Lecture 9: Dynamic Factor Models I

Geert Mesters (Universitat Pompeu Fabra, Barcelona GSE and VU Amsterdam)

Motivation

Large datasets

- During the postwar period statistical agencies have collected monthly or quarterly data on many related macroeconomic, financial, and sectoral variables.
- As a consequence macroeconometricians face data sets that include many time series, but the number of observations on each series is relatively short, for example 20 to 40 years of quarterly data
- Incorporating many variables in time series models is challenging: curse of dimensionality

How dynamic factor models help

- Take the challenge by exploiting that many variables exhibit similar fluctuations over time
- Dynamic factor models are widely applied in economics and finance because of their ability to handle large panels
  - The model for business cycle indicator construction
  - The forecasting model for most central banks
- Empirically they have good forecasting properties and are useful as inputs in several structural models

Alternative dimension reduction methods

- In high dimensional applications the goal is typically to reduce the dimension using a reasonable assumptions
- In macroeconomics there exists a lot of co-movement among the series: THIS is why factor models typically work well
- But the existence of the factor structure remains an assumption, which can be tested only under certain other assumptions
- In contrast, in other fields (e.g. genome studies) researchers have strong evidence that their prediction problem is sparse and they exploit this structure to reduced the dimensionality
- Just to say factor models are not always the best option

Running example

Data series

- Today we illustrate the methodology using 144 disaggregated macroeconomic time series
- All data series are obtained from `https://fred.stlouisfed.org/`
- These types of data panels have become popular ever since the seminal contributions of Forni et al (2000) and Stock & Watson (2002)
- The careful construction of the data panel is important

