

An Introduction to Factor Modelling

Joint Vienna Institute / IMF ICD
Macro-econometric Forecasting and Analysis
JV16.12, L05, Vienna, Austria, May 19, 2016

Presenters
Massimiliano Marcellino
(Bocconi University)
Sam Ouliaris

This training material is the property of the International Monetary Fund (IMF) and is intended for use in IMF Institute courses. Any reuse requires the permission of the IMF Institute.

Why factor models?

- Factor models decompose the behaviour of an economic variable (x_{it}) into a component driven by few unobservable factors (f_t), common to all the variables but with specific effects on them (λ_i), and a variable specific idiosyncratic components (ξ_{it}):

$$x_{it} = \lambda_i f_t + \xi_{it},$$

$$t = 1, \dots, T; \quad i = 1, \dots, N$$

- Idea of few common forces driving all economic variables is appealing from an economic point of view, e.g. in the Real Business Cycle (RBC) and Dynamic Stochastic General Equilibrium (DSGE) literature there are just a few key economic shocks affecting all variables (productivity, demand, supply, etc.), with additional variable specific shocks
- Moreover, factor models can handle large datasets (N large), reflecting the use of large information sets by

Why factor models?

From an econometric point of view, factor models:

- Alleviate the curse of dimensionality of standard VARs (number of parameters growing with the square of the number of variables)
- Prevent omitted variable bias and issues of non-fundamentality of shocks (shocks depending on future rather than past information that cannot be properly recovered from VARs)
- Provide some robustness in the presence of structural breaks
- Require minimal conditions on the errors (can be correlated over time, heteroskedastic etc)
- Are relatively easy to be implemented (though underlying model is nonlinear and with unobservable variables)

What can be done with factor models?

- Use the estimated factors to summarize the information in a large set of indicators. For example, construct coincident and leading indicators as the common factors extracted from a set of coincident and leading variables, or in the same way construct financial condition indexes or measures of global inflation or growth.
- Use the estimated factors for nowcasting and forecasting, possibly in combination with autoregressive (AR) terms and/or other selected variables, or for estimation of missing or outlying observations (getting a balanced dataset from an unbalanced one). Typically, they work rather well.
- Identify the structural shocks driving the factors and their dynamic impact on a large set of economic and financial indicators (impulse response functions and forecast error variance decompositions, as in

An introduction to factor models

In this lecture we will consider:

- Small scale factor models: representation, estimation and issues
- Large scale factor models
 - Representation (exact/approximate, static/dynamic, parametric / non parametric)
 - Estimation: principal components, dynamic principal components, maximum likelihood via Kalman filter, subspace algorithms
 - Selection of the number of factors (informal methods and information criteria)
 - Forecasting (direct / iterated)
 - Structural analysis (FAVAR based)
- Useful references (surveys): Bai and Ng (2008), Stock and Watson (2002, 2011, 2015), Faust et al. (2011)

Some extensions

In the next lecture we will consider some relevant extensions for empirical applications:

- How to allow for parameter time variation
- How to handle $I(1)$ variables: Factor augmented Error Correction Models
- How to handle hierarchical structures (e.g., countries/regions/sectors)
- How to handle nonlinearities
- How to construct targeted factors
- How to handle unbalanced datasets: missing observations, mixed frequencies and ragged edges

Representation

Let us consider the factor model:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \\ \dots \\ x_{Nt} \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1r} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2r} \\ \dots & & \dots & \\ \lambda_{N1} & \lambda_{N2} & \dots & \lambda_{Nr} \end{pmatrix} \begin{pmatrix} f_{1t} \\ f_{2t} \\ \dots \\ f_{rt} \end{pmatrix} + \begin{pmatrix} \xi_{1t} \\ \xi_{2t} \\ \dots \\ \xi_{Nt} \end{pmatrix},$$

$$\begin{pmatrix} f_{1t} \\ f_{2t} \\ \dots \\ f_{rt} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \dots & & \dots & \\ a_{r1} & a_{r2} & \dots & a_{rr} \end{pmatrix} \begin{pmatrix} f_{1t-1} \\ f_{2t-1} \\ \dots \\ f_{rt-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \\ \dots \\ u_{rt} \end{pmatrix}$$

where each (weakly stationary and standardized) variable x_{it}

, $i = 1, \dots, N$, depends on r unobservable factors f_{jt} via the loadings λ_{ij} , $j = 1, \dots, r$, and on its own idiosyncratic error, ξ_{it} . In turn, the factors are generated from a VAR(1) model, so that each factor f_{it} depends on the first lag of all the factors,

Representation

For example, x_{it} , $i = 1, \dots, N$, $t = 1, \dots, T$ can be:

- A set of macroeconomic and/or financial indicators for a country → the factors represent their common drivers
- GDP growth or inflation for a large set of countries the factors capture global movements in these two variables
- All the subcomponents of a price index → the factors capture the extent of commonality among them and can be compared with the aggregate index
- A set of interest rates of different maturities → commonality is driven by level, slope and curvature factors

In general, we are assuming that all the variables are driven by a (small) set of common unobservable factors plus



Let us write the factor model more compactly as:

$$\begin{aligned} X_t &= \Lambda f_t + \xi_t, \\ f_t &= A f_{t-1} + u_t, \end{aligned}$$

where:

- $X_t = (x_{1t}, \dots, x_{Nt})'$ is the $N \times 1$ vector of stationary variables under analysis
- $f_t = (f_{1t}, \dots, f_{rt})'$ is the $r \times 1$ vector of unobservable factors
- $\Lambda = (\lambda'_1, \dots, \lambda'_N)'$ is the $N \times r$ matrix of loadings with (measure effects of factors on variables)
- $\lambda_i = (\lambda_{i1}, \dots, \lambda_{ir})$ is the $N \times r$ matrix of loadings with (measure effects of factors on variables)
- $\xi_t = (\xi_{1t}, \dots, \xi_{Nt})'$ is the $N \times 1$ vector of idiosyncratic shocks
- u_t is the $r \times 1$ vector of shocks to the factors
- ξ_t and u_t are multivariate, mutually uncorrelated, standard orthogonal white noise sequences (hence, uncorrelated over time and with constant variance covariance matrix);
- $|\lambda_{\max}(A)| < 1, |\lambda_{\min}(A)| > 0$ (factors are stationary and

In the factor model:

$$\begin{aligned} X_t &= \Lambda f_t + \xi_t, \\ f_t &= A f_{t-1} + u_t, \end{aligned}$$

- Λf_t is called the common component, and $\lambda_i f_t$ is the common component for each variable i .
- ξ_t is called the idiosyncratic component, and ξ_{it} is the idiosyncratic component for each variable i .
- As f_t has only a contemporaneous effect on X_t , this is a static factor model.

