to cross-sectional correlation and heteroscedasticity but requires covariance stationarity. Bai and Ng (2006a) argue that for small cross-correlation in the errors, constraining them to be zero could sometimes be desirable because the sampling variability from estimating them could generate nontrivial efficiency loss. Consequently, they recommend using the following estimator of Γ_t

$$\widehat{\Gamma}_t = \frac{1}{N}\sum_{i=1}^{N}\widehat{p}_{i.}'\widehat{p}_{i.}\widehat{\varepsilon}_{it}^2, \tag{6}$$

[4]Onatski (2012) describes situations in which PC is inconsistent while Choi (2012) derives the asymptotic distribution of a Generalized Principal Component estimator with smaller asymptotic variance.

where $\widehat{\varepsilon}_t = (\widehat{\varepsilon}_{1t}, ..., \widehat{\varepsilon}_{Nt})'$ is obtained as $\widehat{\varepsilon}_t = Y_t - \widehat{P}\widehat{F}_t$.

In order to illustrate factor extraction using PC and to analyze the roll of N and r on the results, we generate a system of $N = 120$ variables with $T = 200$ observations from the following DFM with $r = 3$ factors

$$Y_t = PF_t + \varepsilon_t, \tag{7}$$

$$F_t = \begin{bmatrix} 1.3 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} F_{t-1} + \begin{bmatrix} -0.36 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} F_{t-2} + \eta_t, \tag{8}$$

where the weights of the first factor, p_{i1}, for $i = 1, ..., N$, are generated by a uniform distribution in $[0, 1]$; see Bai and Ng (2006a) who also carry out simulations generating the weights by a uniform distribution. The weights of the second factor are generated such that $p_{i2} \neq 0$, for $i = 13, ..., 60$ and $p_{i2} = 0$ otherwise. When different from zero, the weights are also generated by a uniform distribution. Note that the second factor only affects the variables from $i = 13, ..., 60$. Finally, the weights of the third factor, $p_{i3} = 0$ for $i = 1, ..., 60$ and generated from an uniform distribution for $i = 61, ..., 120$. Consequently, the third factor affects the last 60 variables in the system. These two latter factors have a block structure as they are specific to subsets of variables. They can be considered as sectorial factors, likely to appear when big-data systems are considered. The covariance matrix of the idiosyncratic errors is given by $\Sigma_\varepsilon = I_N$ so that the errors are homoscedastic and mutually uncorrelated. Finally, the first factor is given by an AR(2) process with roots 0.9 and 0.4 while the second and third factors are stationary AR(1) processes with roots 0.9 and 0.5 respectively. The variances of the three factors are 1. The PC factor extraction procedure is implemented to extract the factors with different number of variables and factors in the system. First, we consider $N = 12$ variables (small-data) with the first 12 variables being selected ($r = 1$); second, $N = 12$ variables are selected from the $13th$ to the $24th$ so that $r = 2$ and the second factor only has weights on a subset of variables; third, $N = 12$ variables are selected from the $55th$ to the $66th$ so that $r = 3$; fourth, $N = 30$ (medium-data) with variables from the $46th$ to the $75th$ being chosen so that $r = 3$; fifth, we consider extracting the factors using all $N = 120$ variables (big-data). Previous to implementing the PC procedure, the number of factors is selected by using the procedure proposed by Onatski (2009)[5].

[5] When the factors are extracted using PC, it is fundamental to have an unbiased estimate of r. Note that the factors are not uniquely identified which means that even when the objective is the estimation of an unique factor, it is important to know r so that the estimated factors can be rotated to obtain the desired interpretable estimation. There is a large number of alternative proposals of estimating the number of factors, mostly based on the eigenvalues of the sample covariance matrix of Y_t; see, for example, Bai and Ng (2002, 2007), Amengual and Watson (2007), Alessi *et al.* (2010), Kapetanios (2010) and Breitung and Pigorsch (2013), among others. These procedures require that the cumulative effect of the factors grows as fast as N. Alternatively, Onatski (2009, 2010) proposes an estimator of the number of factors that works well when the idiosyncratic terms are substantially serially or cross-sectionally correlated. Onatski (2009) formalize the widely used empirical method of the number of

As mentioned above, the factors are estimated up to a rotation. Consequently, we compare the true and PC estimated factors in the scale of the true factors by rotating the estimated factors as follows

$$\widehat{F}^* = \widehat{F}\left[\left(\widehat{P}'\widehat{P}\right)^{-1}\widehat{P}'P\right]^{-1}. \tag{9}$$

Although most authors compare the spaces spanned by the true and the estimated factors, we are also interested in the quality of the estimation of each individual factor. The reason of this interest is that, as mentioned in the Introduction, in many macroeconomic applications, the interest is estimating a particular factor as, for example, the business cycle. Consider first the design when $N = 12$ and $r = 1$. The top panel of Figure 1 plots the true factor, F, together with the (rotated) factor estimated by PC, $\widehat{F}^*$, and the corresponding pointwise 95% intervals constructed using the asymptotic MSE in (5). Observe that the point estimates of the factor follow rather closely the evolution of the true factor with the correlation between the true and the estimated factor being 0.87; see Table 1. However, the asymptotic intervals are extremely tiny with the true factor lying outside them most of the time. The coverage of the 95% asymptotic intervals, also reported in Table 1, is 46%. Note that when $N = 12$ and the number of factors is either $r = 2$ or $r = 3$, Table 1 shows that the estimated first factor is only slightly less correlated with the corresponding true factor. This result seems to contradict Bai and Ng (2008a) who conclude that the precision falls with the number of factors. However, note that the precision in the estimation of the second factor is relatively small and Bai and Ng (2008a) are computing an "average" precision.

Next, consider the illustration when $N = 120$ (big-data) and $r = 3$. The first row of Figure 2 plots the same quantities as above for each of the three factors in the system. As before, we can observe that the PC factors estimates follow very closely the true factors. The correlation between true and (rotated) estimated first factor, reported in Table 1, is 0.97 which is larger than when estimating the first factor using $N = 12$ variables. However, the empirical coverage of the 95% asymptotic intervals is still 46%. For the same number of factors ($r = 3$) and a medium-size data set ($N = 30$), the correlation between the true and estimated first rotated factor is between the previous two ones (0.91). The empirical coverage remains around 46%.

