Dynamic Factor Models

Anna Simoni[2]

[2]CREST, CNRS, Ensae, Ecole Polytechnique

Outline

Introduction. I

- The large datasets available today cover a number of years which is finite thus the number of data points is limited.
- On the other hand, more and more time series are collected and made available by statistical agencies.
- T = number of points in time ; N = number of series. Typically we are in a setting where $N \geq T$.
- The case $N \geq T$ in statistics is known as high-dimensional setting and dealing with such datasets by means of standard techniques as linear regressions constitutes a hard problem to be solved due to the lack of degrees of freedom.
- For example, in macroeconomic datasets we have $N \simeq 100, 1000$ and $T \simeq 100$ (quarterly or monthly series), while in financial datasets we have $N \simeq 100, 1000$ and $T \simeq 1000$ (daily series).

Introduction. II

- Factor models are a dimension reduction technique for analysing a large panel of time series.
- Parsimonious representation of the information.
- The recent availability of large datasets made them increasingly popular in the last twenty years.
- We suppose that the observed variables can be described by a small number of latent factors and that these common factors are the source of the correlation among the observable variables.

Introduction. III

Suppose to have quarterly observations on 207 time series (see Table 1)

- The full span of the dataset is $1959Q1 - 2014Q4$. Only 145 of the 207 series are available for this full period.
- From this dataset (after preliminary transformation of the data) one common component (factor) is extracted by using principal components analysis (see later).
- Then, the quarterly growth rate of each series (real Gross Domestic Product (GDP), total nonfarm employment, IP, and manufacturing and trade sales) is regressed on this common factor.
- None of the 4 series were used to extract the factor.
- As it can be seen from the figure, the single factor explains a large fraction of the 4-quarter variation in these four series.

Introduction. IV

- The R^2s of the four-quarter fits range from 0.73 for GDP to 0.92 for employment.
- At the same time, the estimated factor does not equal any one of these series, nor does it equal any one of the 58 series used to construct it.

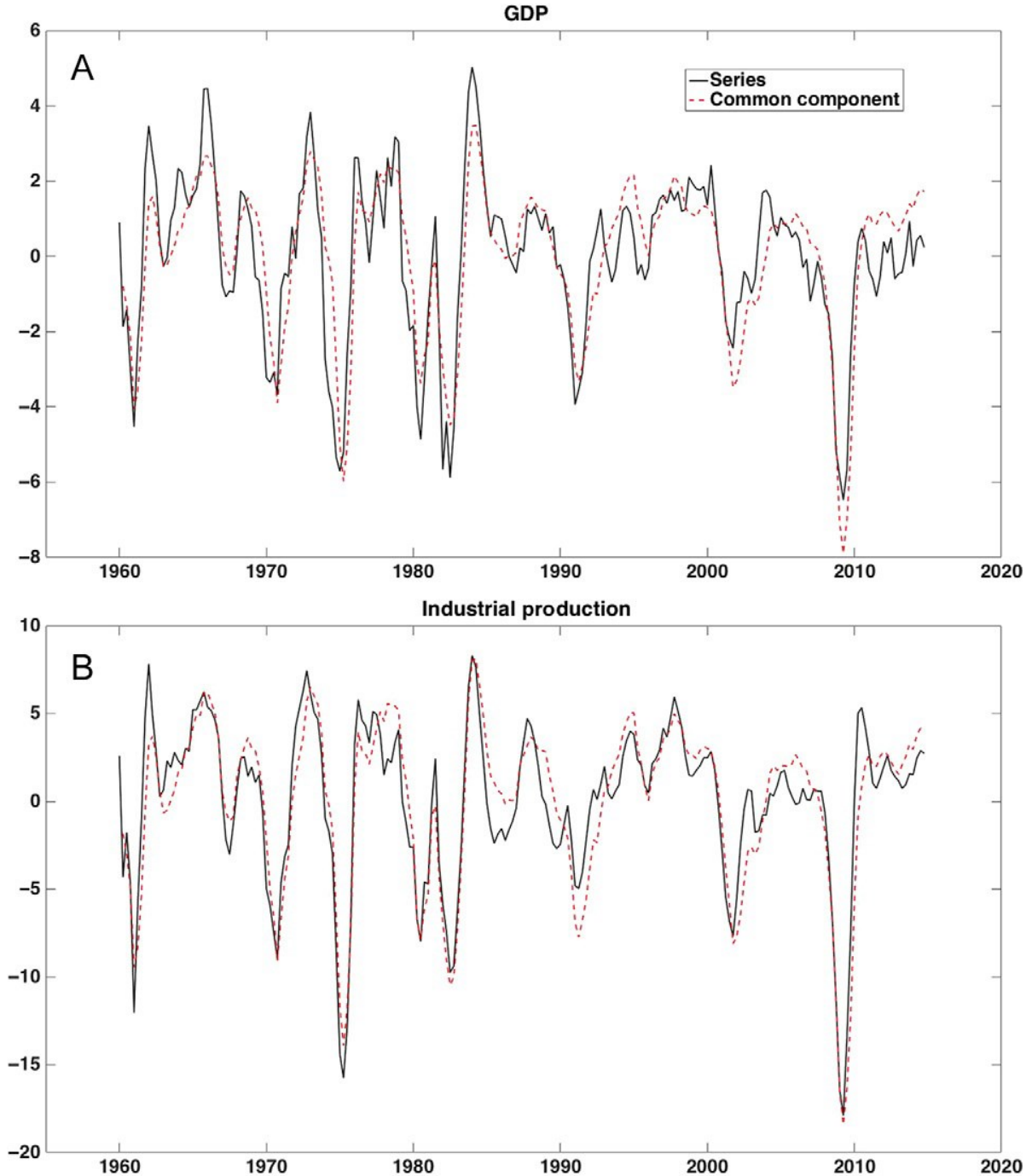
Source : Stock & Watson

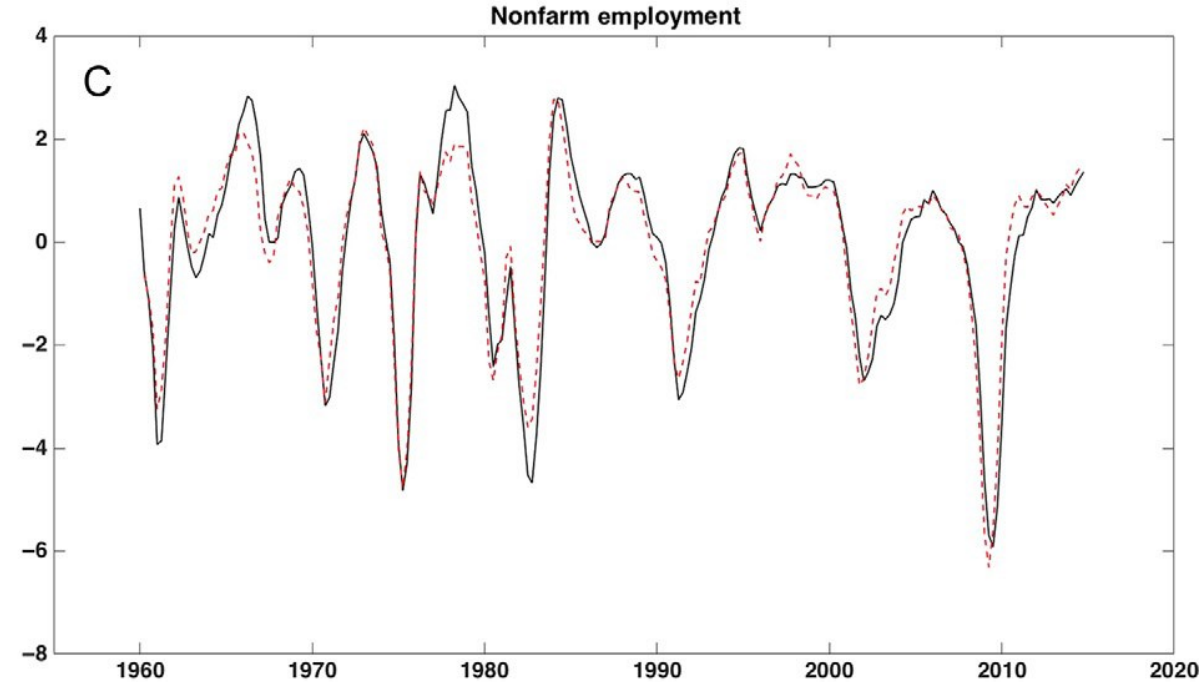
Table 1 Quarterly time series in the full dataset

	Category	Number of series	Number of series used for factor estimation
(1)	NIPA	20	12
(2)	Industrial production	11	7
(3)	Employment and unemployment	45	30
(4)	Orders, inventories, and sales	10	9
(5)	Housing starts and permits	8	6
(6)	Prices	37	24
(7)	Productivity and labor earnings	10	5
(8)	Interest rates	18	10
(9)	Money and credit	12	6
(10)	International	9	9
(11)	Asset prices, wealth, and household balance sheets	15	10
(12)	Other	2	2
(13)	Oil market variables	10	9
	Total	207	139

Notes: The real activity dataset consists of the variables in the categories 1–4.