- Additional lags of f_t in the X_t equations can be easily allowed, and we obtain a dynamic factor model. Additional lags in the f_t equations can be also easily allowed, as well as deterministic components.
- If the variance covariance matrix of ξ_t is diagonal (no correlation at all among the idiosyncratic components), we have a strict factor model. Otherwise, an approximate factor model.
- As we have specified a model for the factors (VAR(1)), and made specific assumption on the error structure (multivariate white noise), we have a parametric factor model.

Let us consider an even more compact formulation of the factor model:

$$X = \Lambda F + \xi$$

where:

- $X = (X_1, \dots, X_T)$ is the $N \times T$ matrix of stationary variables under analysis
- $F = (f_1, \dots, f_T)$ is the $r \times T$ matrix of unobservable factors
- $\Lambda = (\lambda_1', \dots, \lambda_N)'$ is the $N \times r$ matrix of loadings, as before
- $\xi = (\xi_1, \dots, \xi_T)$ is the $N \times T$ matrix of idiosyncratic shocks

Identification

- Let us now consider two factor models:

$$X = \Lambda F + \xi, \text{ and}$$

$$X = \Lambda P^{-1} P F + \xi = \Theta G + \xi$$

where P is an $r \times r$ invertible matrix, $\Theta = \Lambda P^{-1}$ and $G = P F$.

- The two models for X are observationally equivalent (same likelihood), hence to uniquely identify the factors and the loadings we need to impose a priori restrictions on Λ and/or F .
- This is similar to the error correction model where the cointegrating vectors and/or their loadings are properly restricted to achieve identification.

- Typical restrictions are either $\Lambda = (I : \tilde{\Lambda})$ where I_r is the r -dimensional identity matrix and $\tilde{\Lambda}$ is the $(N-r) \times r$ matrix of restricted loadings, or $FF' = I_r$. The latter condition imposes that the factors are orthogonal and with unit variance, as

$$FF' = \begin{pmatrix} \sum_{t=1}^T f_{1t}^2 & \sum_{t=1}^T f_{1t}f_{2t} & \cdots & \sum_{t=1}^T f_{1t}f_{rt} \\ \sum_{t=1}^T f_{2t}f_{1t} & \sum_{t=1}^T f_{2t}^2 & \cdots & \sum_{t=1}^T f_{2t}f_{rt} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=1}^T f_{rt}f_{1t} & \sum_{t=1}^T f_{rt}f_{2t} & \cdots & \sum_{t=1}^T f_{rt}^2 \end{pmatrix}$$

- The condition $FF' = I_r$ is sufficient to get unique estimators for the factors, but not to fully identify the model. For that additional conditions are needed, such as $\Lambda^t \Lambda$ is diagonal with distinct, decreasing diagonal elements. See, e.g., Lutkepohl (2014) for details.

Factor models and VARs

An interesting question:

- Is there a VAR that is equivalent to a factor model (in the sense of having the same likelihood)?

Unfortunately, in general no, at least not a finite order VAR. However, it is possible to impose restrictions on a VAR to make it "similar" to a factor model.

Let us consider the VAR(1) model

$$X_t = BX_{t-1} + \xi_t,$$

assume that the $N \times N$ matrix B can be factored into $B = CD$, where C and D are $N \times r$ and $r \times N$ matrices respectively, and define $g_t = DX_t$. We get:

$$\begin{aligned} X_t &= Cg_{t-1} + \xi_t, \\ g_t &= Qg_{t-1} + v_t, \end{aligned}$$

where $Q = DB$ and $v_t = D\xi_t$.

This is called a Multivariate Autoregressive Index (MAI) model, and g_t plays a similar role as f_t in the factor model, but it is observable (a linear combination of the variables in X_t) and can only affect X_t with a lag. Moreover, estimation of the MAI is complex, as the model is nonlinear (see Carriero, Kapetanios and Marcellino (2011, 2015)). Hence, let us

Estimation by the Kalman filter

Let us consider again the factor model written as:

$$\begin{aligned} X_t &= \Lambda f_t + \xi_t, \\ f_t &= A f_{t-1} + u_t. \end{aligned}$$

In this formulation:

- the factors are unobservable states,
- $X_t = \Lambda f_t + \xi_t$ are the observation equations (linking the unobservable states to the observable variables),
- $f_t = A f_{t-1} + u_t$ are the transition equations (governing the evolution of the states).

- Hence, the model:

$$\begin{aligned} X_t &= \Lambda f_t + \xi_t, \\ f_t &= A f_{t-1} + u_t. \end{aligned}$$

is already in state space form, and therefore we can use the Kalman Filter to obtain maximum likelihood estimators for the factors, the loadings, the dynamics of the factors, and the variance covariance matrices of the errors (e.g., Stock and Watson (1989)).

However, there are a few problems:

- First, the method is computationally demanding, so that it is traditionally considered applicable only when the number of variables, N , is small.
- Second, with N finite, we cannot get consistent estimators for the factors (as the latter are random variables, not parameters).
- Finally, the approach requires to specify a model for the factors, which can be difficult as the latter are not observable. Hence, let us consider alternative estimation approaches.