2.3 Kalman filter and smoothing

In the context of small-data, the factors are usually estimated using the KFS algorithms with the parameters estimated by ML. Running the Kalman filter requires knowing the

factors determination based on the visual inspection of the sree plot introduced by Cattell (1966). In the empirical application, we select r using Alessi *et al.* (2010) due to its good performance. However, this procedure requires monitorization of plots and, consequently, in the simulations, we implement the criterion by Onatski (2009).

specification and parameters of the DFM in equations (1) to (3). Therefore, the factors extracted using the KFS algorithms can be non-robust in the presence of model misspecification. However, if the model is correctly specified, extracting the factors using the Kalman filter is attractive for several reasons. First, it allows to deal with data irregularities as, for example, systems containing variables with different frequencies and/or missing observations; see Aruoba *et al.* (2009), Jungbaker *et al.* (2011), Pinheiro *et al.* (2013), Banbura and Modugno (2014) and Bräuning and Koopman (2014) for some examples. Second, KFS algorithms are not so affected by outliers as PC procedures which are based on estimated covariance matrices. Third, they provide a framework for incorporating restrictions derived from economic theory; see Bork *et al.* (2009) and Doz *et al.* (2012). Fourth, the KFS procedures are more efficient than PC procedures for a flexible range of specifications that include non-stationary DFM and idiosyncratic noises with strong cross-correlations. Finally, they allow obtaining uncertainty measures associated with the estimated factors when the cross-sectional dimension is finite; see Poncela and Ruiz (2015). However, the number of parameters that need to be estimated increase with the cross-sectional dimension in such a way that ML estimation is unfeasible for moderate systems. Jungbacker *et al.* (2011) and Jungbacker and Koopman (in press) propose a computationally feasible device to deal with large dimensional unobserved component models using the Kalman filter. However, if the cross-sectional dimension is large, this procedure is only feasible if the idiosyncratic noises are serially uncorrelated. Fiorentini *et al.* (2014) also propose an alternative spectral EM algorithm capable of dealing with large systems.

The DFM in equations (1) to (3) is conditionally Gaussian. Consequently, when the idiosyncratic noises are serially uncorrelated, the KFS algorithms provide Minimum MSE (MMSE) estimates of the underlying factors which are given by the corresponding conditional means. Denoting by $f_{t|\tau}$ the estimate of F_t obtained with the information available up to time τ, and by $V_{t|\tau}$ its corresponding MSE, KFS delivers

$$f_{t|\tau} = E\left[F_t|Y_1, ..., Y_\tau\right], \tag{10}$$

$$V_{t|\tau} = E\left[(F_t - f_{t|\tau})(F_t - f_{t|\tau})'|Y_1, ..., Y_\tau\right], \tag{11}$$

where $\tau = t-1$, for one-step-ahead predictions, $\tau = t$ for filtered estimates and $\tau = T$ for smoothed factor estimates. It is also important to point out that the KFS algorithms deliver out-of-sample forecasts of the factors together with their corresponding Mean Squared Forecast Errors (MSFE). In this paper, our focus is on smoothed estimates so that they can be compared with those obtained from alternative procedures.

When the idiosyncratic noises are serially correlated, the DFM can be reformulated in two alternative ways to preserve the optimal properties of KFS. First, it is possible

to express the DFM in state space form as follows

$$
\begin{aligned}
Y_t &= \Gamma Y_{t-1} + \begin{bmatrix} P & -\Gamma P \end{bmatrix} \begin{bmatrix} F_t \\ F_{t-1} \end{bmatrix} + a_t \qquad (12) \\
\begin{bmatrix} F_t \\ F_{t-1} \end{bmatrix} &= \begin{bmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{bmatrix} \begin{bmatrix} F_{t-1} \\ F_{t-2} \end{bmatrix} + \begin{bmatrix} \eta_t \\ 0 \end{bmatrix};
\end{aligned}
$$

see Reis and Watson (2010), Jungbacker *et al.* (2011) and Pinheiro *et al.* (2013) for implementations of the model in (12). One can alternatively deal with the autocorrelation of the idiosyncratic noises by augmenting the state vector by ε_t; see, for example, Jungbacker *et al.* (2011), Banbura and Modugno (2014) and Jungbacker and Koopman (in press). Both formulations lead to the same results when the initialization issues are properly accounted for. However, note that, in practice, augmenting the state space is only feasible for relatively small cross-sectional dimensions.

The parameters are usually estimated by Maximum Likelihood (ML) maximizing the one-step-ahead decomposition of the log-Gaussian likelihood; see Engle and Watson (1981) and Watson and Engle (1983). The maximization of the log-likelihood entails nonlinear optimization which restricts the number of parameters that can be estimated and, consequently, the number of series that can be handled when estimating the underlying factors. Even if the number of factors is considered as fixed, the number of parameters to be estimated increases very quickly with N. Consequently, the estimation problem is challenging if not impossible. Although the EM algorithm allows to maximize the likelihood function of very large DFM, it does not allow the estimation of the parameters in Γ; see Doz *et al.* (2012). Alternatively, Jungbacker and Koopman (in press) propose to transform the observation equation into a lower dimension which leads to a computationally efficient approach to parameter and factor estimation.

With respect to the uncertainty associated with the KFS estimates, Poncela and Ruiz (2015) obtain expressions of the finite N and T steady-state MSE associated with the common factors estimated by the KFS procedure both when the model parameters are known and when they are estimated using a consistent estimator. They show that, in the first case, the MSE are decreasing functions of the cross-sectional dimension regardless of whether the idiosyncratic noises are weakly or strongly correlated. Furthermore, if the idiosyncratic noises are weakly correlated, the minimum MSE are zero for filtered and smoothed estimates while if they are strongly correlated, the minimum MSE are different from zero, so the factor estimates are not consistent. However, it is very important to remark that, in any case, the MSE is very close to the minimum when the number of variables in the system is relatively small, approximately around 30 variables. In the latter case, when the parameters are estimated, if the sample size is fixed, the MSE can even be an increasing function of the cross-sectional dimension. Therefore, in this case, which is the most common when dealing with empirical data, one can have more uncertainty about the underlying factors when the number of series used to estimate them increases.

The KFS procedure is illustrated using the same simulated system considered in the previous subsection when illustrating the PC factor extraction. In order to compare the factor estimates obtained by both procedures, we also compute the KFS estimates in units of the true factors using an analogous transformation to (9).

The second panel of Figure 1 plots the KFS estimated factor together with the 95% interval obtained using the MSE provided by the filter. We can observe that the true and estimated factors move closely together with the intervals containing the true factor most of the time. The correlation between the estimated and true factors, reported in Table 1, is 0.95, larger than the correlation observed when the factor is extracted using PC which was 0.87. Furthermore, the coverage of the intervals is 95%, equal to the nominal. Similar conclusions are obtained when there are two or three factors in the system; see Table 1. Obviously, the correlations and coverages are slightly worse than when $r = 1$ but better than when extracting the factors using PC. When the cross dimension increases to $N = 30$ variables and there are three factors ($r = 3$), the correlation between the first true and estimated rotated factors remains at 0.95 (see Table 1). Finally, Figure 2 plots the factors extracted when $N = 120$. In this case, the correlation between the first true and estimated rotated factors is slightly higher, 0.96. Therefore, using the big-data system is only getting very minor improvements with respect to using $N = 30$ variables.

It is important to note that, if the objective is the estimation of the first common factor, the presence of additional factors only affects marginally the extraction of the factor of interest.

2.4 Principal Components-Kalman filter smoothing

Doz *et al.* (2011, 2012) propose two further two-steps procedures to estimate the factors in the presence of big-data systems based on combining the PC and KFS approaches; see Giannone *et al.* (2008) for previous empirical applications and Banbura and Modugno (2014) for an extension to systems with missing data.