Source : Stock & Watson




Nonfarm employment
C
4
2
0
−2
−4
−6
−8
1960
1970
1980
1990
2000
2010
2020


Manufacturing and trade sales
D
10
5
0
−5
−10
−15
−20
1960
1970
1980
1990
2000
2010
2020

Outline

Static Factor Models (SFM). I

The SFM expresses a N-vector X_t of observed zero-mean time series variables as depending on a reduced number $r < N$ of latent factors F_t and a mean-zero idiosyncratic component ε_t :

$$X_t \quad = \quad \Lambda F_t + \varepsilon_t, \qquad t = 1, \ldots, T \tag{1}$$

where :

- F_t is an r-vector and ε_t is an N-vector ;
- $\Lambda := (\lambda_1, \lambda_2, \ldots, \lambda_N)'$ is an $N \times r$ matrix ;
- $\mathbf{E}[\varepsilon_t] = \mathbf{E}[F_t] = 0$;
- $\mathbf{E}[F_t F_t'] = I_r$, $\mathbf{E}[F_t \varepsilon_\tau] = 0$, $\forall (t, \tau)$;
- $\mathbf{E}[F_t F_\tau] = 0$, $\forall (t, \tau)$, $t \neq \tau$;
- $\mathbf{E}[\varepsilon_t \varepsilon_\tau] = 0$, $\forall (t, \tau)$, $t \neq \tau$;
- **Exact** FM : $\mathbf{E}[\varepsilon_t \varepsilon_t'] = D = diag(d_1, \ldots, d_N)$

Static Factor Models (SFM). II

In a **static model :** the factors have only a contemporaneous effect on X_t.

The common component is $\chi_t := \Lambda F_t$.

Previous assumptions imply that $Cov(\chi_{it}, \varepsilon_{js}) = 0$, for all t, s and all $i, j \in \{1, \ldots, N\}$.

***

In an **exact** SFM : since the ε_{it} are pairwise uncorrelated, then all the correlation among the observed variables passes through the factors.

In the **Approximate SFM** : we do not suppose that $\mathbf{E}[\varepsilon_{it}\varepsilon_{jt}] = 0$ but that $\mathbf{E}[\varepsilon_{it}\varepsilon_{jt}]$ is negligible with respect to the correlation explained by the common factors.

That is, as $N \to \infty$, $\mathbf{E}[\varepsilon_t \varepsilon_t']$ remains bounded while $\Lambda\Lambda'$ is unbounded.

So, as $Var(X_t) = \Lambda\Lambda' + D$, we can consider that the part of the correlation between two variables that is not explained by the factors is negligible.

Outline

Dynamic Factor Models (DFMs). I

- The DFM represents the evolution of a vector of N observed time series, X_t, in terms of :
  - a reduced number of unobserved common factors which evolve over time,
  - plus uncorrelated disturbances (represent measurement error and/or idiosyncratic dynamics of the individual series).

- There are two ways to write the model : dynamic form vs. static form

- DFM : example of the class of state-space or hidden Markov models, in which observable variables are expressed in terms of unobserved or latent variables, which in turn evolve according to some lagged dynamics with finite dependence (*i.e.*, the law of motion of the latent variables is Markov).

- DFM important for macroeconometric applications because : the complex **comovements** of a potentially large number of observable series are summarized by a small number of common factors, which drive the **common fluctuations** of all the series.

Dynamic Factor Models (DFMs). II

Notation and Conventions :

- observable and latent variables are assumed to be second-order stationary.
- all data series are assumed to be transformed to have unit standard deviation.
- all the series in X_t are integrated of order zero ($X_t \sim I(0)$). So, remove by differencing the data stochastic trends and potential deterministic trends arising through drift if any.

State-space Models. I

- Consider a time series $\{Y_t\}_{t\geq 1}$.
- Specifying the joint finite-dimensional distributions of $(Y_1, \ldots, Y_t)$, for any $t \geq 1$, is not an easy task (independence and exchangeability not often justified in time series).
- Markovian dependence is the simplest form of dependence among the Y_t's. We say that $\{Y_t\}_{t\geq 1}$ is a *Markov chain* if, for any $t > 1$,

$$\pi(Y_t|Y_1, \ldots, Y_{t-1}) = \pi(Y_t|Y_{t-1}).$$

- Assuming a Markovian structure for the observations is, however, not appropriate in many applications.
- **State space models** build on the simple dependence structure of a Markov chain to define more complex models for $\{Y_t\}_t$.

State-space Models. II

- In a state space model we assume that there is an unobservable Markov chain θ_t, called the state process, and that Y_t is an imprecise measurement of θ_t.
- One can think of $\{\theta_t\}$ as an auxiliary time series that facilitates the task of specifying the probability distribution of the observable time series $\{Y_t\}_t$.
- Formally, a state space model consists of an $\mathbb{R}^p$valued time series $\{\theta_t; t = 0, 1, \ldots\}$ and an $\mathbb{R}^m$-valued time series $\{Y_t; t = 0, 1, \ldots\}$, satisfying the following assumptions :

  1. $\{\theta_t\}_t$ is a Markov chain,
  2. Conditionally on $\{\theta_t\}$, the Y_t's are independent and Y_t depends on θ_t only.

  As a consequence, a *state space model* is completely specified by the initial distribution $\pi(\theta_0)$ and the conditional densities $\pi(\theta_t|\theta_{t-1})$ and $\pi(y_t|\theta_t)$, $t \geq 1$. In fact, for any $t > 0$,

$$\pi(\theta_t, \ldots, \theta_0, y_1, \ldots, y_t) = \pi(\theta_0) \prod_{j=1}^{t} \pi(\theta_j|\theta_{j-1})\pi(y_j|\theta_j).$$

Dynamic form of the DFM. I

The DFM expresses a N-vector X_t of observed time series variables as depending on a reduced number q of latent factors f_t and a mean-zero idiosyncratic component ε_t :

$$X_t = \lambda(L)f_t + \varepsilon_t \quad (2)$$

$$f_t = \Psi(L)f_{t-1} + \eta_t, \quad (3)$$

where :

- f_t and ε_t are in general *serially correlated* (to capture dynamics);
- $\mathbf{E}[f_t] = 0$ and $\mathbf{E}[\varepsilon_t] = 0$;
- $\lambda(L)$ and $\Psi(L)$ are $N \times q$ and $q \times q$ lag polynomial matrices;
- $\lambda(L)f_t = \lambda_0 f_t + \ldots + \lambda_p f_{t-p}$
- η_t is *serially uncorrelated* and has mean zero;
- $\mathbf{E}[\varepsilon_t \eta'_{t-k}] = 0, \forall k.$

Dynamic form of the DFM. II

$(*)$ The i-th row of $\lambda(L)$, $\lambda_i(L)$, is called the dynamic factor loading for the i-th series X_{it}.

$(*)$ If $p = 0$: the **relation** between X_t and f_t is **static** (but X_t and f_t are dynamic processes).

$(*)$ The DFM capture the dynamics in the data in two ways :

- factors are autocorrelated and their dynamic is modeled as a VAR (or VARMA),
- and the observed variables can be affected by contemporaneous as well as past values of the factors.

Particular case for the measurement equation :

$$X_t = \lambda_0 f_t + \varepsilon_t,$$

the factors enter the equation only through their contemporaneous value.

$(*)$ The system of equations (2)-(3) is in the form of a state-space model. (2) is called *measurement equation* while (3) is called *state equation*.

Dynamic form of the DFM. III

$(*)$ The idiosyncratic error ε_t can be serially correlated. A simple and tractable model is the following :

$$\varepsilon_{it} = \delta_i(L)\varepsilon_{i,t-1} + \nu_{it}, \tag{4}$$

where ν_{it} is serially uncorrelated.

Dynamic form of the DFM. IV

1. Exact DFM

When $\mathbf{E}[\varepsilon_{it}\varepsilon_{js}] = 0, \forall t, s, i \neq j$.

So, the correlation of one series with another occurs only through the latent factors f_t.

Forecast : under Gaussianity of (ν_t, η_t) and (4),

$$\mathbf{E}[X_{i,t+1}|X_t, f_t, X_{t-1}, f_{t-1}, \ldots] = \alpha_i^f(L)f_t + \delta_i(L)X_{it}, \tag{5}$$

where $\alpha_i^f(L) = \lambda_{i0}\Psi(L) - \delta_i(L)\lambda_i(L) + L^{-1}(\lambda_i(L) - \lambda_0)$.

Estimation procedure :

- the dynamics of f_t and ε_t are specified;
- the model is cast in a state-space form;
- the likelihood can be computed through the Kalman filter under a Gaussian assumption, for any given value of the parameters of the model;
- the MLE of the parameters is obtained through a numerical procedure.

Dynamic form of the DFM. V

2. Approximate DFM

- The assumption that ε_t is uncorrelated across series is unrealistic in many applications.
- For example, data derived from the same survey might have correlated measurement error and multiple series for a given sector might have unmodeled sector-specific dynamics.
- Chamberlain and Rothschild's (1983) approximate factor model allows for such correlation.
- Under the approximate DFM, (5) would contain additional observable variables relevant for forecasting series X_{it}.
- Used when $N \uparrow \infty$. The components of ε_t might be correlated among them but we assume that the part of the correlation between observables due to the idiosyncratic components is negligible with respect to the part due to the common factors.

Static form of the DFM. I

The dynamic representation in (2)-(3) captures the dependence of the observed variables on the lags of the factors explicitly, while the static representation of the DFM embeds those dynamics implicitly.

The 2 forms lead to different estimation methods.

Two types of Static factor models :

- exact factor models : the factors explain all the correlation among the variables ;
- approximate factor models : adapted to the case where the number of observed variables $\uparrow \infty$, the factors explain most of the correlation among variables (the remaining part of the correlation is negligible).

Static form of the DFM. I

We rewrite the dynamic form above to depend on r static factors F_t instead of the q dynamic factors f_t, where $r > q$.

This rewriting makes the model amenable to principal components analysis (PCA).