Non-parametric, large N , factor models

- There are two competing approaches in the factor literature that are non-parametric, allow for very large N (in theory $N \rightarrow \infty$) and produce consistent estimators for the factors and/or the common components. They were introduced by Stock and Watson (2002a, 2002b, SW) and Forni, Hallin, Lippi and Reichlin (2000, FHLR), and later refined and extended in many other contributions, see e.g. Bai and Ng (2008) for an overview.
- We will now review their main features and results, starting with SW (which is simpler) and then moving to FHLR.

The SW approach - PCA

- The Stock and Watson (2002a,2002b) factor model is

$$X_t = \Lambda f_t + \xi_t,$$

where:

- X_t is $N \times 1$ vector of stationary variables
- f_t is $r \times 1$ vector of common factors, can be correlated over time
- Λ is $N \times r$ matrix of loadings
- ξ_t is $N \times 1$ vector of idiosyncratic disturbances, can be mildly cross-sectionally and temporally correlated
- conditions on Λ and ξ_t guarantee that the factors are pervasive (affect most variables) while idiosyncratic errors are not.

The SW approach - PCA

- Estimation of Λ and f_t in the model $X_t = \Lambda f_t + \xi_t$ is complex because of nonlinearity (Λf_t) and the fact that f_t is a random variable rather than a parameter.
- The minimization problem we want to solve is

$$\min_{\Lambda, f_1, f_2, \dots, f_T} \left(\sum_{t=1}^T \|X_t - \Lambda f_t\|^2 \right)$$

- Under mild regularity conditions, it can be shown that the (space spanned by the) factors can be consistently estimated by the first r static principal components of X (**PCA**).

The SW approach - Choice of r

Choice of the number of factors, r :

- Fraction of explained variance of X_t : should be large (though decreasing) for the first r principal components, very small for the remaining ones
- Information criteria (Bai and Ng (2002): r should minimize properly defined information criteria (cannot use standard ones as now not only T but also N can diverge)
- Testing: Kapetanios (2010) provides some statistics and related distributions, not easy

The SW approach - Properties of PCA

- Need both N and T to grow large, and not too much cross-correlation among idiosyncratic errors.
- As a basic example, consider case with one factor and uncorrelated idiosyncratic errors (exact factor model):

$$x_{it} = \lambda_i f_t + e_{it}. \quad (1)$$

Then, use simple cross-sectional average as factor estimator:

$$\frac{1}{N} \sum_{i=1}^N x_{it} = \bar{x}_t = \left(\frac{1}{N} \sum_{i=1}^N \lambda_i \right) f_t + \frac{1}{N} \sum_{i=1}^N e_{it}$$

$$\lim_{N \rightarrow \infty} \bar{x}_t = \bar{\lambda} f_t$$

And \bar{x}_t is consistent for $\bar{\lambda} f_t$ (up to a scalar). We can also get factor loadings by OLS regression of x_{it} on \bar{x}_t and

$$\lim_{T \rightarrow \infty} \hat{\lambda}_i = \frac{\lambda_i}{\bar{\lambda}}$$

So, if both N and T diverge $\hat{\lambda}_i \bar{x}_t \rightarrow \lambda_i f_t$.

The SW approach - Properties of PCA

- PCA are weighted rather than simple averages of the variables, where weights depend on λ_i and $\text{var}(e_{it})$.
- Under general conditions and with proper standardization, PCA and estimated loadings have asymptotic Normal distributions (Bai Ng (2006))
- If N grows faster than T (such that $T^{1/2}/N$ goes to zero), the estimated factors can be treated as true factors when used in second-step regressions (e.g. for forecasting, factor augmented VARs, etc.). Namely, there are no generated regressor problems.
- If the factor structure is weak (first factor explains little percentage of overall variance), PCA is no longer consistent (Onatski (2006)).

The SW approach - Properties of PCA based forecasts

Suppose the model is

$$\begin{aligned} Y_{t+1} &= f_t \beta + \\ X_t' v_t &= \Lambda f_t + \xi_t, \end{aligned}$$

then we can construct a forecast as

$$\hat{y}_{t+1} = \hat{f}_t' \hat{\beta},$$

where \hat{f}_t are the PCA factor estimators and $\hat{\beta}$ the OLS estimator of β , obtained by regressing y_{t+1} on

- The asymptotic distribution of factor based forecasts is also Normal, under general conditions, and its variance depends on the variance of the loadings and on that of the factors, so you need both n and T large to get a precise forecast (Bai and Ng (2006)). This results can be used to derive interval and density factor based forecasts.

The FHLR approach - DPCA

- The FHLR factor model is

$$X_t = B(L)u_t + \xi_t = x_t + \xi_t,$$

where:

- X_t is the $N \times 1$ vector of stationary variables
- u_t is the $q \times 1$ vector of i.i.d. orthonormal common shocks. These are the drivers of the common factors in the SW formulation, but in FHLR the focus is on the common shocks rather than the common factors)
- $B(L) = 1 + B_1L + B_2L^2 + \dots + B_pL^p$
- $x_t = B(L)u_t$ is the $N \times 1$ vector of common components. It is estimated by Dynamic Principal Components (DPCA), details in Appendix A.
- ξ_t is the $N \times 1$ vector of idiosyncratic shocks, can be mildly correlated across units and over time
- Conditions on $B(L)$ and ξ_t guarantee that the factors are pervasive (affect most variables) while idiosyncratic errors

The FHLR approach - static and dynamic factors

- q can be different from r : the former is usually referred to as the number of dynamic factors while r is the number of static factors, with $q \leq r$.
- Let us assume for simplicity that there is a single factor f_t , but it has both a contemporaneous and lagged effect on X_t :

$$X_t = \Lambda_1 f_t + \Lambda_2 f_{t-1} + \xi_t,$$

We can define $g_t = (f_t, f_{t-1})'$ and write the model in static form as

$$X_t = \Lambda g_t + \xi_t.$$

In this case we have $r = 2$ static factors (those in g_t), which are all driven by $q = 1$ common shock (u_t). Typically, FHLR focus on q (and the common shocks u_t), while SW on r (and the common factors g_t). The distinction matters more for structural analysis than for

The FHLR approach - Choice of q

- Informal methods:

- Estimate recursively the spectral density matrix of a subset of

- the dynamic eigenvalues for a grid of frequencies, increasing the number of variables at each step; choose when the number of variables increases the average

- over frequencies of the first q dynamic eigenvalues diverges, while the average of the $q + 1^{th}$ does not.