The 2SKF procedure proposed by Doz *et al.* (2011) starts extracting the factors by PC. Then, the factors' dynamics are estimated after fitting a VAR(1) model which is estimated by Least Squares (LS), i.e. $\widehat{\Phi}^{(0)} = \sum_{t=1}^{T}(\widehat{F}_{t-1}^{(0)}\widehat{F}_{t-1}^{(0)\prime})^{-1}(\widehat{F}_{t-1}^{(0)}\widehat{F}_{t}^{(0)\prime})$ where $\widehat{F}_t^{(0)} = \widehat{F}_t^{PC}$. The parameters in Σ_ε are estimated using the sample covariance matrix of the residuals as follows

$$\widehat{\Sigma}_\varepsilon^{(0)} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\varepsilon}_t^{(0)}\widehat{\varepsilon}_t^{(0)\prime},$$

where $\widehat{\varepsilon}_t^{(0)} = Y_t - \widehat{P}^{(0)}\widehat{F}_t^{(0)}$and $\widehat{P}^{(0)} = \widehat{P}^{PC}$. Setting $\Sigma_\eta = I$ for identification purposes, in the second step, the factors are estimated by running the smoothing algorithm of the Kalman filter implemented in the DFM in equations (1) to (3) with $\Gamma = 0$ and the parameters substituted by $\widehat{P}^{(0)}$, $\widehat{\Phi}^{(0)}$ and $\widehat{\Sigma}_\varepsilon^{(0)}$. In the second step, the factors are

estimated implementing the Kalman filter smoother with the estimated parameters and assuming that the idiosyncratic noises are serially and contemporaneously uncorrelated. The MSE of the factors are directly obtained from the Kalman filter.

Consider now the simulated system used above for illustrating the PC and KFS procedures. The third row of Figure 1, which plots the (rotated) estimates of the factor obtained using the 2SKF procedure together with the 95% intervals and the true simulated factor, shows that the factor estimated with $N = 12$ variables has a behavior similar to that of the PC and KFS estimates. The correlation with the true factor is 0.92. However, the coverage of the intervals is closer to that of KFS. In practice, the performance of the 2SKF estimates is a compromise between PC and ML estimates. The same conclusions can be obtained for all other cases considered.

Doz *et al.* (2012) propose a QML procedure based on iterating the 2SKF. Actually, this is equivalent to maximum likelihood estimation implemented through the EM algorithm when the idiosyncratic noises are white noise. Given $\widehat{F}_t^{(i)}$, obtained at step i, the two steps of the 2SKF procedure are iterated by re-estimating the VAR parameters, the factor loadings and the variance of the error term in equation (1) as explained above. At each iteration, the algorithm ensures higher values of the log-likelihood. The process converges when the slope between two consecutive log-likelihood values is lower than a given threshold. The MSE of the factors are directly obtained from the Kalman filter in the last step; see Banbura and Runstler (2011) for an application in which they use the MSE to compute the weights for the predictions of a variable of interest.

Consider the simulated system used as an illustration. The last row of Figure 1 plots the (rotated) factor extracted using the iterated QML-EM procedure together with the corresponding 95% intervals and the true simulated factor. Once more, we can observed that the performance of the QML-EM estimates is very similar to that of alternative procedures in terms of point estimates. However, the coverages of the confidence intervals is similar to that of KFS; see Table 1.

3 Monte Carlo experiments

In this section, we carry out Monte Carlo experiments to compare the finite sample performance of the four procedures considered both in terms of point and interval factor estimates. The results are based on $R = 500$ replicates generated from the same system considered in the illustration. As in the illustration, we consider situations with different number of variables and factors. Previous to the factor extraction, the number of factors is determined using the criterion by Onatski (2009). Table 2 reports the percentage of failures of the test. Observe that the performance is appropriate when the number of series is relatively large with respect to the number of factors. When $N = 12$ and there are 3 factors in the system, Onatski (2009) procedure only detects them in 18% of the replicates. When the number of factors detected implementing the Onatski (2009) procedure coincides with the true number of factors in the system, we

extract the factors using each of the four procedures considered.

In order to assess the precision of the point factor estimates, Table 2 reports the Monte Carlo averages and standard deviations of the corresponding trace R^2 of the regression of the estimated on the true factors given by[6]

$$\frac{Trace\left(F'\widehat{F}\left(\widehat{F}'\widehat{F}\right)^{-1}\widehat{F}'F\right)}{Trace(F'F)}. \tag{13}$$

The trace measure in (13), which is smaller than 1 and tends to 1 with the increasing canonical correlations between the estimated and true factors[7], has been implemented by, for example, Stock and Watson (2002), Boivin and Ng (2006), Doz *et al.* (2008, 2012) and Banbura and Modugno (2014), to assess the precision of factor estimates.

First of all observe that, as expected, regardless of the procedure, if the number of variables is fixed, the trace statistic decreases when the number of factors increases. On the other hand, if r is fixed, the trace statistic increases with the number of variables. Also, it is important to note that the trace statistics of the KFS and QML procedures are very similar in all cases. On the other hand, the trace statistics of PC are clearly smaller while 2SKF is somehow in between. If $N > 30$, depending on the number of variables and factor in the system, it seems that just one iteration of the Kalman filter is enough to obtain similar factor estimates as with the KFS. Only when $N = 120$ and $r = 3$, the trace statistics of all procedures are similar and over 0.9. Furthermore, note that with $N = 30$, the Kalman based procedures have statistics over 0.8. Finally, Table 2 shows that, when using the KFS or QML procedures to extract the factors, a remarkably large precision is obtained even with $N = 12$ if there is just one single factor in the system. If by adding more variables, the number of factors increases, the precision is similar. Regardless of the procedure and number of factors, all procedures considered are adequate to estimate the space spanned by the factors if $N > 30$.

As the objective of this paper is not only to assess the accuracy of point factor estimates but also of interval estimates, Table 3 reports the Monte Carlo means and standard deviations of the empirical coverages of the pointwise intervals of the factors extracted by each of the four procedures. The MSE used to construct the PC intervals are the asymptotic MSE in equation (5) while the MSE of the other three procedures are obtained from the Kalman smoother when the model parameters are substituted by the corresponding estimated parameters. Note that these MSE do not incorporate the uncertainty associated with the parameter estimation. The nominal coverage of the intervals is 95%. Table 3 shows that the asymptotic MSE used to construct the intervals for the PC factor estimates are clearly small to represent the uncertainty associated with these estimates. Furthermore, it is possible to observe that the undercoverage is

[6] It is important to point out that the results reported in Table 2 corresponds to those replicates in which the number of factors detected by Onastki (2009) is correct. Therefore, they are conditional on the number of factors and based on a number of replicates smaller than 500.