- Let p =degree of the lag polynomial matrix $\lambda(L)$;
- let $F_t := (f_t', \ldots, f_{t-p}')'$ de an r-vector of *static* factors ($r = q(p+1)$);
- let $\Lambda := (\lambda_0, \lambda_1, \ldots, \lambda_p)$, where λ_h is a $N \times q$ matrix of coefficients on the h-th lag in $\lambda(L)$;
- let $\Phi(L)$ be the matrix consisting of 0s, 1s, and the elements of $\Psi(L)$ such that the VAR in (3) can be written as

$$X_t = \Lambda F_t + \varepsilon_t \tag{6}$$
$$F_t = \Phi(L) F_{t-1} + G\eta_t, \tag{7}$$

where $G = [I_q \; 0_{q\times(r-q)}]'$.

Static form of the DFM. II

- If $\varepsilon_t \sim AR(1)$ as in (4) and if $(\nu_t, \eta_t) \sim \mathcal{N}$, the one step ahead forecast of the i-th variable in the static factor model is

$$\mathbf{E}[X_{i,t+1}|X_t, F_t x_{t-1}, F_{t-1}, \ldots] = \alpha_i^F(L)F_t + \delta_i(L)X_{it}, \tag{8}$$

where $\alpha_i^F(L) = \Lambda_i\Phi(L) - \delta_i(L)\Lambda_i$.

Identification. I

Because the factors are unobserved, they are identified only up to arbitrary normalizations :

- In the static DFM, the space spanned by F_t is identified, but F_t itself is not identified :

$$\Lambda F_t = \Lambda Q^{-1} Q F_t,$$

  where Q is any invertible $r \times r$ matrix.

- Indeed, we will see that the method of PCs estimates the space spanned by F_t instead of F_t :

  if $\widehat{F}_t$ denotes the r-vector of factor estimates, there exists an $r \times r$ invertible matrix Q such that $\widehat{F}_t$ estimates $Q' F_t$.

- For many applications (*e.g.* macro monitoring and forecasting), the object of interest is the conditional mean and so it is necessary only to identify the space spanned by the factors (*i.e.* their linear combinations), not the factors themselves. So, Q is irrelevant.

Identification. II

- For such applications, the lack of identification is resolved by imposing a convenient normalization : the *principal components* normalization.

Principal Components Normalization :
the columns of Λ are orthogonal and are scaled to have unit norm :

$$\frac{1}{N}\Lambda'\Lambda = I_r \qquad \text{and} \qquad F'F \text{ diagonal.}$$

Alternatively,

$$\frac{1}{T}F'F = I_r \qquad \text{and} \qquad \Lambda'\Lambda \text{ diagonal,}$$

where $F := (F_1, \ldots, F_T)$ is the $(T \times r)$ matrix of factors.

While these restrictions identify the space spanned by the columns of F and the space spanned by the columns of Λ, they do not necessarily identify the individual columns of F or of Λ.

Normalization of the factors. III

IDENTIFICATION OF STATIC FACTORS :

- We are interested in conditions under which we can identity the columns of F and the columns of Λ from the product $F\Lambda'$.
- Since $F\Lambda' = FQQ^{-1}\Lambda$ for any $(r \times r)$ invertible matrix Q, and Q has r^2 free parameters, we need at least r^2 restrictions to identify F and Λ.
- Bai & Ng (2013, JoE) propose three sets of restrictions that lead to exact identification :

Identifying restrictions:		
	Restrictions on F	Restrictions on Λ
(2.1): PC1	$\frac{1}{T}F'F = I_r$	$\Lambda'\Lambda$ is a diagonal matrix with distinct entries
(2.2): PC2	$\frac{1}{T}F'F = I_r$	$\Lambda = \begin{pmatrix} \Lambda_1 \\ \Lambda_2 \end{pmatrix}$, $\Lambda_1 = \begin{pmatrix} \lambda_{11} & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{r1} & \lambda_{r2} & \cdots & \lambda_{rr} \end{pmatrix}$, $\lambda_{ii} \neq 0, i = 1, \ldots, r$
(2.3): PC3	Unrestricted	$\Lambda = \begin{pmatrix} I_r \\ \Lambda_2 \end{pmatrix}$

Normalization of the factors. IV

If the true F and true Λ satisfy these restrictions, then the corresponding rotation matrix is asymptotically I_r :

PC1. $H = I_r + \mathcal{O}_p(\delta_{NT}^{-2})$, with $\delta_{NT} := \min\{\sqrt{N}, \sqrt{T}\}$.

PC2. If $r = 1$, $H = I_r + \mathcal{O}_p(\delta_{NT}^{-2})$, with $\delta_{NT} := \min\{\sqrt{N}, \sqrt{T}\}$.
If $r > 1$, $H = I_r + \mathcal{O}_p(T^{-1/2})$.

PC3. H converges in probability to I_r.

Both PC1 and PC2 identify F and Λ up to a column sign change.

Outline

Estimation of the Factors and Parameters. I

Estimation of : the factors, the loadings, the number of factors, other parameters.

- Exact Static factor model :

  $\Rightarrow$ Estimation : MLE.

- Approximate Static factor model :

  $\Rightarrow$ Estimation : Principal Component Analysis (PCA)

  Chamberlain-Rotschild (1983) : large N case + strong assumption. Stock & Watson (2002)

- Exact Dynamic factor model :

  $\Rightarrow$ Estimation : MLE using Kalman filter

  Sargent & Sims (1977), Geweke (1977), Engle & Watson (1980, 1981, 1983), Stock & Watson (1989, 1991, 1993) : small N case.

Estimation of the Factors and Parameters. II

- Approximate Dynamic factor model :

  $\Rightarrow$ Estimation : PCA in the time domain , PCA in the frequency domain, Kalman Filter, QMLE

  Stock & Watson (1989, 1991, 1993), Forni and Reichlin (1996, 1998), Forni, Hallin, Lippi & Reichlin (1999), Bai & Ng : large N case.

Outline

ML estimation of small Exact SFMs. I

We consider an **Exact Static Factor Model** written in the matrix notation :

$$X = F\Lambda' + \varepsilon,$$

where

$$X := \begin{pmatrix} X_1' \\ \vdots \\ X_T' \end{pmatrix}, \qquad F := \begin{pmatrix} F_1' \\ \vdots \\ F_T' \end{pmatrix}, \qquad \varepsilon := \begin{pmatrix} \varepsilon_1' \\ \vdots \\ \varepsilon_T' \end{pmatrix}.$$

- Assume $\varepsilon_t | F \sim i.i.d. \mathcal{N}(0, \Omega)$, Ω p.d. and not related to F.
- The conditional log-likelihood function is :

$$-2\ell(\Lambda, F, \Omega) = TN\log(2\pi) + T\log(|\Omega|) + tr\left\{\Omega^{-1}(X - F\Lambda')'(X - F\Lambda')\right\}. \tag{9}$$

ML estimation of small Exact SFMs. II

- Assume for the moment that Ω is known. The conditional MLE (CMLE) of Λ is

$$\widehat{\Lambda} = X'F(F'F)^{-1}. \tag{10}$$

Replace this in (9) and use the standardization $F'F = TI_r$ to obtain the conditional MLE (CMLE) for F :

$$\widehat{F} = \sqrt{T} \times \Big(\text{matrix of eigenvectors corresponding to the } r \text{ largest eigenvalues of } X\Omega^{-1}X' \Big).$$

Hence, $\widehat{\Lambda} = \frac{1}{T}X'\widehat{F}$.

ML estimation of small Exact SFMs. III

Alternatively :

- The log-likelihood (9) can be maximized with respect to F first. In this case (by using the standardization $\widetilde{\Lambda}'\Omega^{-1}\widetilde{\Lambda} = NI_r$) we get :

$$\widetilde{\Lambda} = \sqrt{N} \times \Big(\text{matrix of eigenvectors corresponding to the } r \text{ largest eigenvalues of } \Omega^{-1}X'X\Omega^{-1}\Big)$$

$$\widetilde{F} = \frac{1}{N} X\Omega^{-1}\widetilde{\Lambda}.$$

- It holds : $\widehat{F}\widehat{\Lambda}' = \widetilde{F}\widetilde{\Lambda}'$.

ML estimation of small Exact SFMs. IV

- The conditional MLE of Ω is

$$\widehat{\Omega} = \frac{1}{N}(X - \widehat{F}\widehat{\Lambda})'(X - \widehat{F}\widehat{\Lambda}).$$

- In practice, an initial estimator of Ω is required (for instance use PCA).
- Notice that $\widehat{F}$ and $\widehat{\Lambda}$ are not MLEs since the MLEs require knowing the probabilistic structure of $\{F_t\}$ and use iterative computation methods.
- Without the normality assumption, the CMLEs $\widehat{\Lambda}$ and $\widehat{F}$ are called *generalized principal component estimators* (GPCEs) since they take into account the nonspherical structure of the error's var-cov matrix (as do the GLS estimator), see Choi 2012 and Stock & Watson 2006.

Principal Components (PCs) estimation of large Approximate SFMs. I

Nonparametric methods estimate the static factors directly without specifying a model for the factors or assuming specific distributions for the disturbances.

- PCs solve the least-squares problem in which Λ and F_t in (6) are treated as unknown parameters :

$$\min_{F,\Lambda} V_r(\Lambda, F), \qquad \text{where } V_r(\Lambda, F) := \frac{1}{NT}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t),$$
$$\text{s.t. } \frac{1}{N}\Lambda'\Lambda = I_r \quad \text{or} \quad \frac{1}{T}F'F = I_r. \qquad (11)$$

- The data have to be centered and standardized.

Principal Components (PCs) estimation of large Approximate SFMs. II

- If we concentrate out Λ and use the normalization that $\frac{1}{T}F'F = I_r$, the optimization problem is identical to

$$\max tr(F'(XX')F).$$

The estimated factor matrix is

$\check{F} = \sqrt{T} \times$ eigenvectors corresponding to the r largest eigenvalues of **XX'**.