- For the whole X_t there should be a big gap between the

- variance of X_t explained by the first q dynamic principal components and that explained by the $q + 1^{th}$ component.

- Formal methods:

- Information criteria: Hallin Liska (2007); Amengual

The FHLR approach - Forecasting

- Consider now the model (direct estimation, the common shocks have an h-period delay in effecting X_t):

In this context, an optimal linear forecast for X_{t+h} is $\hat{\chi}_t$ that can be obtained, as said, by

- A problem with using this method for forecasting is the use of future information in the computation of the DPCA. To overcome this issue, which prevents a real time implementation of the procedure, Forni, Hallin, Lippi and Reichlin (2005) propose a modified one-sided estimator (which is however too complex for implementation in EViews).

Parametric estimation - quasi MLE

- Kalman filter produces (quasi-) ML estimators of the factors, but considered not feasible for large N . No longer true: Doz, Giannone, Reichlin (2011, 2012).
- Model has the form

$$X_t = \Lambda f_t + \xi_t, \quad (2)$$

$$\Psi(L)f_t = B \eta_t, \quad (3)$$

where q -dimensional vector η_t contains the orthogonal dynamic shocks driving the r factors f_t , and the matrix B is $(r \times q)$ -dimensional, with $q \leq r$.

- For given r and q , estimation proceeds in the following steps:

Parametric estimation - quasi MLE

1. Estimate \hat{f}_t by PCA and $\hat{\Lambda}$ by regressing X_t on \hat{F}_t . The covariance of $\hat{\zeta}_t = X_t - \hat{\Lambda}\hat{F}_t$, denoted as $\hat{\Sigma}_{\zeta}$, is also estimated.
2. Estimate a VAR(p) on the factors \hat{f}_t , yielding $\hat{\Psi}(L)$ and the residual covariance of $\hat{\zeta}_t = \hat{\Psi}(L)\hat{F}_t$, denoted as $\hat{\Sigma}_{\zeta}$.
3. To estimate B , given the number of dynamic shocks q , apply an eigenvalue decomposition of $\hat{\Sigma}_{\zeta}$. Let M be the $(r \times q)$ matrix of the eigenvectors corresponding to the q largest eigenvalues, and let the $(q \times q)$ -dimensional matrix P contain the largest eigenvalues on the main diagonal and zero otherwise. Then, $\hat{B} = M \times P^{-1/2}$.
4. The Kalman filter (or smoother) then yields new estimates of the factors, and the procedure can be iterated.
5. Forecasts can be obtained either by the Kalman filter or as

$$\hat{X}_{T+h} = \hat{\Lambda}\hat{f}_{T+h},$$

where \hat{f}_{T+h} are obtained from the VAR in (3).

Parametric estimation - Subspace algorithms (SSS)

- Let us now consider again the factor model:

$$\begin{aligned} X_t &= Cf_t + Du_t, \quad t = 1, \dots \\ \cdot f_t^T & \quad Af_{t-1} + Bu_{t-1} \end{aligned} \quad (4)$$

Kapetanios and Marcellino (2009, KM) show that (4) can be written as regression of future on past, with particular reduced rank restrictions on the coefficients (similar to reduced rank VAR seen above):

$$X_t^f = OKX_t^p + EE_t^f \quad (5)$$

Where

$$X_t^f = (X_t', X_{t+1}', X_{t+2}', \dots)', \quad X_t^p = (X_{t-1}', X_{t-2}', \dots),$$

$$E_t^f = (u_t', u_{t+1}', \dots)'$$

- Note that (i) $X_t^f = OKX_t^p + EE_t^f$ and (ii) $\hat{f}_t = \hat{K}X_t^p$. Hence, best linear predictor of future X is OKX_t^p , and we need and estimator for K (and for the loadings $X_t^f = OKX_t^p$).

Parametric estimation - SSS

- KM show how to obtain the SSS factor estimates $\hat{\xi}_i = \hat{K} X_i^p$. See Appendix A for details.
- Once estimates of the factors are available, estimates of the other parameters (including the factor loadings, $\hat{\alpha}$) can be obtained by OLS.
- Choice of number of factors can be done by information criteria, similar to those by Bai and Ng (2002) for PCA but with different penalty function, see KM.

Parametric estimation - SSS forecasts

- The SSS forecasts are $X_t^f = \hat{O} \hat{K} X_t^p$, where \hat{O} is OLS regression on the estimated factors, as in
- **PCA**. MLE forecasts are obtained by iterated method (VAR for factors is iterated forward to produce forecasts for the factors, which are then inserted into the static model for X_t). Forecasts obtained by PCA, DPCA and SSS use direct method (variable of interest is regressed on the estimated factors lagged h periods, and parameter estimates are combined with current value of the estimated factors to produce h-step ahead forecast of variable(s) of interest).
- If model is correctly specified, MLE plus iterated method produces better (more efficient) forecasts. If there is mis-specification, as it is often the case, the ranking is not clear-cut, other factor estimation approaches plus direct estimation can be better. See, e.g., Marcellino, Stock and

Factor estimation methods - Monte Carlo

Comparison

- Comparison of PCA, DPCA, MLE and SSS (based on Kapetanios and Marcellino (2009, KM)).
- The DGP is:

$$x_t = Cf_t + \varepsilon_t, \quad t = 1, \dots, T$$

$$A(L)f_t = B(L)u_t \tag{6}$$

Where

$B(L) = I + B_1(L) + \dots + B_q(L)$, with $(N, T) = (50, 50)$, $(50, 100)$, $(100, 50)$, $(100, 100)$, $(50, 500)$, $(100, 500)$ and $(200, 50)$. MLE for $(50, 50)$ only, due to computational burden.