[7] The results based on canonical correlations are available upon request.

more severe as more variables are used to extract the factors. There are two potential reasons for this counterintuitive fact. First, note that as more variables are considered, the number of failures of the Onastki (2009) procedure to select the number of factors is smaller. Therefore, the Monte Carlo results reported in Table 3 are based on more replicates as N increases. One can expect the strength of the factors to be larger when less replicates are used. Second, as N increases, the parameter uncertainty also increases and, consequently, the intervals that do not incorporate this uncertainty are less reliable. The interaction between both effects can explain why the coverages of the PC intervals can deteriorate as N increases. The coverages of the three other procedures considered are appropriate when $N = 12$ and $r = 1$. However, when the number of factors is larger than one, the empirical coverages are well under the nominal. Note that, in each case, the coverages are similar for all the factors in the system. Therefore, if the interest is estimating just one factor (for example, the business cycle), having more factors in the system could deteriorate the interval estimation. As when looking at the PC intervals, given the number of factors in the system, the performance of the intervals deteriorates when N increases. Therefore, according to the Monte Carlo results reported in Table 3, if one wants to obtain interval estimates of a single factor, it is better to keep the number of variables to be relatively small so that no new additional factors are introduced in the system.

It is important to point out that, if the interest is the estimation of a single factor common to all variables in the system in a multifactor model, it is not straightforward to find an adequate rotation of the factors that estimates properly this particular factor. In our experience, the estimated factors are interchanged in a number of replicates, i.e. the common factor is estimated as the sectorial factor and vice versa. Consequently, the individual correlations between each estimated factor and the corresponding true factor could be much smaller than what the statistics reported in Table 2 might suggest.

The results in this section are based on a medium sample size, $T = 200$, and in a model in which the idiosyncratic noises are temporally uncorrelated. Of course, large sample sizes will deliver even better results. On the other hand, smaller sample sizes are not very likely to be used in any sensible empirical application. With respect to the lack of temporal dependence of the idiosyncratic noises, it could be of interest to compare the four procedures in this case.

4 Empirical analysis

In this section, we analyze the monthly series contained in the data base considered by Stock and Watson (2012), which consists of an unbalanced panel of 137 US macroeconomic variables (two of which are deflators and other six not included in the analysis) observed monthly from January 1959 to December 2011. These variables have been deflated, transformed to stationarity and corrected by outliers as described by Stock and Watson (2012). Furthermore, as it is usual in this literature, they have been stan-

dardized to have zero mean and variance one. The variables can be classified into the following 12 economic categories: industrial production (13); employment and unemployment (44); housing starts (8); inventories, orders and sales (8); prices (16); earnings (3); interest rates and spreads (18); money and credit (14); stock prices (3); housing prices (2); exchange rates (6); and other (2), with the number of the series in the category given in parentheses. In order to obtain a balanced panel, we select those variables observed without missing values over the whole observation period. The resulting balanced panel has $N = 103$ variables, classified into 11 categories and $T = 628$ observations. All variables belonging to the *Housing Prices* category disappear from the panel. The objective of this empirical exercise is to answer the following questions: i) One the interest is to estimate just one factor as, for example, the business cycle, is it worth to use all available variables to extract it?; ii) Are the factor extraction procedure and number of variables used relevant to estimate the factors?; iii) Is the number of factors in the system independent of the number of variables?

We start the analysis by extracting the factors from a system with 11 variables each of them representing one of the categories. Each variable has been chosen as that exhibiting the highest averaged correlation with respect to the remaining series in the same category; see Alvarez *et al.* (2012) for this criterion. In this system, we start selecting the number of factors as proposed by Alessi *et al.* (2010) who, following Hallin and Liska (2007), introduce a tuning multiplicative constant to improve the performance of the procedure proposed by Bai and Ng (2002). The number of factors selected is one. The factor is then extracted by each of the four procedures described above. Figure 3 plots the extracted factor together with its corresponding 95% pointwise intervals. The corresponding Root MSE (RMSE), computed without incorporating the parameter uncertainty, have been reported in the main diagonal of Table 4 for each of the four procedures considered. As already concluded from the simulated system used in the illustration, we can observe that the asymptotic RMSE of the PC procedure are unrealistically small. Figure 3 illustrates that, regardless of the factor extraction procedure, the point estimates of the factors extracted using the information contained in the 11 variables selected from the original data base are very similar.

Next, we add into the set of variables used to extract the factors the variables with second highest correlation, with $N = 21$ variables. In this case, the number of factors identified using Alessi *et al.* (2010) is again 1. Figure 4 plots the factors extracted by each procedure together with the corresponding pointwise 95% pointwise intervals. Then, we extract the factors with $N = 91$ variables and $\widehat{r} = 4$; see Figure 5 which plots the first extracted factor. Finally, the factors are extracted using all $N = 103$ variables. For each procedure, Table 4 reports the RMSE together with the correlations between the factors extracted when the cross-sectional dimension changes. We can observe that, in general, the RMSE decrease with N. However, the MSE when $N = 91$ and when $N = 103$ are very similar. For each procedure, Table 4 also reports

the correlations between the factors estimated with different cross-sectional dimensions. These correlations are very high, being always over 0.85. It seems that regardless of the procedure implemented for factor extraction, increasing the number of variables only pays a very marginal increase in terms of factor estimation accuracy.

Finally, we compare the factors extracted using different procedures with the same number of variables. Table 5, that reports the correlations between the estimated factors obtained by the alternative procedures, shows that there is a high correlation between the factor estimates extracted using KFS and QML-EM, which is always over 0.95. The same happens with the correlations between the factors extracted using PC and 2SKF which are always over 0.97. These results confirm the conclusions obtained with the simulated system used in the illustration and the Monte Carlo experiment.

5 Conclusions

In this paper, we compare small-data and big-data factor extraction procedures implementing the alternative procedures considered to the same data sets. Using simulated and real data, we compare PC, KFS, 2SKF and QML, given the sample size. We also compare the performance of each procedure for different cross sectional dimensions. We conclude that, regardless of the procedure implemented and the number of variables used for the factor estimation, (the spaces spanned by) the factors extracted are very similar. When using simulated data, all procedures extract (conveniently rotated) factors highly correlated with the true unobserved factors. If the objective is estimating a given factor (as, for example, the business cycle) adding more variables into the system may increase the number of factors but the increase in accuracy of point estimates is relatively small. We also show that the asymptotic bounds of PC are too narrow being inadequate to represent the finite sample uncertainty associated with estimated factors. A closer look to an illustration shows that both ML and QML procedures extract very similar point and interval factors which have, in general, higher correlations with the true factors than PC and 2SKF estimates when the cross-sectional dimension is relatively small.

In this manuscript, we did not consider the effect of parameter estimation on the construction of intervals for the factors; see Poncela and Ruiz (2015) for this effect in the context of the ML procedure. However the empirical coverages reported in Table 3 are smaller than the nominal coverages. Furthermore, the interval coverages of all procedures decrease with the number of series, probably as a result of increasing the number of parameters that we have to estimate as Poncela and Ruiz (2015) point out for Kalman filter estimations. Recall that the estimation is only carried out for those cases where the true number of factors is detected. At this regard, the number of factors correctly found by Onatski's (2009) test increases with the number of series. On the contrary, the interval coverages decrease with the number of series. Therefore, it seems that incorporating the parameter uncertainty could be important to get more adequate

confidence intervals. When dealing with ML or the hybrid procedures, this uncertainty can be incorporated in practice using bootstrap procedures as those proposed by Rodríguez and Ruiz (2009, 2012) in the context of state space models. However, as far as we know, there are not procedures proposed in the literature to incorporate the parameter uncertainty in the context of PC procedures. Looking at the effects of parameter uncertainty when constructing intervals for estimated factors in empirical applications is within our research agenda. Also, the analysis of real data systems can be extended to consider unbalanced data bases by using, for example, the computationally efficient procedures by Jungbacker *et al.* (2011) and Jungbacker and Koopman (in press).