Given $\check{F}$, $\check{\Lambda}' = (\check{F}'\check{F})^{-1}\check{F}'X = \check{F}'X/T$.

- Another solution to the least-squares problem (11) is given by

$\widehat{\Lambda} = \sqrt{N} \times$ eigenvectors corresponding to the r largest eigenvalues of **X'X**.

Given $\widehat{\Lambda}$, $\widehat{F} = X\widehat{\Lambda}/N$.

- The first solution is less computationally costly when $N > T$, while the second one when $T > N$.

Principal component analysis. I

We now look at the justification of the previous procedure.

- N fixed and finite.
- Assume to have a panel of i.i.d. data, *i.e.* with no serial correlation : $Cov(X_{it}, X_{j,t-k}) = 0$ for any k and any $i, j = 1, \ldots, n$.
- Assume $\mathbf{E}[X_t] = 0$ and define $\Sigma_X = \mathbf{E}[X_t X_t']$.

The covariance matrix Σ_X is a symmetric and positive definite matrix and can always be factorised as

$$\Sigma_X = PDP',$$

where $P = (p_1, \ldots, p_N)$ is an $N \times N$ orthogonal matrix of eigenvectors and $D = diag(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N > 0)$ contains the ordered eigenvalues. So $PP' = P'P = I_N$,

Then,

$$\Sigma_X p_j = \lambda_j p_j$$

Principal component analysis. II

and

$$\Sigma_X = \sum_{j=1}^{N} \lambda_j p_j p_j'.$$

Recall that the eigenvalues solve : $|\Sigma_X - \lambda_j I_N| = 0$, then once we have found the eigenvalues we can find the eigenvectors by solving :

$$(\Sigma_X - \lambda_j I_N) p_j = 0.$$

Equivalently, eigenvalues and eigenvectors are the solutions of the problem :

$$\begin{aligned} p_1 &= \arg\max_a a'\Sigma_X a \qquad \text{such that } a'a = 1 \\ p_j &= \arg\max_a a'\Sigma_X a \qquad \text{such that } a'a = 1 \quad \text{and } a'p_1 = \ldots = a'p_{j-1} = 0,\ j > 1. \end{aligned}$$

We can also write :

$$p_1 = \arg\max_a Var(\sum_{i=1}^{N} a_i X_{it}) \qquad \text{such that } \sum_{i=1}^{N} a_i^2 = 1.$$

Principal component analysis. III

- $p_j' X_t$ is called j-th principal component of X_t and has variance λ_j.
- The projection of X_t onto the j-th principal component is the vector

$$y_j := (p_j' X_t) p_j$$

  which has length λ_j and direction p_j.

Assume that we want to reduce the dimension of the data from N to $r < N$, consider the decomposition dictated by principal components :

$$X_t = \sum_{j=1}^{N} (p_j' X_t) p_j = \underbrace{\sum_{j=1}^{r} (p_j' X_t) p_j}_{X_{t,[r]}} + \underbrace{\sum_{j=r+1}^{N} (p_j' X_t) p_j}_{X_t - X_{t,[r]}}. \tag{12}$$

This dimension reduction technique is called **Principal Component Analysis (PCA)**. PCA is usually used to estimate static factor models.

Principal component analysis. IV

Denote $\xi_t := X_t - X_{t,[r]}$ with covariance Σ_ξ. Then, consider any other projection $\widetilde{X}_{t,[r]}$ with residuals $\widetilde{\xi}_t := \widetilde{X}_t - \widetilde{X}_{t,[r]}$ and covariance matrix $\Sigma_{\widetilde{\xi}}$. We can show that $X_{t,[r]}$ is the best r-dimensional representation of X_t in the sense that

$$tr\Sigma_\xi \leq tr\Sigma_{\widetilde{\xi}}.$$

So, this decomposition minimises the sum of the residual variances (which is equivalent to minimising the sum of all eigenvalues of the residual covariance Σ_ξ.

The covariance matrix of $X_{t,[r]}$ has just r eigenvalues different from zero and therefore has rank r, while the covariance matrix of ξ_t has rank $N - r$.

Principal component analysis. V

Now assume the static exact factor model for x_t :

$$X_t = \Lambda F_t + \varepsilon_t,$$

where F_t is a r-vector of static factors, Λ is a $N \times r$ matrix of loadings, and ε_t is an idiosyncratic error component as opposed to the common component ΛF_t.

- Assume orthogonality between the common and idiosyncratic components : $\mathbf{E}[F_t \varepsilon_t'] = 0$,
- Assume exact factor structure : $\mathbf{E}[\varepsilon_t \varepsilon_t'] = \Gamma_\varepsilon$ to be a positive definite and diagonal matrix.
- Without loss of generality, we can impose $\mathbf{E}[F_t F_t'] = I_r$. Therefore,

$$\Sigma_X = \Lambda\Lambda' + \Gamma_\varepsilon$$

where $\Lambda\Lambda'$ has rank r and Γ_ε has rank N.

Outline

Principal Components (PCs) estimation of large Approximate DFMs. I

Stock & Watson (2002, JASA) : $N, T \to \infty$.

- The model considered is the Static form of DFM (recall the dynamic form : $x_{it} = \lambda_i(L) f_t + \varepsilon_{it}$, with $\lambda_i(L) = \lambda_{0,i} + \lambda_{1,i} L + \ldots + \lambda_{p,i} L^p$) :

$$X_t \quad = \quad \Lambda F_t + \varepsilon_t, \tag{13}$$

  with $F_t := (f_t', \ldots, f_{t-p}')'$ an r-vector of *static* factors, and $\Lambda := (\lambda_0, \lambda_1, \ldots, \lambda_p)$, where λ_h is a $N \times q$ matrix of coefficients on the h-th lag in $\lambda(L)$ and $\lambda_h = (\lambda_{h,1}', \ldots, \lambda_{h,N}')'$.

- No restrictions on the relative rates of N and T.

- Main assumptions about the factors and factor loadings :
  1. $(\Lambda'\Lambda)/N \to I_r$;
  2. $\mathbf{E}(F_t F_t') = \Sigma_{FF}$, with Σ_{FF} diagonal with elements $\sigma_{ii} > \sigma_{jj} > 0$ for $i < j$;
  3. $\frac{1}{T} \sum_t F_t F_t' \to^p \Sigma_{FF}$.

Principal Components (PCs) estimation of large Approximate DFMs. II

- Main assumptions about the errors ε_t :
  1. $\mathbf{E}[\varepsilon_t'\varepsilon_{t+h}/N] = \gamma_{N,t}(h)$ and $\lim_{N\to\infty} \sup_t \sum_{h=-\infty}^{+\infty} |\gamma_{N,t}(h)| < \infty$;
  2. $\mathbf{E}[\varepsilon_{it}\varepsilon_{jt}] = \tau_{ij,t}$, and $\lim_{N\to\infty} \sup_t \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |\tau_{ij,t}| < \infty$;
  3. $\lim_{N\to\infty} \sup_{t,s} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |Cov(\varepsilon_{is}\varepsilon_{it}, \varepsilon_{js}\varepsilon_{jt})| < \infty$.

Principal Components (PCs) estimation of large Approximate DFMs. III

1) When $N < T$: consider the nonlinear least squares objective function

$$V(\widetilde{F}, \widetilde{\Lambda}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{T=1}^{T} (x_{it} - \widetilde{\lambda}_i \widetilde{F}_t)^2, \tag{14}$$

written as a function of hypothetical values of the factors $\widetilde{F} := (\widetilde{F}_1, \ldots, \widetilde{F}_T)'$ and factor loadings $\widetilde{\Lambda}$, where $\widetilde{\lambda}_i$ is the i-th row of $\widetilde{\Lambda}$.

- Let $\widehat{F}$ and $\widehat{\Lambda}$ denote the minimizers of $V(\widetilde{F}, \widetilde{\Lambda})$.
- After concentrating out $\widehat{F}$, minimizing (14) is equivalent to maximizing

$$tr\left[\widetilde{\Lambda}' X' X \widetilde{\Lambda}\right] \qquad \text{subject to } \widetilde{\Lambda}' \widetilde{\Lambda}/N = I_r,$$

where X is the $T \times N$ data matrix with t-th row X_t'.

Principal Components (PCs) estimation of large Approximate DFMs. IV

- This is the classical principal component problem which is solved by setting

  $\widehat{\Lambda} = \sqrt{N}$ eigenvectors of $X'X$ corresponding to its r largest eigenvalues.

- The resulting PCs estimator of F is

$$\widehat{F} = X\widehat{\Lambda}/N.$$

2) When $N > T$: we can use the following computational simpler approach :

   - concentrating out $\widehat{\Lambda}$, minimizing (14) is equivalent to maximizing

$$tr\left[\widetilde{F}'XX'\widetilde{F}\right] \qquad \text{subject to } \widetilde{F}'\widetilde{F}/T = I_r,$$

Principal Components (PCs) estimation of large Approximate DFMs. V

- this gives the estimator $\check{F}$ which is the matrix of the first r eigenvectors of XX' (times $\sqrt{T}$).
- The column spaces of $\widehat{F}$ and $\check{F}$ are equivalent and so they can be used interchangeably for forecasting purposes.

Principal Components (PCs) estimation of large Approximate DFMs. VI

Consistent Estimation of Factors :

- Because F and Λ are both unknown, both N and $T \to \infty$ are needed for the following result.