- Experiments differ for number of factors (one or several), A and B matrices, choice of s ($s = m$ or $s = 1$), factor loadings (static or dynamic), choice of number of factors (true number or misspecified), properties of idiosyncratic errors (uncorrelated or serially correlated), and the way C matrix is generated (standard normal or uniform with

Factor estimation methods - MC Comparison, summary

- Appendix B provides more details on the DGP and detailed results. The main findings are the following:
 - DPCA shows consistently lower correlation between true and estimated common components than SSS and PCA. It shows, in general, more evidence of serial correlation of idiosyncratic components, although not to any significant extent.
 - SSS beats PCA, but gains are rather small, in the range 5-10%, and require a careful choice of s .
 - SSS beats MLE, which is only slightly better than PCA.
 - All methods perform very well in recovering the common components. As PCA is simpler, it seems reasonable to use it.

Factor models - Forecasting performance

- Really many papers on forecasting with factor models in the past 15 years, starting with Stock and Watson (2002b) for the USA and Marcellino, Stock and Watson (2003) for the euro area. Banerjee, Marcellino and Masten (2006) provide results for ten Eastern European countries. Eickmeier and Ziegler (2008) provide nice summary (meta-analysis), see also Stock and Watson (2006) for a survey of the earlier results.
- Recently used also for nowcasting, i.e., predicting current economic conditions (before official data is released). More on this in the next lecture.

Factor models - Forecasting performance

Eickmeier and Ziegler (2008):

- "Our results suggest that factor models tend to outperform small models, whereas factor forecasts are slightly worse than pooled forecasts. Factor models deliver better predictions for US variables than for UK variables, for US output than for euro-area output and for euro-area inflation than for US inflation. The size of the dataset from which factors are extracted positively affects the relative factor forecast performance, whereas pre-selecting the variables included in the dataset did not improve factor forecasts in the past. Finally, the factor estimation technique may matter as well."

Structural Factor Augmented VAR (FAVAR)

- To illustrate the use of the FAVAR for structural analysis, we take as starting point the FAVAR model as proposed by Bernanke, Boivin and Eliasch (2005, BBE), see also Eickmeier, Lemke and Marcellino (2015, ELM) for extensions and Lutkepohl (2014), Stock and Watson (2015) for surveys.

- The model for a large set of stationary macroeconomic and financial variables is:

$$x_{i,t} = \lambda_i F_t + e_{i,t}, \quad i = 1, \dots, N \quad (7)$$

where the factors are orthonormal ($F'F = I$) and uncorrelated with the idiosyncratic errors, and $E(e_t) = 0$, $E(e_t e_t') = R$, where R is a diagonal matrix. As we have seen, these assumptions identify the model and are common in the FAVAR literature.

- The dynamics of the factors are then modeled as a

$$F_t = A_1 F_{t-1} + \dots + A_p F_{t-p} + w_t, \quad E(w_t) = 0, E(w_t w_t') = W. \quad (8)$$

Structural FAVAR

- The VAR equations in (8) can be interpreted as a reduced-form representation of a system of the

$$PF_t = K_1 F_{t-1} + \dots + K_p F_{t-p} + u_t, \quad E(u_t) = 0, \quad E(u_t u_t') = S, \quad (9)$$

where P is lower-triangular with ones on the main diagonal, and S is a diagonal matrix.

- The relation to the reduced-form parameters in (8) is $B_i = P^{-1} K_i$ and $W = P^{-1} S P^{-1}$. This system of equations is often referred to as a 'structural VAR' (SVAR) representation, obtained with Choleski identification.
- For the structural analysis, BBE assume that X_t is driven by G latent factors F_t^* and the Federal Funds rate (i_t) as a $(G + 1)$ th observable factor, as they are interested in measuring the effects of monetary policy shocks in the economy. ELM use $G = 5$

Structural FAVAR - Monetary policy shock identification

- The space spanned by the factors can be estimated by PCA using, as we have seen, the first $G + 1$ PCs of the data X_t (BBE also consider other factor estimation methods).
- To remove the observable factor i_t from the space spanned by all $G + 1$ factors, dataset is split into slow-moving variables (expected to move with delay after an interest rate shock), and fast-moving variables (can move instantaneously).
Slow-moving variables comprise, e.g., real activity measures, consumer and producer prices, deflators of GDP and its components and wages, whereas fast-moving variables are financial variables such as asset prices, interest rates or commodity prices.

Structural FAVAR - Monetary policy shock identification

- In line with BBE, ELM estimate the first G PCs from the set of slow-moving variables, denoted by \hat{F}_t^{slow} .

- Then, they carry out a multiple regression of F_t on \hat{F}_t^{slow} and i_t ,
i.e.

$$F_t = a \hat{F}_t^{slow} + b i_t + v_t.$$

- An estimate of F_t^* is then given by $\hat{a} \hat{F}_t^{slow}$.

Structural FAVAR - Monetary policy shock identification

- In the joint factor vector $F_t \equiv [\hat{F}_t^*, i_t]$ the Federal Funds rate i_t is ordered last. Given this ordering, the VAR representation with lower-triangular contemporaneous-relation matrix P in (8) directly identifies the monetary policy shock as the last element of the innovation vector u_t , say $u_{int,t}$. Hence, the shock identification works via a Cholesky decomposition, which is here readily given by the lower triangular P^{-1} .