Finally, it is important to mention that, in practice, the models fitted could be misspecified. In stationary DFMs, Doz *et al.* (2011, 2012) show the consistency of the factors estimated using the PC-KFS procedures so that he misspecification of the idiosyncratic noise serial correlation does not jeopardize the consistent estimation of the factors. Considering the effects of misspecification both in the number of factors and/or in the dynamics of factors and idiosyncratic noises is also left for further research.

References

[1] Alessi, L., M. Barigozzi and M. Capasso (2010), Improved penalization for determining the number of factors in approximate factor models, *Statistics and Probability Letters,* 80, 1806-1813.

[2] Altissimo, F., R. Cristadoro, M. Forni, M. Lippi and G. Veronese (2010), New eurocoin: tracking economic growth in real time, *The Review of Economics and Statistics,* 92(4), 1024-1034.

[3] Alvarez, R., M. Camacho and G. Perez-Quiros (2012), Finite sample performance of small versus large scale dynamic factor models, WP 1204, Banco de España.

[4] Amengual, D. and M.W. Watson (2007), Consistent estimation of the number of dynamic factors in a large N and T panel, *Journal of Business and Economic Statistics*, 25(1), 91-96.

[5] Aruoba, S.B., F.X. Diebold, and C. Scotti (2009), Real-time measurement of business conditions, *Journal of Business & Economic Statistics* 27, 417-427.

[6] Bai, J. (2003), Inferential theory for factor models of large dimensions, *Econometrica*, 71(1), 135-171.

[7] Bai, J. and S. Ng (2002), Determining the number of factors in approximate factor models, *Econometrica*, 70(1), 191-221.

[8] Bai, J. and S. Ng (2006a), Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions, *Econometrica*, 74(4), 1133-1150.

[9] Bai, J. and S. Ng (2006b), Evaluating latent and observed factors in macroeconomics and finance, *Journal of Econometrics*, 131, 507-537.

[10] Bai, J. and S. Ng (2007), Determining the number of primitive shocks in factor models, *Journal of Business & Economic Statistics*, 25(1), 52-60.

[11] Bai, J. and S. Ng (2008a), Large dimensional factor analysis, *Foundations and Trends in Econometrics*, 3, 89-163.

[12] Bai, J. and S. Ng (2008b), Forecasting economic time series using targeted predictors, *Journal of Econometrics,* 146, 304-3.

[13] Bai, J. and P. Wang (2014), Identification theory for high dimensional static and dynamic factor and estimation models, *Journal of Econometrics,* 178(2), 794-804.

[14] Banbura, M. and M. Modugno (2014), Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data, *Journal of Applied Econometrics,* 29, 133-160.

[15] Banbura, M. and G. Runstler (2011), A look into the factor model blackbox: Publication lags and the role of hard and soft data in forecasting, *International Journal of Forecasting*, 27(2), 333-346.

[16] Banerjee, A., M. Marcellino and I. Masten (2014), Forecasting with factor-augmented error correction models, *International Journal of Forecasting,* 30(3), 589-612.

[17] Bernanke, B., J. Boivin and P. Eliasz (2005), Factor augmented vector autoregressions (FAVARs) and the analysis of monetary policy, *Quarterly Journal of Economics,* 120, 387-422.

[18] Boivin, J. and S. Ng (2006), Are more data always better for factor analysis?, *Journal of Econometrics*, 132, 169-194.

[19] Bork, L., H. Dewachter and R. Houssa (2009), Identification of macroeconomic factors in large panels, CREATES research paper 2009-43, School of Economics and Management, University of Aarhus.

[20] Bräuning, F. and S.J. Koopman (2014), Forecasting macroeconomic variables using collapsed dynamic factor analysis, *International Journal of Forecasting,* 30(3), 572-584.

[21] Breitung, J. and S. Eickemeir (2006), Dynamic Factor Models, in Hübler, O. and J. Frohn (eds.), *Modern Econometric Analysis*, Chapter 3, Springer.

[22] Breitung, J. and S. Eickemeir (2014), Analyzing business and financial cycles using multi-level factors, Discussion paper No. 11/2014, Deutsche Bundesbank.

[23] Breitung, J. and U. Pigorsch (2013), A canonical correlation approach for selecting the number of dynamic factors, *Oxford Bulletin of Economics and Statistics,* 75(1), 23-36.

[24] Buch, C.M., S. Eickmeier and E. Prieto (2014), Macroeconomic factors and microlevel bank behavior, *Journal of Money, Credit and Banking,* 46(4), 715-751.

[25] Caggiano, G., G. Kapetanios and V. Labhard (2011), Are more data always better for factor analysis? Results from the euro area, the six largest euro area countries and the UK, *Journal of Forecasting*, 30, 736-752.

[26] Camacho, M. and G. Perez-Quiros (2010), Introducing the Euro-STING: Short Term INdicator of Euro Area Growth, *Journal of Applied Econometrics*, 25(4), 663-694.

[27] Cattell, R. (1966), The scree test for the number of factors, *Multivariate Behavioral Research,* 1, 245-276.

[28] Chamberlain, G. and M. Rothschild (1983), Arbitrage, factor structure and mean-variance analysis in large asset markets, *Econometrica*, 51, 1305-1324.

[29] Choi, I. (2012). Efficient estimation of factor models, *Econometric Theory,* 28:274-308.

[30] Diebold, F.X. (2003), "Big data" dynamic factor models for macroeconomic measurement and forecasting (discussion of Reichlin and Watson papers), in Dewatripont, M., L.P. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics*, Cambridge University Press, Cambridge.

[31] Diebold, F.X. and M. Nerlove (1989), The dynamics of exchange rate volatility: a multivariate latent factor ARCH model, *Journal of Applied Econometrics,* 4(1), 1-21.

[32] Doz, C., D. Giannone and L. Reichlin (2011), A two-step estimator for large approximate dynamic factor models based on Kalman filtering, *Journal of Econometrics,* 164, 188-205.

[33] Doz, C., D. Giannone and L. Reichlin (2012), A Quasi Maximum Likelihood approach for large approximate dynamic factor models, *Review of Economics and Statistics*, 94(4), 1014-1024.

[34] Eickmeier, S., W. Lemke and M. Marcellino (in press), Classical time-varying FAVAR models - Estimation, forecasting and structural analysis, *Journal of the Royal Statistical Society, Series A.*

[35] Engle, R.F. and M.W. Watson (1981), A one-factor multivariate time series model of metropolitan wage rates, *Journal of the American Statistical Association,* 76, 774-781.

[36] Fiorentini, G., A. Galesi and E. Sentana (2014), A spectral EM algorithm for dynamic factor models, Manuscript.

[37] Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000), The generalized dynamic-factor model: identification and estimation, *Review of Economics and Statistics,* 82(4):540-554.