  Theorem : (Stock & Watson, 2002 JASA)
  Let S_i denote a variable with values of ± 1, let $N, T \to \infty$ and suppose that assumptions on factors, factor loadings and error hold. Suppose that k factors are estimated, where k may be $\leq$ or $> r$, the true number of factors. Then, S_i can be chosen so that the following holds :

  (a). For $i = 1, \ldots, r$, $T^{-1} \sum_{t=1}^{T} (S_i \widehat{F}_{it} - F_{it})^2 \to^p 0$;

  (b). For $i = 1, \ldots, r$, $S_i \widehat{F}_{it} \to^p F_{it}$ for every t;

  (c). For $i = 1, \ldots, r$, $T^{-1} \sum_{t=1}^{T} \widehat{F}_{it}^2 \to^p 0$ for every t.

Estimating the Number of Factors. I

Bai & Ng (2002, Econometrica).

- Suppose that r is the true number of factors.
- Consistent estimation of the number of factors when $N, T \to \infty$.
- Allow approximate factor structure.
- If all potentially informative factors are observed, but not the loadings : this would be a classical selection problem.
- In practice this is not the case.

Let

$$V(k, \widehat{F}^k) := \min_{\Lambda} \frac{1}{NT} \sum_{i=1} \sum_{t=1}^{T} (X_{it} - \lambda_i^k \widehat{F}_t^k)^2$$

denote the SSR when k factors are estimated.

Aim : to find penalty functions $g(N, T)$ such that criteria of the form

$$PC(k) := V(k, \widehat{F}^k) + kg(N, T)$$

Estimating the Number of Factors. II

can consistently estimate r.

- Let k_{max} be a bounded integer such that $r \leq k_{max}$.
- Let $\widehat{k} = \arg\min_{0 \leq k \leq k_{max}} PC(k)$. Under some assumption,

$$\lim_{N,T \to \infty} Prob[\widehat{k} = r] = 1$$

if :

(i) $g(N,T) \to 0$ and

(ii) $C_{NT}^2 g(N,T) \to \infty$ as $N, T \to \infty$, where $C_{NT} := \min\{\sqrt{N}, \sqrt{T}\}$.

Estimating the Number of Factors. III

Let $\widehat{\sigma}^2$ be a consistent estimator of $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{E}[\varepsilon_{it}^2]$.

Proposed $g(N,T)$ (Bai & Ng, 2002) :

$$\begin{aligned}
PC_{p1}(k) &:= V(k,\widehat{F}^k) + k\widehat{\sigma}^2\frac{N+T}{NT}\ln\left(\frac{NT}{N+T}\right);\\
PC_{p2}(k) &:= V(k,\widehat{F}^k) + k\widehat{\sigma}^2\frac{N+T}{NT}\ln C_{NT}^2;\\
PC_{p1}(k) &:= V(k,\widehat{F}^k) + k\widehat{\sigma}^2\frac{\ln C_{NT}^2}{C_{NT}^2}.
\end{aligned}$$

- These criteria generalize the C_p criterion of Mallows (1973) developed for selection of models in strict time-series or cross-section contexts to a panel data setting.
- All these criteria satisfy conditions (i) and (ii) above.
- The 3 criteria, although asymptotically equivalent, will have different properties in finite samples.

Estimating the Number of Factors. IV

Other valid criteria are the following :

$$
\begin{aligned}
IC_{p1}(k) &:= \ln(V(k,\widehat{F}^k)) + k\frac{N+T}{NT}\ln\left(\frac{NT}{N+T}\right); \\
IC_{p2}(k) &:= \ln(V(k,\widehat{F}^k)) + k\frac{N+T}{NT}\ln C_{NT}^2; \\
IC_{p1}(k) &:= \ln(V(k,\widehat{F}^k)) + k\frac{\ln C_{NT}^2}{C_{NT}^2}.
\end{aligned}
$$

The main advantage of these information criteria is that they do not depend on the choice of $k_{\max}$ through $\widehat{\sigma}^2$.

The scaling by $\widehat{\sigma}^2$ is implicitly performed by the logarithmic transformation of $V(k,\widehat{F}^k)$ and thus not required in the penalty term.

EM estimation with an unbalanced panel and data irregularities. I

Stock & Watson (2002, JBES).

- In practice, when N is large one encounters data irregularities : *e.g.* missing observations, unbalanced panels, and mixed frequency (for example, monthly and quarterly) data.
- In this case, a modification of standard PC estimation is necessary.
- To motivate the modification, consider the least squares estimators of Λ and F, from (13) from a balanced panel. The objective function is

$$V(F,\Lambda) := \sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \lambda_i' F_t)^2, \tag{15}$$

  where λ_i is the i-th row of Λ.

EM estimation with an unbalanced panel and data irregularities. II

- When the panel is unbalanced, LSs estimators of F_t can be calculated from the objective function

$$V^{\dagger}(F,\Lambda) := \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{1}_{it}(X_{it} - \lambda_i' F_t)^2, \tag{16}$$

  where $\mathbb{1}_{it} = 1$ if X_{it} is available and 0 otherwise.

- Minimization of (16) requires iterative methods. The following iterative method based on the EM algorithm has proved to be easy and effective.

EM estimation with an unbalanced panel and data irregularities. III

The j-th iteration of the algorithm is the following :

- Let $\widehat{\Lambda} := \widehat{\Lambda}^{(j-1)}$ and $\widehat{F} := \widehat{F}^{(j-1)}$ denote estimates of Λ and F constructed from the $(j-1)$st iteration, and let

$$Q(X^{\dagger}, \widehat{F}, \widehat{\Lambda}, F, \Lambda) := \mathbf{E}_{\widehat{F},\widehat{\Lambda}}[V(F,\Lambda)|X^{\dagger}], \tag{17}$$

  where $X^{\dagger}$ denotes the full set of *observed* data and $\mathbf{E}_{\widehat{F},\widehat{\Lambda}}[V(F,\Lambda)|X^{\dagger}]$ is the expectation of the complete data log-likelihood $V(F,\Lambda)$, conditional on $X^{\dagger}$, evaluated using the conditional density of $X|X^{\dagger}$ evaluated at $\widehat{F}$, $\widehat{\Lambda}$.
- Compute

$$\begin{aligned}(\widehat{F}^{(j)}, \widehat{\Lambda}^{(j)}) &= \arg\min_{F,\Lambda} Q(X^{\dagger}, \widehat{F}^{(j-1)}, \widehat{\Lambda}^{(j-1)}, F, \Lambda) \\ &= \arg\min_{F,\Lambda} \sum_{i=1}^{N}\sum_{t=1}^{T} \left\{ \mathbf{E}_{\widehat{F},\widehat{\Lambda}}[X_{it}^2|X^{\dagger}] + (\lambda_i' F_t)^2 - 2\widehat{X}_{it}(\lambda_i' F_t) \right\}\end{aligned}$$

  where $\widehat{X}_{it} := \mathbf{E}_{\widehat{F},\widehat{\Lambda}}[X_{it}|X^{\dagger}]$.

EM estimation with an unbalanced panel and data irregularities. IV

- We notice that $(\widehat{F}^{(j)}, \widehat{\Lambda}^{(j)})$ that minimize $Q(X^{\dagger}, \widehat{F}^{(j-1)}, \widehat{\Lambda}, F^{(j-1)}, \Lambda)$ can be equivalently calculated as the minimizers of

$$\widehat{V}(F, \Lambda) := \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{X}_{it} - \lambda_i' F_t)^2.$$

  At the j-th step this reduces to the usual PC calculation where the missing value are replaced by their expectation conditional on the observed data and using the parameter values from the previous iteration.

- If the full dataset contains a subset that constitutes a balanced panel, then starting values for $\widehat{F}$ in the EM iteration can be obtained using estimates from the balanced panel subset.

- Let $X_i := (X_{i1}, \ldots, X_{iT})'$ and $X_i^{\dagger}$ be the T-vector of observations on the i-th variable.
  - Suppose that $X_i^{\dagger} = A_i X_i$ for some known matrix A_i.
  - Then $\mathbf{E}[X_i | X^{\dagger}] = \mathbf{E}[X_i | X_i^{\dagger}] = F\lambda_i + A_i'(A_i A_i')^{-}(X_i^{\dagger} - A_i F \lambda_i)$, where $(A_i A_i')^{-}$ is the generalized inverse of $A_i A_i'$.

EM estimation with an unbalanced panel and data irregularities. V

Special cases :

1) Missing Observations.

- Suppose some observations for X_{it} are missing.
- Then, in iteration j, the elements of the estimated balanced panel are constructed as :

$$\widehat{X}_{it} = \begin{cases} X_{it} & \text{if } X_{it} \text{ observed} \\ \widehat{\lambda}_i' \widehat{F}_t & \text{otherwise.} \end{cases}$$

- The estimate of F is updated by computing the eigenvectors corresponding to the largest r eigenvalues of $\frac{1}{N} \sum_{i=1}^{N} \widehat{X}_i \widehat{X}_i'$, where $\widehat{X}_i := (\widehat{X}_{i1}, \ldots, \widehat{X}_{iT})'$.
- The estimate of Λ is updated by the OLS regression of $\widehat{X}$ on this updated estimate of F.

EM estimation with an unbalanced panel and data irregularities. VI

2) Mixed Monthly and Quarterly Data-I(0) Stock Variables.

- A series that is observed quarterly and is a stock variable would be the point-in-time level of a variable at the end of the quarter, say, the level of inventories at the end of the quarter.

- If this series is I(0), then it is handled as in case 1) (*i.e.* it is treated as a monthly series with missing observations in the 1^{st} and 2^{nd} month of the quarter).

EM estimation with an unbalanced panel and data irregularities. VII

3) Mixed Monthly and Quarterly Data-I(0) Flow Variables.