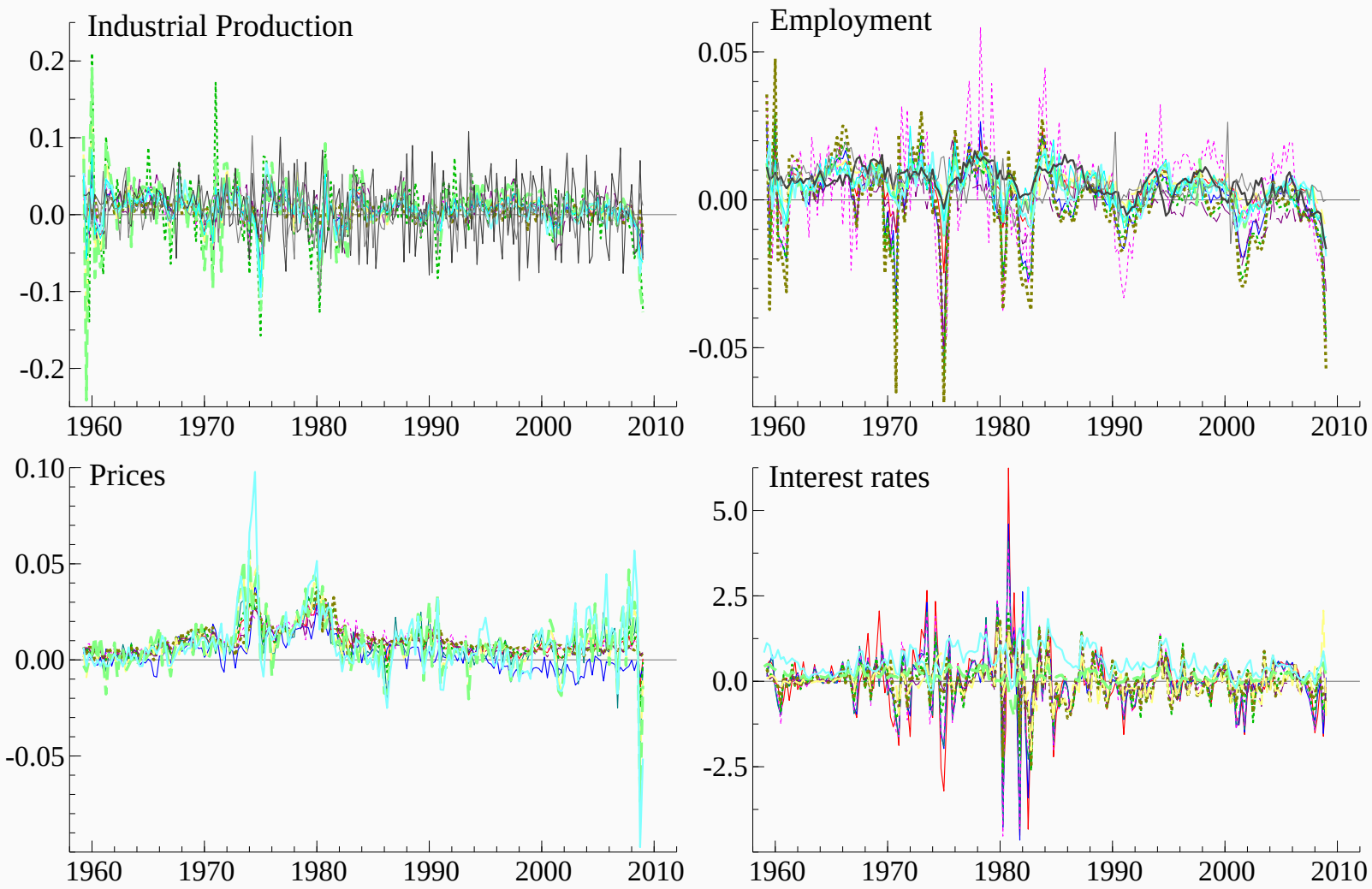
Data series

	Category	number of series (144)
A	GDP components	16
B	Industrial production	14
C	Employment	20
D	Unemployment rate	7
E	Housing starts	6
F	Inventories	6
G	Prices	37
H	Wages	6
I	Interest rates	13
J	Money	8
K	Exchange rates	5
L	Stock prices	5
M	Consumer expectations	1

Some comments

- All series are disaggregate
- All series are transformed to stationarity
- All series are aggregated to quarterly frequency

Examples


Industrial Production
0.2
0.1
0.0
-0.1
-0.2
1960
1970
1980
1990
2000
2010
Employment
0.05
0.00
-0.05
1960
1970
1980
1990
2000
2010
Prices
0.10
0.05
0.00
-0.05
1960
1970
1980
1990
2000
2010
Interest rates
5.0
2.5
0.0
-2.5
1960
1970
1980
1990
2000
2010

Goals

Such disaggregate data panels can be useful for

- Forecasting
- Business cycle measurement
- Structural analysis

Model

Set up

- Let the panel sizes be
  - N the number of time series
  - T the number of time periods
- Dynamic factor models can be used even when $N > T$ and are thus useful for data rich environments

Model formulation

$$y_t = \mathbf{\Lambda} f_t + \epsilon_t$$

where

- y_t: $N \times 1$ vector of observations
- f_t: $r \times 1$ vector of common factors
- $\mathbf{\Lambda}$: $N \times r$ loading matrix
- ϵ_t: $N \times 1$ disturbance vector

Some comments

- The premise of a dynamic factor model is that a few latent dynamic factors, f_t, drive the comovements of a high-dimensional vector of time-series variables, y_t, which is also affected by a vector of mean-zero idiosyncratic disturbances, ϵ.
- These idiosyncratic disturbances arise from measurement error and from special features that are specific to an individual series
- A central empirical finding is that a few factors can explain a large fraction of the variance of many macroeconomic series

Some more comments

- Λ, f_t and ϵ_t are all not observed!!!
- We only observe the vectors y_t
- The number of factors r is typically considered small, say $r = 1, 2, 3, ...$
- The DFM decomposes variables y_t into two components
  1. Λf_t the common component
  2. ϵ_t the idiosyncratic component (time series specific)
- The common component is moving trough time according to the common factors f_t that are loaded to the variables via Λ

Matrix notation

$$Y = \Lambda F + \epsilon$$

where

- Y: $N \times T$ matrix of observations
- $F = (f_1, \ldots, f_T)$: $r \times T$ matrix of common factors
- Λ: $N \times r$ loading matrix
- ϵ: $N \times T$ disturbance matrix

Identification

Identification

- Without further restrictions the dynamic factor model is not identified

For any non-singular $r \times r$ matrix A we have

$$\begin{aligned} y_t &= \Lambda f_t + \epsilon_t \\ &= \Lambda A A^{-1} f_t + \epsilon_t \\ &= \Lambda^* f_t^* + \epsilon_t \end{aligned}$$

where $\Lambda^* = \Lambda A$ and $f_t^* = A^{-1} f_t$.

- There does not exists a unique representation
- To pin down the factors we need r^2 restrictions

Identification

Some options

- Named factors

$$\Lambda = \begin{bmatrix} I_r \\ \Lambda_2 \end{bmatrix}$$

- Normalized factors

$$\Lambda'\Lambda / N = I_r \qquad FF' / T = D_r$$

  where D_r is a diagonal matrix

- Choleski type

$$\Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} \qquad FF' / T = I_r$$

  where Λ_1 is lower-triangular

Some comments

- Notice that these are statistic identification restrictions, in contrast with the structural identification restrictions discussed in lecture 6
- The appropriate identification scheme depends largely on the goal of the study
- For instance, for forecasting normalized restrictions are common, whereas for structural analysis named restrictions are more common
- Note that one can always rotate the factors ex-post to obtain the desired rotation

Estimation of common factors

Estimation methods

- Maximum likelihood via the Kalman filter
- Principal components analysis
- Hybrid methods

Likelihood based estimation

- Likelihood based estimation is a based on the state space representation of the factor model
- Estimation by maximum likelihood falls into two parts:
  - part 1: filtering, smoothing and prediction
  - part 2: parameter estimation

The dynamic factor model in state space representation

- We assume that the factors can be written in state space form

- The complete model can be written as

$$\begin{aligned} y_t &= \Lambda f_t + \epsilon_t & \epsilon_t &\sim NID(0, \Sigma_\epsilon) \\ f_t &= \Phi f_{t-1} + \epsilon_t & \epsilon_t &\sim NID(0, \Sigma_\eta) \end{aligned}$$

where the deterministic parameters are

- Λ: $N \times r$ loading matrix
- Σ_ϵ: $N \times N$ variance matrix of ϵ_t
- Σ_η: $r \times r$ variance matrix of η_t
- Φ: $r \times r$ transition matrix