[38] Forni, M., M. Hallin, M. Lippi and L. Reichlin (2005), The generalized dynamic factor model: one sided estimation and forecasting, *Journal of the American Statistical Association*, 100, 830-840.

[39] Frale, C., G. Mazzi, M. Marcellino and T. Proietti (2011), EUROMIND: a monthly indicator of the euro area economic conditions, *Journal of the Royal Statistical Society*, 174, 439-470.

[40] Geweke, J. (1977), The dynamic factor analysis of economic time series, in D.J. Aigner and A.S. Goldberger (eds.), *Latent Variables in Socio-Economic Models,* North-Holland, Amsterdam.

[41] Giannone, D., L. Reichlin and D. Small (2008), Nowcasting: The real-time informational content of macroeconomic data, *Journal of Monetary Economics*, 55, 665-676.

[42] Hallin, M. and R. Liska (2007), Determining the number of factors in the general dynamic factor model, *Journal of the American Association,* 102, 603-617.

[43] Han, X. (in press), Tests for overidentifying restrictions in factor-augmented VAR models, *Journal of Econometrics.*

[44] Harvey, A.C. (1989), *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

[45] Harvey, A.C., E. Ruiz and N.G. Shephard (1994), Multivariate stochastic variance models, *The Review of Economic Studies*, 61(2), 247-264.

[46] Jungbaker, B. and S.J. Koopman (in press), Likelihood-based analysis for dynamic factor models, *Econometrics Journal.*

[47] Jungbacker, B., S.J. Koopman and M. van der Wel (2011), Maximum likelihood estimation for dynamic factor models with missing data, *Journal of Economic Dynamics & Control*, 35, 1358-1368.

[48] Kapetanios, G. (2010), A testing procedure for determining the number of factors in approximate factor models with large data sets, *Journal of Business & Economic Statistics*, 28, 397-409.

[49] Koopman, S.J. and M. van der Wel (2013), Forecasting the US term structure of interest rates using a macroeconomic smooth dynamic factor model, *International Journal of Forecasting,* 29, 676-694.

[50] Kose. M.A., C. Otrok and C.H. Whiteman (2003), International business cycles: world, region and country-specific factors, *American Economic Review,* 93, 1216-1239.

[51] Moench, E., S. Ng and S. Potter (2013), Dynamic hierarchical factor models, *Review of Economics & Statistics,* 95(5), 1811-1817.

[52] Onatski, A. (2009), Testing hypothesis about the number of factors in large factor models, *Econometrica*, 77(5), 1447-1479.

[53] Onatski, A. (2010), Determining the number of factors from empirical distribution of eigenvalues, *Review of Economics and Statistics*, 92(4), 1004-1016,

[54] Onatski, A. (2012), Asymptotics of the Principal Components estimator of large factor models with weakly influential factors, *Journal of Econometrics*, 168, 244-258.

[55] Pinheiro, M., A. Rua and F. Dias (2013), Dynamic factor models with jagged edge panel data: Taking on board the dynamics of the idiosyncratic components, *Oxford Bulletin of Economics and Statistics,* 75(1), 80-102.

[56] Poncela, P. and E. Ruiz (2015), More is not always better: back to the Kalman filter in dynamic factor models, in Koopman, S.J and N.G. Shephard (eds.), *Unobserved Components and Time Series Econometrics,* Oxford University Press.

[57] Reis, R. and M.W. Watson (2010), Relative good's prices, pure inflation, and the Phillips correlation, *American Economic Journal: Macroeconomics*, 2(3), 128-157.

[58] Rodríguez, A. and E. Ruiz (2009), Bootstrap prediction intervals in state space models, *Journal of Time Series Analysis,* 30(2), 167-178.

[59] Rodríguez, A. and E. Ruiz (2012), Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters, *Computational Statistics & Data Analysis,* 56(1), 62-74.

[60] Sargent, T.J. and C.A. Sims (1977), Business cycle modeling without pretending to have too much a priory economic theory, in C.A. Sims (ed.), *New Methods in Business Cycle Research*, Federal Reserve Bank of Minneapolis, Minneapolis.

[61] Stock, J.H. and M.W. Watson (1991), A probability model of the coincident economic indicators, in Lahiri, K. and G.H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge University Press, Cambridge.

[62] Stock J.H. and M.W. Watson (2002a), Forecasting using Principal Components from a large number of predictors, *Journal of the American Statistical Association*, 97(460), 1167-79.

[63] Stock J.H. and M.W. Watson (2006), Forecasting with many predictors, in Elliot, G., C. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting,* vol.1, Elsevier.

[64] Stock, J.H. and M.W. Watson (2011), Dynamic factor models, in Clements, M.P. and D.F. Hendry (eds.), *The Oxford Handbook of Economic Forecasting*, Oxford University Press, Oxford.

[65] Stock, J.H. and M. W. Watson (2012), Disentangling the channels of the 2007-09 recession, *Brooking Papers on Economic Activity*, Spring, 81-135.

[66] Watson, M.W. and R.F. Engle (1983), Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression, *Journal of Econometrics*, 23, 385-400.

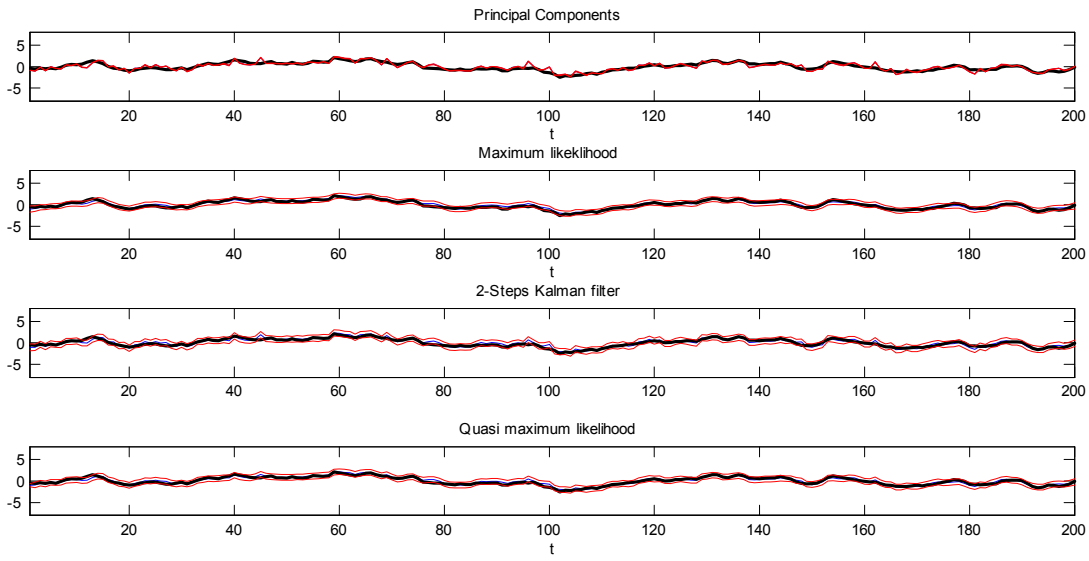

Figure 1: Small-data ($N = 12$ and $r = 1$) illustration of factor estimates (blue line) obtained using PC (top panel), KFS (second panel), 2SKF (third panel) and QML-EM (bottom panel) together with their corresponding 95% confidence intervals (red lines) and the true factor (black line).