- A quarterly flow variable is the average (or sum) of unobserved monthly values.

- If this series is I(0), it can be treated as follows :
  - The unobserved monthly series, X_{it}, is measured only as the time aggregate X_{it}^q, where

$$X_{it}^q = \frac{1}{3}(X_{i,t-2} + X_{i,t-1} + X_{i,t}), \qquad \text{for } t = 3, 6, 9, 12, \ldots$$

    and X_{it}^q is missing for all other values of t.
  - In this case estimation proceeds as in case 1) but with $\widehat{X}_{it} = \widehat{\lambda}_i' \widehat{F}_t + \widehat{\varepsilon}_{it}$, where $\widehat{\varepsilon}_{it} = X_{i\tau}^q - \widehat{\lambda}_i'(\widehat{F}_{\tau-2} + \widehat{F}_{\tau-1} + \widehat{F}_\tau)/3$, where $\tau = 3$ when $t = 1, 2, 3$, $\tau = 6$ when $t = 4, 5, 6$, and so forth.

EM estimation with an unbalanced panel and data irregularities. VIII

4) Mixed Monthly and Quarterly Data-I(1) Stock Variables.

- Suppose that underlying monthly data are I(1).
- Let X_{it}^q denote the quarterly 1^{st} differences stock variable, assumed to be measured in the 3^{rd} month of every quarter.
- Let X_{it} denote the monthly 1^{st} difference of the variable.
- Then, $X_{it}^q = (X_{i,t-2} + X_{i,t-1}, +X_{i,t})$ for $t = 3, 6, 9, 12, \ldots$ and X_{it}^q is missing for all the other values of t.
- In this case estimation proceeds as in case 1) but with $\widehat{X}_{it} = \widehat{\lambda}_i' \widehat{F}_t + \frac{1}{3}\widehat{\varepsilon}_{it}$, where $\widehat{\varepsilon}_{it} = X_{i\tau}^q - \widehat{\lambda}_i'(\widehat{F}_{\tau-2} + \widehat{F}_{\tau-1} + \widehat{F}_\tau)$, where $\tau = 3$ when $t = 1, 2, 3$, $\tau = 6$ when $t = 4, 5, 6$, and so forth.

EM estimation with an unbalanced panel and data irregularities. IX

5) Mixed Monthly and Quarterly Data-I(1) Flow Variables.

- Let X_{it}^q denote the quarterly 1^{st} differences, assumed to be observed at the end of every quarter.
- The vector of observations is then $X_i^\dagger = (X_{i3}^q, X_{i6}^q, \ldots, X_{i\tau}^q)$, where τ denotes the month of the last quarterly observation.
- Let X_{it} denote the monthly 1^{st} difference.
- If the underlying quarterly data are averages of monthly series, then

$$X_{it}^q = \frac{1}{3}\left(X_{i,t} + 2X_{i,t-1} + 3X_{i,t-2} + 2X_{i,t-2} + X_{i,t-4}\right) \qquad \text{for } t = 3, 6, 9, 12, \ldots,$$

  and this implicitly defines the rows of A_i.
- Then, the estimate of X_i is given by

$$\widehat{X}_i = F\lambda_i + A_i'(A_iA_i')^-(X_i^\dagger - A_iF\lambda_i).$$

Hybrid Methods. I

Recall, with parametric state-space methods :

- State-space estimation entails specifying a full parametric model for X_t, ε_t, and f_t in the dynamic (or static) form of the DFM, so that the likelihood can be computed.

- A common treatment is to model the elements of ε_t as following the independent univariate ARs in (4) and

$$\{\nu_{it}\}_{i=1,\ldots,N} \sim i.i.d.\mathcal{N}(0, \sigma^2_{\nu_i}),$$

  $\eta_t \sim i.i.d.\mathcal{N}(0, \Sigma_\eta)$ and $\{\nu_t\} \perp\!\!\!\perp \{\eta_t\}$.

- Equations (2)-(4) constitute a complete linear state-space model.

- Alternatively, the static DFM can be written in state-space form using (6), (7) and (4).

- Given the parameters, the Kalman filter can be used to compute the likelihood and the Kalman smoother can be used to compute estimates of f_t given the full-sample data on $\{X_t\}$.

Hybrid Methods. II

- The likelihood can be maximized to obtain MLEs of the parameters.
- Alternatively, with the addition of a prior distribution, the Kalman filter can be used to compute the posterior distribution of the parameters and posterior estimates of the unobserved factors can be computed from the Kalman smoother.
- Parametric state-space methods have several advantages :
  - the use of quasi-maximum likelihood estimation,
  - the possibility of performing Bayes inference,
  - efficient treatment of missing observations,
  - the use of intertemporal smoothing to estimate the factors.
- Drawbacks :
  - their implementation becomes numerically challenging when N is large.
  - State-space methods require specifying the degree of the factor loading lag polynomial and models for the factors and for the idiosyncratic terms.

Hybrid Methods. III

- One way to handle the computational problem of ML estimation of the state-space parameters is to adopt a 2-step hybrid approach that combines the speed of PCs and the efficiency of the Kalman filter.

- So, state space/Kalman filter estimates can produce substantial improvements in estimates of the factors and common components if the *signal* of the common component is persistent and small.

2-step estimation of large approximate DFMs. I

Consider the following model :

- Assume that :

$$x_t = \Lambda F_t + \varepsilon_t$$

where

- F_t is the vector of common factors, assumed to be a zero-mean stationary process ;
- Λ is the $(N \times r)$-matrix of factor loadings ;
- $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})'$ = (idiosyncratic component) N-dimensional stationary process with mean 0 and covariance $\mathbf{E}[\varepsilon_t \varepsilon_t'] = \Psi$;
- ε_t and F_t uncorrelated at all leads and lags.

The factors are assumed to follow a VAR process :

$$F_t = AF_{t-1} + Bu_t,$$

where B is a $(r \times q)$ matrix of full rank q and $u_t \sim WN(0, I_q)$.

2-step estimation of large approximate DFMs. II

- Estimation procedure :

  STEP 1 :

  1. Obtain $\widehat{\Lambda}$ and $\widehat{F}_t$ by PCA.
  2. Obtain $\widehat{\varepsilon}_{it} = x_{it} - \widehat{\lambda}_i \widehat{F}_t$ and estimate their variance by the associated empirical variance $(\widehat{\Psi})_{ii}$.
  3. Obtain $\widehat{A}$ and an estimate of $\Sigma := Var(Bu_t)$, by using the estimate $\widehat{F}_t$ in the VAR specification.

  STEP 2 : Cast the model in the state-space form. Produce a new estimate of the factor by using Kalman filter :

$$\widehat{F}_{t|T} = \mathbf{E}(F_t | x_1, \ldots, x_T, \widehat{\theta})$$

  where θ is a vector collecting all the parameters and $\widehat{\theta}$ and estimator of it.

  Remark : The estimates of the factors in the 2^{nd} step are more efficient since the Kalman filter performs the best linear projection on the present and past observations.

2-step estimation of large approximate DFMs. III

Remark : In practice, the procedure outlined above is applied to standardized data

Alternative to STEP 1 : Estimate the parameters θ by QML.
This allows to exploit the advantage of MLE over PCs : efficiency gains, it allows to incorporate restrictions derived from economic theory.

Outline

Macroeconomic Forecasting with diffusion indexes. I

Stock & Watson (2002, JBES)

- The idea is to use the estimated factors for prediction.
- Forecasting is carried-out in two steps :
  1. the factors are estimated (by principal components) using X_t ;
  2. these estimated factors are used to forecast y_{t+1}.
- Focusing on the forecasts implied by the factors rather than on the factors themselves permits sidestepping the problem of identification (or rotation) inherent in factor models.
- One interpretationo of the estimated factors is in terms of diffusion indexes to measure common movement in a set of macroeconomic variables.
- So, the estimated factors are called diffusion indexes.

The statistical model is the following :

Macroeconomic Forecasting with diffusion indexes. II

- Let X_t be N-vector of predictor variables each one with mean zero.
- Let y_{t+1} be the scalar series to be forecast with zero mean.
- Suppose that (X_t, y_{t+1}) admit a dynamic factor model representation with q common dynamic factors f_t :

$$\begin{aligned} y_{t+1} &= \beta(L)f_t + \gamma(L)y_t + e_{t+1}, \qquad \mathbf{E}[e_{t+1}|f_t, y_t, X_t, f_{t-1}, y_{t-1}, X_{t-1}, \ldots] = 0, \\ X_{it} &= \lambda_i(L)'f_t + \varepsilon_{it}, \qquad i = 1, \ldots, N, \end{aligned}$$

where $\beta(L), \gamma(L)$ and $\lambda_i(L)$ are lag polynomials in nonnegative powers of L.

Macroeconomic Forecasting with diffusion indexes. III

Forecasting :

- In small N dynamic factor models, forecasts are generally constructed using a 3-step process (Stock and Watson 1989).

  1. Parametric models are postulated for the joint stochastic process $\{Y_{t+h}, X_t, w_t, \varepsilon_t\}$ in the model

$$X_t = \Lambda F_t + \varepsilon_t, \tag{18}$$

$$y_{t+h} = \beta_F' F_t + \beta_w' w_t + u_{t+h} \tag{19}$$

     and the sample data $\{Y_{t+h}, X_t, w_t\}_{t=1}^{T}$ are used to estimate the parameters of this process, typically using a Gaussian Maximum likelihood estimator (MLE).