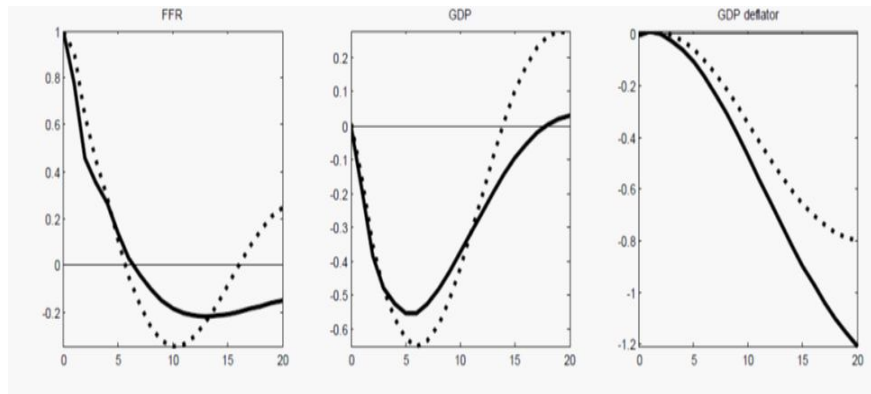
- Naturally, the methodology also allows for other identification approaches, such as short/long run or sign restrictions. These can be just applied to the VAR for

- Impulse responses of the factors to the monetary policy shock, estimated loading equations $x_{i,t} = \Lambda_i' F_t + e_{i,t}$,

To get $\partial x_{i,t+h} / \partial u_{int,t}$ the Proper confidence bands for the impulse response functions can be computed by using the bootstrap method.

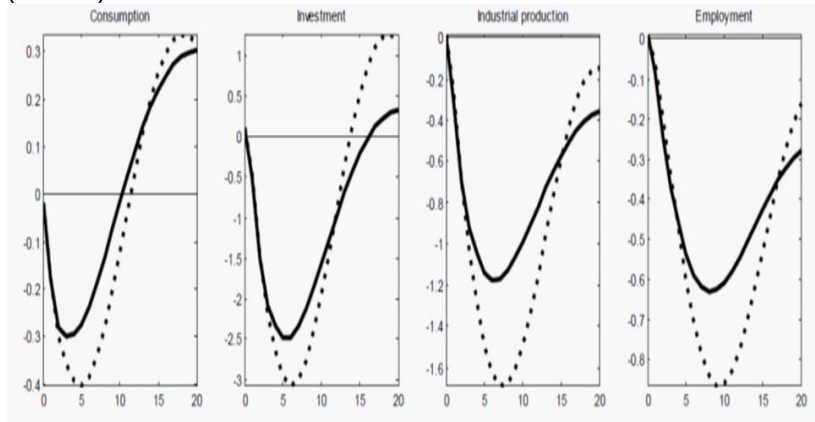
Structural FAVAR - Monetary policy (FFR) shock

Impulse responses from constant parameter FAVAR (solid) and time varying FAVAR (averages over all periods, dotted) for key variables, taken from ELM (who developed the TV-FAVAR, discussed in next lecture)



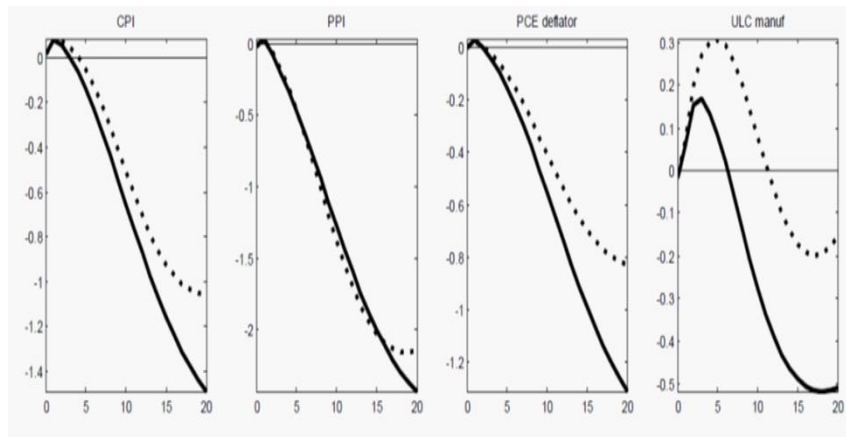
Structural FAVAR - Monetary policy (FFR) shock

Impulse responses from FAVAR (solid) and TV-FAVAR (dotted)



Structural FAVAR - Monetary policy (FFR) shock






Impulse responses from FAVAR (solid) and TV-FAVAR (dotted)





Structural FAVAR: Summary


- Structural factor augmented VARs are a promising tool as they address several issues with smaller scale VARs, such as omitted variable bias, curse of dimensionality, possibility of non-fundamental shocks, etc.
- FAVAR estimation and computation of the responses to structural shocks is rather simple, though managing a large dataset is not so simple
- Some problems in VAR analysis remain also in FAVARs, in particular robustness to alternative identification schemes, parameter instability, nonlinearities, etc.
- In the next lecture we will consider some extensions of the basic model that will address some of these issues.


References


-  Amengual, D. and Watson, M.W. (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
-  Bai, J. and S. Ng (2002). "Determining the number of factors in approximate factor models". Econometrica, 70, 191-221.
-  — Bai, J. and Ng, S., (2006). "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," Econometrica, 74(4), 1133-1150.
-  — Bai, J., and S. Ng (2008), "Large Dimensional Factor Analysis," Foundations and Trends in Econometrics, 3(2): 89-163.
-  — Bauer, D. (1998), Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood

 Banerjee, A., Marcellino, M. and I. Masten (2006). "Forecasting macroeconomic variables for the accession countries", in Artis, M., Banerjee, A. and Marcellino, M. (eds.), The European Enlargement: Prospects and Challenges, Cambridge: Cambridge University Press.

 Bernanke, B.S., Boivin, J. and P. Elias (2005). "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach", The Quarterly Journal of Economics, 120(1), 387–422.


 — Carriero, A., Kapetanios, G. and Marcellino, M. (2011), "Forecasting Large Datasets with Bayesian Reduced Rank Multivariate Models", Journal of Applied Econometrics, 26, 736-761.

 — Carriero, A., Kapetanios, G. and Marcellino, M. (2016), "Structural Analysis with Classical and Bayesian Large Reduced Rank VARs", Journal of Econometrics,






 Doz, C., Giannone, D. and L. Reichlin (2011). "A two-step estimator for large approximate dynamic factor models based on Kalman filtering," *Journal of Econometrics*, 164(1),

 188-205.


— Doz, C., Giannone, D. and L. Reichlin (2012). "A Quasi-Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models," *The Review of Economics and Statistics*, 94(4), 1014-1024.