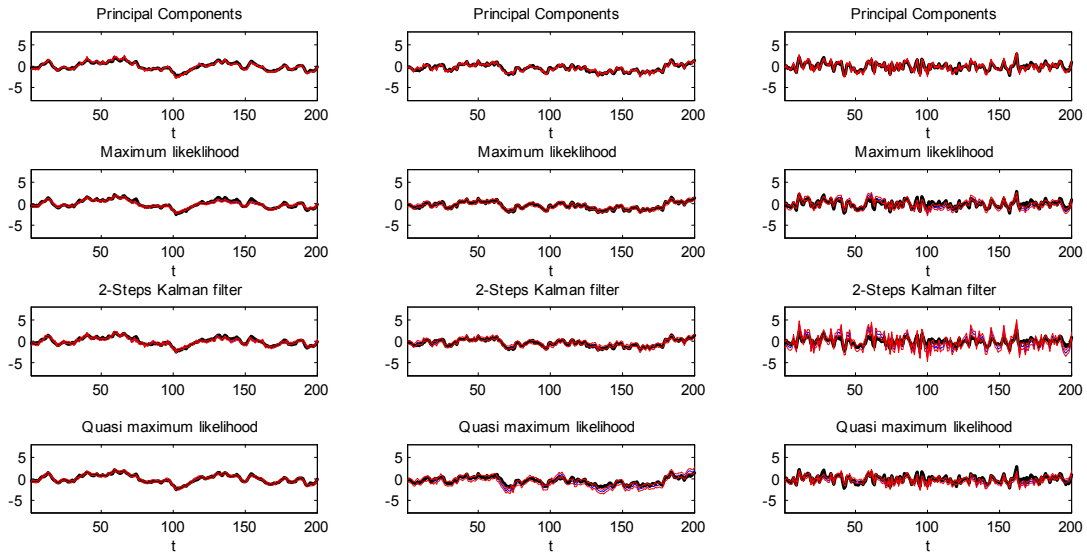

Figure 2: Big-data ($N = 120$ and $r = 3$) illustration of factor estimates (blue line) obtained using PC (top panel), KFS (second panel), 2SKF (third panel) and QML-EM (bottom panel) together with their corresponding 95% confidence intervals (red lines) and the true factor (black line).

N	12	12		12			30			120		
r	1	2		3			3			3		
Correlations												
PC	0.87	0.79	0.80	0.82	0.69	0.81	0.91	0.84	0.85	0.97	0.94	0.95
KFS	0.95	0.92	0.87	0.87	0.72	0.63	0.95	0.85	0.67	0.96	0.95	0.81
2SKF	0.92	0.81	0.81	0.70	0.52	0.65	0.90	0.73	0.80	0.94	0.94	0.87
QML	0.94	0.91	0.87	0.83	0.73	0.35	0.95	0.82	0.65	0.98	0.92	0.81
Coverages												
PC	46	44	40	45.5	29	37.5	46.5	33.5	35	46	40	32
KFS	95	92	85.5	83.5	91.5	86.5	93	87	76.5	58	82	55
2SKF	95.5	71	69.5	78.5	58.5	74	66.5	61	78	48	82	50.5
QML	96	81	81.5	93	75.5	64.5	90.5	76	64	77.5	69	51.5

Table 1. Comparison of procedures with a simulated system. Top panel reports the correlations between the true and estimated factors. The low panel reports percentage coverages of pointwise factor intervals when the nominal is 95%.

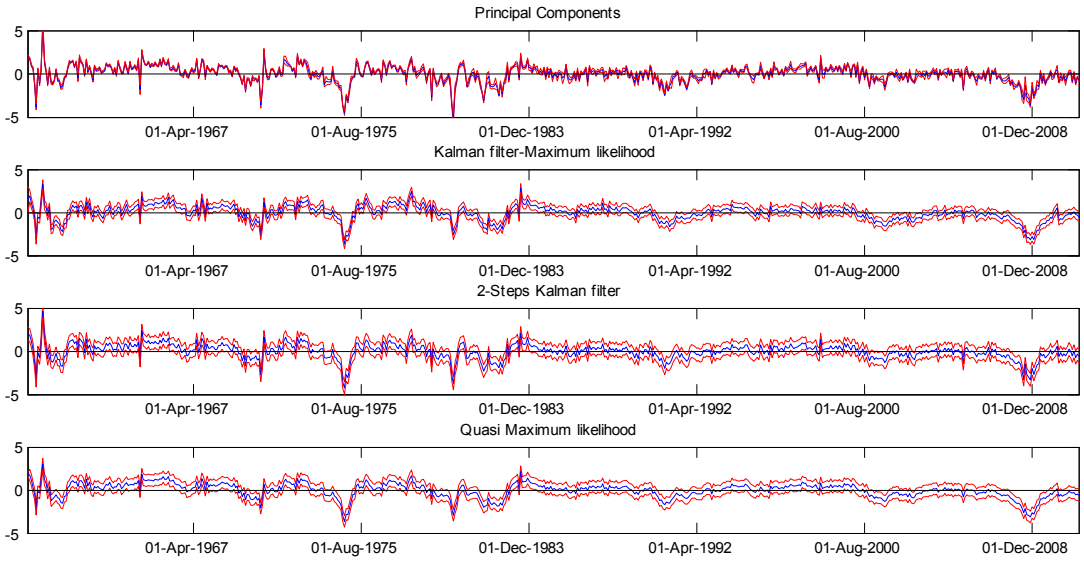

Figure 3: Factor extracted by each of the four procedures using 11 variables selected as the more correlated in average within its class together with the corresponding 95% intervals.

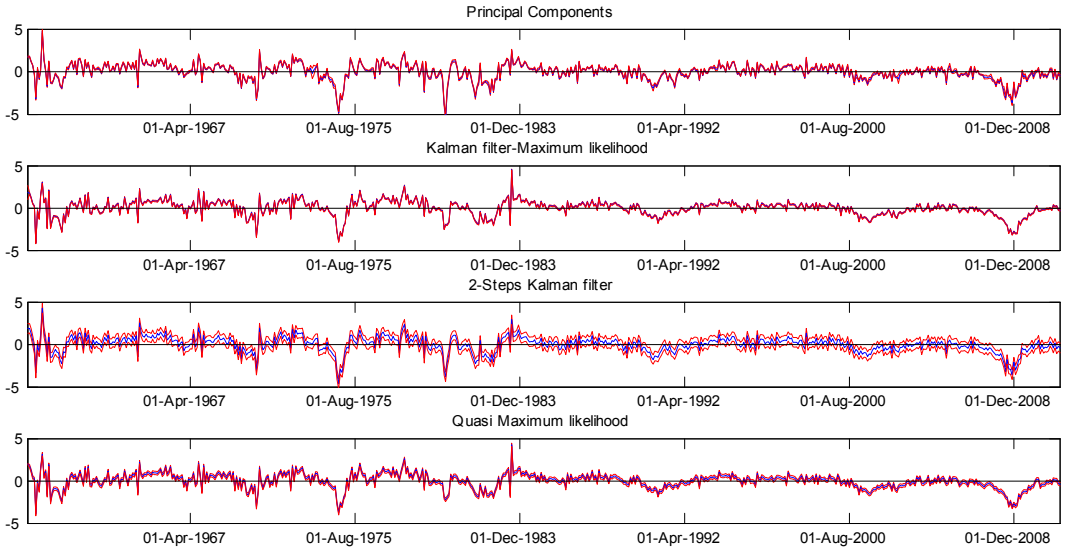

Figure 4: Factor extracted by each of the four procedures using 21 variables selected as the two more correlated in average within its class together with the corresponding 95% intervals. There is a unique factor selected when implementing the criteria proposed by Onatski (2009).