  2. These estimated parameters are used in signal extraction algorithms to estimate the unknown value of F_t.

  3. The forecast of y_{t+h} is constructed using this estimated value of the factor and the estimated parameters.

Macroeconomic Forecasting with diffusion indexes. IV

- The first step in implementing this approach is to write the DFM as a linear state space model.
- For this, consider the static form of the DFM (6)-(7).
- The linear state space model is completed by specifying a process for ε_t and for η_t. Typically, ε_t are assumed to follow univariate autoregressions,

$$\varepsilon_{it} = \delta_i(L)\varepsilon_{i,t-1} + \nu_{it}, \qquad i = 1, \ldots, N.$$

- With this and the further assumptions : $\nu_{it} \sim i.i.d.\mathcal{N}(0, \sigma^2_{\nu_i})$, $i = 1, \ldots, N$, and $\eta_{jt} \sim i.i.d.\mathcal{N}(0, \sigma^2_{\eta_j})$, $j = 1, \ldots, q$, $\{\nu_t\} \perp\!\!\!\perp \{\eta_t\}$, equations (6)-(7) constitute a complete *linear state space model*.
- An advantage of this parametric state space formulation : it can handle data irregularities.

Macroeconomic Forecasting with diffusion indexes. V

- Example : if some series are observed weekly and some are observed monthly, the VAR process for the factor (7) can be formulated as evolving on a weekly time scale, but the dimension of the measurement equation (6) depends on which series are actually observed, that is, the row dimension of Λ would change depending on the variables actually observed at a given date.
- If some series are available for only a subset of the sample, the dimension of the measurement equation can change as time series become available.

- Drawback : the number of parameters is proportional to N, so direct estimation of the coefficients by MLE is cumbersome and historically was prohibitive for large systems. Therefore take a different approach and estimate the dynamic factors nonparametrically using the method of principal components.

Macroeconomic Forecasting with diffusion indexes. VI

Main assumptions about the forecasting equation : (let $z_t := (F_t', w_t')'$ and $\beta := (\beta_F', \beta_w')'$)

1. $\Sigma_{zz} := \mathbf{E}[z_t z_t']$ is a positive definite matrix.
2. $\frac{1}{T}\sum_{t=1}^{T} z_t z_t' \to^p \Sigma_{zz}$.
3. $\frac{1}{T}\sum_{t=1}^{T} z_t u_{t+h} \to^p 0$.
4. $\frac{1}{T}\sum_{t=1}^{T} u_{t+h}^2 \to^p \sigma^2$.
5. $|\beta| < \infty$.

Theorem : (Stock & Watson, 2002 JASA)
Suppose that the assumption about the forecasting equation and the conditions of the previous theorem hold. Let $\widehat{\beta}_F$ and $\widehat{\beta}_w$ denote the OLS estimates of β_F and β_w from the regression of $\{y_{t+h}\}_{t=1}^{T-h}$ onto $\{\widehat{F}_t, w_t\}_{t=1}^{T-h}$. Then the following hold :

(a). $(\widehat{\beta}_F \widehat{F}_T + \widehat{\beta}_w w_T) - (\beta_F F_T + \beta_w w_T) \to^p 0$;

(b). $\widehat{\beta}_w - \beta_w \to^p 0$ and S_i (defined in the previous Theorem) can be chosen so that $S_i \widehat{\beta}_{iF} - \beta_{iF} \to^p 0$ for $i = 1, \ldots, r$.

Macroeconomic Forecasting with diffusion indexes. VII

Empirical Example. (Stock & Watson 2002, JASA)

- Forecasting experiment for the Federal Reserve Board's Index of Industrial Production (IP_t) : $y_{t+12} := \ln(IP_{t+12}/IP_t)$.
- The variables making up X_t are 149 monthly macroeconomic variables representing facets of the macroeconomy (*e.g.* production, consumption, employment, price inflation, interest rates).
- Sample period : January 1959- December 1998.
- Principal components of X_t are used to construct forecast of y_{t+12}.
- The 12-month-ahead forecasts of y_{t+12} are constructed in each month starting $1970:1$ and extending through $1997:12$, using previously available data to estimate unknown parameters and factors.

Macroeconomic Forecasting with diffusion indexes. VIII

- Use data dated T and earlier in all calculations for constructing forecasts at time T :
  - For example, to compute the forecast in $T = 1970:1$, the variables making up X_t were standardized using data from $t = 1959:1 - 1970:1$ and PCs were computed.
  - These estimated values of F_t were used together with y_{t+12} for $t = 1959:1 - 1969:1$ to estimate β in (14).
  - Model selection based on AIC and BIC were used to determine the number of factors to include in the regression.
  - The forecasts constructed in $T = 1970:1$ were formed as : $\widehat{y}_{t+12/t} = \widehat{\beta}\widehat{F}_t$.
  - This process was repeated for $1970:2 - 1997:12$.

Macroeconomic Forecasting with diffusion indexes. IX

- For comparison : compute forecasts using four other methods :
  - a univariate autoregression in which Y_{t+12} was regressed on lags of $\ln(IP_t/IP_{t-l})$,
  - a VAR that included the rate of price inflation and short-term interest rates in addition to the rate of growth of the industrial production index,
  - a leading-indicator model in which Y_{t+12} was regressed on 11 leading indicators chosen by Stock and Watson (1989) as good predictors of aggregate macroeconomic activity ;
  - and an autoregressive-augmented PCs model in which Y_{t+12} was regressed on the estimated factors and lags of $ln(IP_t/IP_{t-l})$.

Macroeconomic Forecasting with diffusion indexes. X

Table :

- The first PC is executed with the number of factors determined by IC_{p3} ;
- The other PC forecasts use a fixed number of factors.
- Last row : PC forecasting model (with the number of PC determined by IC_{p3}) augmented with BIC-selected lags of the growth rate of IP.

Results :

- Leading indicators and VAR performed slightly better than the univariate AR in OOS ;
- The factor models offer substantial improvement.
- Nearly all of the forecasting gain come from the first 2 or 3 factors and once these factors are included there is no additional gain from including lagged values of IP growth.

Source : Stock & Watson 2002, JASA

Macroeconomic Forecasting with diffusion indexes. XI

Table 2. Simulated Out-of-Sample Forecasting Results Industrial Production, 12-Month Horizon

Forecast method	*Relative MSE*
Univariate autoregression	1.00
Vector autogression	.97
Leading indicators	.86
Principal components	.58
Principal components, $k = 1$	.94
Principal components, $k = 2$	.62
Principal components, $k = 3$	.55
Principal components, $k = 4$	.56
Principal components, AR	.69
Root MSE, AR model	.049

NOTE: For each forecast method, this table shows the ratio of the MSE of the forecast made by the method for that row to the MSE of a univariate autoregressive forecast with lag length selected by the BIC. The final line presents the root MSE for the autoregressive model in native (decimal growth rate) units at an annual rate.

DFMs for macroeconomic monitoring and forecasting. I

Two applications of DFMs :

1. real-time macroeconomic monitoring,
2. forecasting.

- DFMs have resulted in meaningful forecasting improvements, especially for measures of real economic activity.
- They have also proven useful for the important task of macroeconomic monitoring, *i.e.*, tracking economies in real time.

1) Macroeconomic monitoring :

Economists at central banks, executive branches of government, and in the private sector track the evolution of the economy in real time.

- Key part of macroeconomic monitoring is following and interpreting data releases to gain insights as to where the economy is at present, and where the economy is going.

DFMs for macroeconomic monitoring and forecasting. II

- Two challenges :
  - data are releases throughout the month and quarter, so that the available data change from day to day or even within a day, a feature referred to as the *ragged edge* problem ;
  - the number of data releases and series contained within those releases is vast. Handling this flow of large volumes of disparate data requires judgment and knowledge of idiosyncratic events.

- DFMs are used for 2 related macro monitoring tasks :
  - The first is the **construction of indices** that distill the currently available data into a concise summary of economic conditions. $\Rightarrow$ construction of an index of indicators of economic activity.

    In the DFM, the latent factor summarizes the comovements of the observed variables, so in a DFM with a single factor, the estimate of the latent factor is a natural index of the movements of the relevant time series.

DFMs for macroeconomic monitoring and forecasting. III

- The second is **nowcasting** : task of *forecasting* the current value of a specific series which has not yet been released, (*i.e.* forecasting the value of fourth-quarter GDP in November).

  DFMs permit specifying an internally consistent model that can be used for nowcasting multiple variables while placing appropriate weight on new data releases.

Outline

The Kalman Filter : introduction.

- The Kalman filter is an algorithm for sequentially updating a linear projection for a dynamic system (in the *state-space representation*).
- It provides a way to :
  - calculate exact finite-sample forecasts and the exact likelihood function for Gaussian ARMA processes ;
  - factor matrix autocovariance-generating functions or spectral densities ;
  - estimate vector autoregressions with coefficients that change over time.