 — Eickmeier, S., W. Lemke, M. Marcellino, (2014). "Classical time-varying FAVAR models - estimation, forecasting and structural analysis", *Journal of the Royal Statistical Society*, forthcoming.

— Eickmeier, S. and Ziegler, C. (2008). "How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach", *Journal of Forecasting*, (27),

-  Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2000), "The generalised factor model: identification and estimation", *The Review of Economic and Statistics*, 82, 540-554.
-  — Forni, M., M. Hallin, M. Lippi, L. Reichlin (2005), "The Generalized Dynamic Factor Model: One-sided estimation and forecasting", *Journal of the American Statistical Association*, 100, 830-840.
-  — Hallin, M., and Liška, R., (2007), "The Generalized Dynamic Factor Model: Determining the Number of Factors," *Journal of the American Statistical Association*, 102, 603-617
-  — Kapetanios, G. (2010), "A Testing Procedure for Determining the Number of Factors in Approximate Factor Models With Large Datasets". *Journal of Business and Economic Statistics*, 28(3), 397-409.
-  — Kapetanios, G., Marcellino, M. (2009). "A parametric

-  Lutkepohl, H. (2014), "Structural vector autoregressive analysis in a data rich environment", DIW WP.
-  — Marcellino, M., J.H. Stock and M.W. Watson (2003), "Macroeconomic forecasting in the Euro area: country specific versus euro wide information", European Economic Review, 47, 1-18.
-  — Marcellino, M., J. Stock and M.W. Watson, (2006), "A Comparison of Direct and Iterated AR Methods for Forecasting Macroeconomic Series h-Steps Ahead", Journal of Econometrics, 135, 499-526.
-  — Onatski, A. (2006). "Asymptotic Distribution of the Principal Components Estimator of Large Factor Models when Factors are Relatively Weak". Mimeo.
-  — Stock, J.H and M.W. Watson (1989), "New indexes of coincident and leading economic indicators." In NBER Macroeconomics Annual, 351–393, Blanchard, O. and S.

 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, 1167-1179.

 — Stock, J. H. and Watson, M. W. (2002b), “Macroeconomic Forecasting Using Diffusion Indexes”, *Journal of Business and Economic Statistics* 20(2), 147-162.

 — Stock, J.H., and M.W. Watson (2006), “Forecasting with Many Predictors,” ch. 6 in *Handbook of Economic Forecasting*, ed. by Graham Elliott, Clive W.J. Granger, and Allan Timmermann, Elsevier, 515-554.

— Stock, J. H. and Watson, M. W. (2011), *Dynamic Factor Models*, in Clements, M.P. and Hendry, D.F. (eds), *Oxford Handbook of Forecasting*, Oxford: Oxford University Press.



Stock, J.H. and Watson, M. W. (2015), "Factor Models for Macroeconomics," in J. B. Taylor and H. Uhlig (eds), Handbook of Macroeconomics, Vol. 2, North Holland.

The FHLR approach - DPCA

- The FHLR estimation procedure (assuming q known) is based on the so-called Dynamic Principal Components (DPC) and can be summarized as follows:
 - Estimate the spectral density matrix of X_t by periodogram-smoothing:

$$\Sigma^T(\theta_h) = \sum_{k=-M}^M \Gamma_k^T \omega_k e^{-ik\theta_h}$$

$$\theta_h = \frac{2\pi h}{2M+1}, \quad h = 0, \dots, 2M,$$

where M is the window width, ω_k are kernel weights and Γ_k^T is an estimator of $E(X_t - \bar{X}, X_{t-k} - \bar{X})$

- Calculate the first q eigenvectors of $\Sigma^T(\theta_h)$, $p^T(\theta_h)$, $j = 1, \dots, q$, for $h = 0, \dots, 2M$.

The FHLR approach - DPCA

-Define $p^T(L)$

as

$$p_j^T(L) = \sum_{k=-M}^M p_{j,k}^T L^k,$$

$$p_{j,k}^T = \frac{1}{2M+1} \sum_{h=0}^{2M} p_j^T(\theta_h) e^{ik\theta_h}, \quad k = -M, \dots, M.$$

- $p_j^T(L)x_t, j = 1, \dots, q$, are the first q dynamic principal components of x_t .

-Regress x_t on present, past, and future $p^T(L)x_t$. The fitted value is the estimated common component $\hat{\chi}_t$.

Parametric estimation - Subspace algorithms (SSS)

- The model $X_t^f = \mathcal{O}KX_t^p + \mathcal{E}E_t^f$ involves infinite dimensional vectors. In practice, use truncated versions, $X_{s,t}^f = (X_t', X_{t+1}', \dots, X_{t+s-1}')'$ and $X_{p,t}^p = (X_{t-1}', X_{t-2}', \dots, X_{t-p}')'$. Then, regress $X_{s,t}^f$ on $X_{p,t}^p$, and apply a singular value decomposition to $\hat{\Gamma}^f \hat{\mathcal{F}} \hat{\Gamma}^p$, where $\mathcal{F} = \mathcal{O}K$ and $\hat{\Gamma}^f$, and $\hat{\Gamma}^p$ are the sample covariances of $X_{s,t}^f$ and $X_{p,t}^p$ respectively. These weights are used to determine the importance of certain directions in $\hat{\mathcal{F}}$. Then, the estimate of K is given by

$$\hat{K} = \hat{S}_m^{1/2} \hat{V}_m \hat{\Gamma}^p^{-1}$$

where $\hat{U} \hat{S} \hat{V}$ represents the singular value decomposition of $\hat{\Gamma}^f \hat{\mathcal{F}} \hat{\Gamma}^p$, \hat{S} contains the singular values of $\hat{\Gamma}^f \hat{\mathcal{F}} \hat{\Gamma}^p$ in decreasing order, \hat{S}_m denotes the matrix containing the first m columns of \hat{S} and \hat{V}_m denotes the heading $m \times m$ submatrix of \hat{V} .

- Therefore, the SSS factor estimates are $\hat{f}_t = \hat{K} X_t^p$.