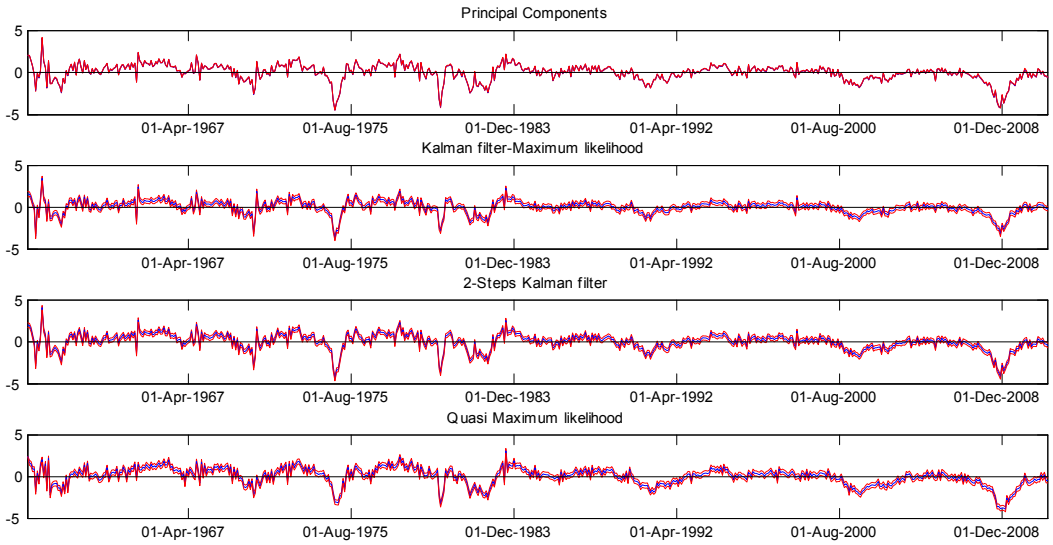

Figure 5: First factor extracted by each of the four procedures using 91 variables selected as the more correlated in average within its class together with the corresponding 95% intervals. Four factors are selected.

	N=12 r=1	N=12 r=2	N=12 r=3	N=30 r=2	N=30 r=3	N=120 r=3
Failures	0.4%	45%	82%	4.2%	20%	0.6%
Trace						
PC	0.77 (0.08)	0.68 (0.06)	0.63 (0.04)	0.81 (0.04)	0.78 (0.04)	0.93 (0.01)
KFS	0.91 (0.04)	0.86 (0.03)	0.77 (0.05)	0.91 (0.03)	0.85 (0.03)	0.92 (0.01)
2SKF	0.86 (0.06)	0.76 (0.06)	0.69 (0.05)	0.87 (0.04)	0.83 (0.04)	0.94 (0.01)
QML	0.90 (0.04)	0.85 (0.04)	0.76 (0.05)	0.91 (0.03)	0.86 (0.03)	0.94 (0.01)

Table 2. Percentage of failures of Onatski (2009) test to detect the true number of factors. Monte Carlo means and standard deviations (in parenthesis) of the trace statistic.

N	12	12		12			30		30			120		
r	1	2		3			2		3			3		
Coverages														
P	0.14	0.10	0.10	0.10	0.07	0.09	0.13	0.13	0.12	0.09	0.11	0.11	0.09	0.08
	(0.07)	(0.08)	(0.09)	(0.07)	(0.07)	(0.08)	(0.11)	(0.12)	(0.07)	(0.10)	(0.09)	(0.07)	(0.11)	(0.09)
K	0.94	0.75	0.74	0.69	0.75	0.71	0.58	0.59	0.65	0.70	0.56	0.39	0.42	0.32
	(0.03)	(0.22)	(0.21)	(0.20)	(0.18)	(0.19)	(0.25)	(0.23)	(0.21)	(0.18)	(0.15)	(0.13)	(0.19)	(0.15)
S	0.96	0.56	0.62	0.62	0.68	0.64	0.44	0.48	0.54	0.60	0.58	0.40	0.42	0.33
	(0.02)	(0.16)	(0.13)	(0.13)	(0.14)	(0.09)	(0.15)	(0.11)	(0.13)	(0.16)	(0.12)	(0.12)	(0.19)	(0.12)
Q	0.97	0.64	0.68	0.70	0.79	0.69	0.48	0.50	0.64	0.68	0.54	0.52	0.43	0.33
	(0.01)	(0.22)	(0.20)	(0.18)	(0.15)	(0.13)	(0.20)	(0.17)	(0.22)	(0.19)	(0.15)	(0.21)	(0.20)	(0.14)

Table 3. Percentage coverages of pointwise factor intervals when the nominal is 95% for the succeed simulations reported in Table 2. P stands for PC estimation, K for Kalman filter ML, S for 2SKF and Q for QML.

	PC				KFS				2SKF				QML			
N	11	21	91	103	11	21	91	103	11	21	91	103	11	21	91	103
11	0.11	0.96	0.89	0.89	0.27	0.95	0.96	0.96	0.38	0.98	0.94	0.94	0.34	0.91	0.88	0.85
21		0.05	0.94	0.94		0.04	0.96	0.95		0.28	0.98	0.98		0.10	0.93	0.89
91			0.02	1			0.13	0.97			0.13	1.			0.15	0.99
103				0.02				0.15				0.13				0.18

Table 4. Empirical application of the Stock and Watson (2012) data base. Main diagonal: RMSE of extracted factors. The KFS, 2SKF and QML are computed using the steady-state RMSE obtained from the Kalman filter with estimated parameters. The PC RMSE are obtained using the asymptotic approximation and averaging over time. Off-diagonal elements: correlations between the factors estimated using alternative number of variables.

	$N = 11$			$N = 21$			$N = 91$			$N = 103$		
	KFS	2SKF	QML	KFS	2SKF	QML	KFS	2SKF	QML	KFS	2SKF	QML
PC	0.84	0.97	0.91	0.85	0.98	0.86	0.97	0.99	0.92	0.94	0.99	0.88
KFS		0.92	0.95		0.92	1		0.98	0.95		0.96	0.97
2SKF			0.96			0.93			0.93			0.89

Table 5. Empirical application of the Stock and Watson (2012) data base. Correlations between the factor estimated by alternative procedures given the number of variables in the system.