The State-Space representation. I

- Let X_t be an $(N \times 1)$ vector of variables observed at date t and f_t a $(r \times 1)$ vector of unobservable variables (*state vector*).
- The state-sapce representation of the dynamics of X_t is given by the following system of equations :

$$\begin{aligned} X_t &= \Lambda_0 f_t + \Lambda_1 f_{t-1} + \ldots + \Lambda_s f_{t-s} + \varepsilon_t, && \text{observation equation} \\ f_t &= \Psi_1 f_{t-1} + \ldots + \Psi_p f_{t-p} + \eta_t, && \text{state equation.} \end{aligned}$$

- Assumptions :
  1. η_t is vector white noise : $\mathbf{E}[\eta_t \eta_\tau'] = Q$ for $t = \tau$ and $= 0$ otherwise, where Q is a $(r \times r)$ matrix ;
  2. ε_t is vector white noise : $\mathbf{E}[\varepsilon_t \varepsilon_\tau'] = R$ for $t = \tau$ and $= 0$ otherwise, where R is a $(N \times N)$ matrix ;
  3. $\mathbf{E}[\eta_t \varepsilon_\tau'] = 0$ for all t and τ ;

The State-Space representation. II

4. Initial conditions : $\mathbf{E}[\eta_t f_1'] = 0$ for $t = 2, \ldots, T$, and $\mathbf{E}[\varepsilon_t f_1'] = 0$ for $t = 2, \ldots, T$;

- The state equation implies that f_t can be written as a linear function of $(f_1, \eta_2, \eta_3, \ldots, \eta_t)$ for $t = 2, 3, \ldots, T$.
- So, if $s \geq p - 1$ the state equation writes :

$$\begin{bmatrix} f_t \\ f_{t-1} \\ \vdots \\ f_{t-p+1} \\ \vdots \\ f_{t-s} \end{bmatrix} = \begin{bmatrix} \Psi_1 & \cdots & \Psi_p & 0 & \cdots & 0 \\ I_r & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & I_r & 0 & \cdots & \cdots & 0 \\ \cdots & \ddots & \ddots & \ddots & & \vdots \\ \cdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & \cdots & 0 & I_r & 0 \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-2} \\ \vdots \\ f_{t-p} \\ \vdots \\ f_{t-s-1} \end{bmatrix} + \begin{bmatrix} I_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \eta_t. \tag{20}$$

The State-Space representation. III

- If $s < p - 1$ the state equation writes :

$$\begin{bmatrix} f_t \\ f_{t-1} \\ \vdots \\ f_{t-p+1} \end{bmatrix} = \begin{bmatrix} \Psi_1 & \Psi_2 & \cdots & \Psi_p \\ I_r & 0 & \cdots & 0 \\ \cdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I_r & 0 \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-2} \\ \vdots \\ f_{t-p} \end{bmatrix} + \begin{bmatrix} I_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \eta_t. \qquad (21)$$

- Once the model is written in state-space form, the Gaussian likelihood can be computed using the Kalman filter for any value of the parameters.

- The likelihood can be maximized by any numerical optimization procedure over the parameter space (Watson & Engle (1983) proposed to use a score algorithm, or the EM algorithm, or a combination of both.

- Given the parameter estimates, $\widehat{\theta}$, the Kalman smoother provides an approximation of f_t using information from all observations :

$$\widehat{f}_{t|T} = \widehat{\mathbf{E}}[f_t | x_1, \ldots, x_T, \widehat{\theta}].$$

Overview of the Kalman Filter. I

- Suppose that the parameters of the system (Λ, Ψ, Q, R) are known. Then, the Kalman Filter is an algorithm to calculate linear least squares **forecasts** of the state vector f_t on the basis of data observed through date t :

$$\widehat{f}_{t+1|t} = \widehat{\mathbf{E}}_L[f_{t+1}|x_1, \ldots, x_t, \widehat{\theta}].$$

  where $\widehat{\mathbf{E}}_L[\cdot]$ denotes the linear projection of f_{t+1} on $x_1, \ldots, x_t$ and a constant.

- The Kalman Filter calculates these forecasts recursively, generating $\widehat{f}_{1|0}, \widehat{f}_{2|1}$, $\ldots, \widehat{f}_{T|T-1}$.

- Associated with each of these forecasts is a MSE matrix :

$$P_{t+1|t} := \mathbf{E}[(f_{t+1} - \widehat{f}_{t+1|t})(f_{t+1} - \widehat{f}_{t+1|t})'].$$

Starting the recursion.

- The recursion starts with $\widehat{f}_{1|0}$ which is just the unconditional mean of f_1 :

$$\widehat{f}_{1|0} = \mathbf{E}(f_1) = 0$$

with associated *MSE* (unconditional variance) :

$$P_{1|0} := \mathbf{E}[(f_1 - \widehat{f}_{1|0})(f_1 - \widehat{f}_{1|0})'] = \mathbf{E}[(f_1 - \mathbf{E}(f_1))(f_1 - \mathbf{E}(f_1))'].$$

Next steps. I

- Given starting values $\widehat{f}_{1|0}$ and $P_{1|0}$ the next step is to calculate similar quantities for the following date : $\widehat{f}_{t+1|t}$ and $P_{t+1|t}$ for every $t = 2, \ldots, T-1$.
- For $t > 1$ do :

  Forecasting X_t **:** given $\widehat{f}_{t|t-1}$ and $P_{t|t-1}$,

  1. forecast the value of X_t : $\widehat{X}_{t|t-1} = \Lambda(L)\widehat{f}_{t|t-1}$,
  2. the error of this forecast is :

$$X_t - \widehat{X}_{t|t-1} = \Lambda(L)(f_t - \widehat{f}_{t|t-1}) + \varepsilon_t$$

  with MSE :

$$\begin{aligned}\mathbf{E}[(X_t - \widehat{X}_{t|t-1})(X_t - \widehat{X}_{t|t-1})'] &= \mathbf{E}[\Lambda(L)(f_t - \widehat{f}_{t|t-1})(f_t - \widehat{f}_{t|t-1})'\Lambda(L)'] + \mathbf{E}[\varepsilon_t \varepsilon_t'] \\ &= \Lambda(L)P_{t|t-1}\Lambda(L)' + R.\end{aligned}$$

Next steps. II

Updating the Inference about f_t :

1. using the formula for updating a linear projection

$$\widehat{f}_{t|t} = \widehat{f}_{t|t-1} + P_{t|t-1}\Lambda(L)'(\Lambda(L)P_{t|t-1}\Lambda(L)' + R)^{-1}(X_t - \Lambda(L)\widehat{f}_{t|t-1}),$$

2. the MSE associated with this updated projection :

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}\Lambda(L)'(\Lambda(L)P_{t|t-1}\Lambda(L)' + R)^{-1}\Lambda(L)P_{t|t-1}.$$

Producing a Forecast of f_t :

1. using the state equation :

$$\widehat{f}_{t+1|t} = \Psi(L)\widehat{f}_{t|t}$$
$$= \Psi(L)\widehat{f}_{t|t-1} + \underbrace{\Psi(L)P_{t|t-1}\Lambda(L)'\left(\Lambda(L)P_{t|t-1}\Lambda(L)' + R\right)^{-1}}_{=:K_t=\text{gain matrix}}\left(X_t - \Lambda(L)\widehat{f}_{t|t-1}\right)$$
$$= \Psi(L)\widehat{f}_{t|t-1} + K_t\left(X_t - \Lambda(L)\widehat{f}_{t|t-1}\right),$$

2. the MSE of this forecast is :

$$P_{t+1|t} = \Psi(L)P_{t|t}\Psi(L)' + Q = \Psi(L)P_{t|t-1}\Psi(L)' + K_t\Lambda(L)P_{t|t-1}\Psi(L)' + Q$$
$$= \left(\Psi(L) - K_t\Lambda(L)\right)P_{t|t-1}\left(\Psi(L) - K_t\Lambda(L)\right)' + K_tRK_t' + Q.$$

Smoothed inference of f_t :

- So far, we have been concerned with a forecast of f_t given information available at date $t-1$.
- But we can use information through the end of the sample (date T) to improve the inference.
- First we run the data through the Kalman filter, storing the sequences $\{\widehat{f}_{t|t}\}_{t=1}^{T}$, $\{\widehat{f}_{t|t-1}\}_{t=1}^{T}$, $\{P_{t|t}\}_{t=1}^{T}$, $\{P_{t|t-1}\}_{t=1}^{T}$.
- The sequence of smoothed estimates is then calculated as in reverse order by iterating on

$$\widehat{f}_{t|T} = \widehat{f}_{t|t} + C_t(\widehat{f}_{t+1|T} - \widehat{f}_{t+1|t}), \qquad \text{for } t = T-1, T-2, \ldots, 1$$

  where $C_t = P_{t|t}\Psi(L)' P_{t+1|t}^{-1}$.
- The corresponding MSE are found by iterating on

$$P_{t|T} = P_{t|t} + C_t(P_{t+1|T} - P_{t+1|t})C_t'$$

  in reverse order for $t = T-1, T-2, \ldots, 1$.

Outline

Next steps. I

- The EM algorithm was proposed by Dempster et al. (1977) as a general solution to problems for which incomplete or latent data make the likelihood intractable or difficult to deal with.

- Essential idea : to write the likelihood as if the data were complete and to iterate between 2 steps :

  - in the expectation step we "fill in" the missing data in the likelihood,

  - in the maximization step we reoptimize this expectation.

  More precisely, let $\ell(X, F; \theta)$ be the joint log-likelihood of $X := (X_1', \ldots, X_T')$ and $F = (F_1', \ldots, F_T')$.

- Given the available data $\Omega_T \subseteq X$, for the model given by (6)-(7) the EM algorithm proceeds in a sequence of 2 steps :

Next steps. II

1. **E-step :** the expectation of the log-likelihood conditional on the data is calculated using the estimates from the previous iteration, $\theta(j)$:

$$L(\theta, \theta(j)) = \mathbf{E}_{\theta(j)}[\ell(X, F; \theta)].$$

2. **M-step :** the parameters are re-estimated through the maximization of the expected log-likelihood with respect to θ :

$$\theta(j+1) = \arg\max_{\theta} L(\theta, \theta(j)).$$

Next steps. III

Consider (6)-(7) with $\varepsilon_t \sim i.i.d.\mathcal{N}(0, R)$ where R is a diagonal matrix. In that case, $\theta = \{\Lambda, \Psi, Q, R\}$, where $\Phi = [\Phi_1, \ldots, \Phi_p]$ and $\eta_t \sim i.i.d.\mathcal{N}(0, Q)$.