Parametric estimation - SSS, T asymptotics

- p must increase at a rate greater than $\ln(T)^\alpha$, for some $\alpha > 1$, but Np at a rate lower than $T^{1/3}$. N is fixed for the moment. A range of α between 1.05 and 1.5 provides a satisfactory performance.
- s is required to satisfy $sN > m$. As N is large this restriction is not binding, $s = 1$ is enough.
- If we define $\hat{f}_t = KX_t^p$, then \hat{f}_t converges to (the spanned by) f_t . The speed of convergence is between $T^{1/3}$ and $T^{1/2}$ because p grows. Note that consistency is possible because f_t depends on u_{t-1} . If f_t depends on $Af_{t-1} + u_t$, it does not converge.
- The asymptotic distribution of $\sqrt{T}^*(\text{vec}(\hat{f}) - \text{vec}(H_m f))$ with $f = (f_1, \dots, f_T)'$ is $N(0, V_f)$.
- Once estimates of the factors are available, estimates of the other parameters (including the factor loadings) can be obtained by OLS. Bauer (1998) proves \sqrt{T} consistency and asymptotic normality

Parametric estimation - SSS, T and N asymptotics

- If Np is $o(T^{1/3})$, p is $O(T^{1/z})$, $z > 3$, then when N and T diverge $\hat{f}_t = \hat{\mathcal{K}}X_t^p$ converges to (the space spanned by) f_t . The speed of convergence is $(T/Np)^{1/2}$. The intuition is that the estimator of $\mathcal{F} = \mathcal{OK}$ in $X_{s,t}^f = \mathcal{F}X_{p,t}^p + \mathcal{E}E_t^f$ remains consistent if $Np = o(T^{1/3})$.
- With a proper standardization, \hat{f}_t remains asymptotically normal
- Choice of number of factors can be done by information criteria, similar to those by Bai and Ng (2002) for PCA but with different penalty function.

Factor estimation methods - MC Comparison

- First set of experiments: a single VARMA factor with different specifications:

1 $a_1 = 0.2, b_1 = 0.4j$

2 $a_1 = 0.7, b_1 = 0.2j$

3 $a_1 = 0.3, a_2 = 0.1, b_1 = 0.15, b_2 = 0.15j$

4 $a_1 = 0.5, a_2 = 0.3, b_1 = 0.2, b_2 = 0.2j$

5 $a_1 = 0.2, b_1 = -0.4j$

6 $a_1 = 0.7, b_1 = -0.2j$

7 $a_1 = 0.3, a_2 = 0.1, b_1 = -0.15, b_2 = -0.15j$

8 $a_1 = 0.5, a_2 = 0.3, b_1 = -0.2, b_2 = -0.2j$

9 As 1 but $C = C_0 + C_1L$.

10 As 1 but one factor assumed instead of $p + q$

Factor estimation methods - MC Comparison

- Second group of experiments: as in 1-10 but with each idiosyncratic error being an AR(1) process with coefficient 0.2 (exp. 11-20). Experiments with cross correlation yield similar ranking of methods.
- Third group of experiments: 3 dimensional VAR(1) for the factors with diagonal matrix with elements equal to 0.5 (exp. 21).
- Fourth group of experiments: as 1-21 but the C matrix is $U(0,1)$ rather than $N(0,1)$.
- Fifth group of experiments: as 1-21 but using $s = 1$ instead of $s = m$.

Factor estimation methods - MC Comparison

- KM compute the correlation between true and estimated common component and the spectral coherency for selected frequencies. They also report the rejection probabilities of an LM(4) test for no correlation in the idiosyncratic component. The values are averages over all series and over all replications.
- Detailed results are in paper: for exp. 1-21, groups 1-3, see Tables 1-7; for exp. 1-21, group 4, see Table 8 for (N=50, T=50); for exp. 1-21, group 5, see Tables 9-11.

Factor estimation methods - MC Comparison, N=T=50

- Single ARMA factor (exp. 1-8): looking at correlations, SSS clearly outperforms PCA and DPCA. Gains wrt PCA rather limited, 5-10%, but systematic. Larger gains wrt DPCA, about 20%. Little evidence of correlation of idiosyncratic component, but rejection probabilities of LM(4) test systematically larger for DPCA.
- Serially correlated idiosyncratic errors (exp. 11-18): no major changes. Low rejection rate of LM(4) test due to low power for $T = 50$.
- Dynamic effect of factor (exp. 9 and 19): serious deterioration of SSS, a drop of about 25% in the correlation values. DPCA improves but it is still beaten by PCA. Choice of s matters:
for $s = 1$ SSS becomes comparable with PCA (Table 9).

Factor estimation methods - MC Comparison, $N=T=50$

- Misspecified number of factors (exp. 10 and 20): no major changes, actually slight increase in correlation. Due to reduced estimation uncertainty.
- Three autoregressive factors: (exp. 21): gap PCA-DPCA shrinks, higher correlation values than for one single factor. SSS deteriorates substantially, but improves and becomes comparable to PCA when $s = 1$ (Table 11).
- Full MLE gives very similar and only very slightly better results than PCA, and is dominated clearly by SSS.

Factor estimation methods - MC Comparison, other results

- Larger temporal dimension ($N=50, T=100, 500$), Correlation between true and estimated common component increases monotonically for all the methods, ranking of methods across experiments not affected. Performance of LM tests for serial correlation gets closer and closer to the theoretical one. (Tab 2,3)
- Larger cross-sectional dimension ($N=100, 200, T=50$), SSS is not affected (important, $N > T$), PCA and DPCA improve systematically, but SSS still yields the highest correlation in all cases, except exp. 9, 19, 21. (Tab 4,7).
 - Larger temporal and cross-sectional dimension ($N=100, T=100$ or $N=100, T=500$), The performance of all methods improves, more so for PCA and DPCA that benefit more for the larger value of N . SSS is in general the best in terms of correlation (Tab 5,6).
 - Uniform loading matrix, No major changes (Tab 8)
 - Choice of s , PCA and SSS perform very similarly (Tab 9)