Estimation of f_t

- We collect the deterministic model parameters in $\psi = \{\Lambda, \Phi, \Sigma_\epsilon, \Sigma_\eta\}$
- We treat these parameters as given when we consider computing
  - predictive estimates: $\hat{f}_{t|t-1} = \mathrm{E}(f_t|y_1, \ldots, y_{t-1})$, $P_{t|t-1} = \mathrm{Var}(f_t|y_1, \ldots, y_{t-1})$
  - filtered estimates: $\hat{f}_{t|t} = \mathrm{E}(f_t|y_1, \ldots, y_t)$, $P_{t|t} = \mathrm{Var}(f_t|y_1, \ldots, y_t)$
  - smoothed estimates: $\hat{f}_{t|T} = \mathrm{E}(f_t|y_1, \ldots, y_T)$, $P_{t|T} = \mathrm{Var}(f_t|y_1, \ldots, y_T)$
- These estimates are computed using the Kalman filter and smoothing recursions

Kalman filter recursions for dynamic factor model

$$
\begin{aligned}
v_t &= y_t - \Lambda \hat{f}_{t|t-1} \\
F_t &= \Lambda P_{t|t-1} \Lambda' + \Sigma_\epsilon \\
\hat{f}_{t|t} &= \hat{f}_{t|t-1} + P_{t|t-1} \Lambda' F_t^{-1} v_t \\
P_{t|t} &= P_{t|t-1} - P_{t|t-1} \Lambda' F_t^{-1} \Lambda P_{t|t-1} \\
f_{t+1|t} &= \Phi \hat{f}_{t|t-1} + K_t v_t \\
P_{t+1|t} &= \Phi P_{t|t-1} (\Phi - K_t \Lambda)' + \Sigma_\eta
\end{aligned}
$$

for $t = 1, \ldots, T$, where $K_t = \Phi P_{t|t-1} \Lambda' F_t^{-1}$.

Smoothing recursions for dynamic factor model

$$
\begin{aligned}
\hat{f}_{t|T} &= \hat{f}_{t|t-1} + P_{t|t-1} r_{t-1} \\
P_{t|T} &= P_{t|t-1} - P_{t|t-1} N_{t-1} P_{t|t-1} \\
r_{t-1} &= \Lambda' F_t^{-1} v_t + L_t' r_t \\
N_{t-1} &= \Lambda' F_t^{-1} \Lambda + L_t' N_t L_t
\end{aligned}
$$

for $t = T, T-1, \ldots, 1$, where $L_t = \Phi - K_t \Lambda$.

High dimensionality

- A concern is that the high dimension of the filter and smoother is too large to be handled computationally

To solve this consider the following GLS transformation

$$\tilde{y}_t = (\Lambda' \Sigma_\epsilon^{-1} \Lambda)^{-1} \Lambda' \Sigma_\epsilon^{-1} y_t$$

and note that $\tilde{y}_t$ is $r \times 1$ and given by

$$\begin{aligned} \tilde{y}_t &= f_t + \epsilon_t & \epsilon_t &\sim NID(0, (\Lambda' \Sigma_\epsilon^{-1} \Lambda)^{-1}) \\ f_t &= \Phi f_{t-1} + \epsilon_t & \epsilon_t &\sim NID(0, \Sigma_\eta) \end{aligned}$$

High dimensionality

The low dimensional model

$$\begin{aligned} \tilde{y}_t &= f_t + \epsilon_t & \epsilon_t &\sim NID(0, (\Lambda' \Sigma_\epsilon^{-1} \Lambda)^{-1}) \\ f_t &= \Phi f_{t-1} + \epsilon_t & \epsilon_t &\sim NID(0, \Sigma_\eta) \end{aligned}$$

- Contains all information for extracting the common factors
- Intuition, this works if $(\Lambda' \Sigma_\epsilon^{-1} \Lambda)^{-1} \Lambda' \Sigma_\epsilon^{-1}$ is full rank
- Thus, filtering and smoothing can be done based on the model for $\tilde{y}_t$ only

Parameter estimation by maximum likelihood

We know that the log likelihood can be written as

$$\ell(y;\psi) = -\frac{nT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\left(\log\det F_t + v_t' F_t^{-1} v_t\right)$$

and the ML estimates are obtained by

$$\hat{\psi} = \arg\max_{\psi} \ell(y;\psi)$$

- In its current version the log likelihood requires the output of the Kalman filter for the complete model
- Can we also compute the log likelihood based the low-dimensional model?

Parameter estimation by maximum likelihood

Turns out yes

$$\ell(y;\psi) = \ell(\tilde{y};\psi) - \frac{T}{2}\log|\Sigma_\epsilon| - \frac{T}{2}\log|(\Lambda'\Sigma_\epsilon^{-1}\Lambda)| - \frac{1}{2}\sum_{t=1}^{T} e_t'\Sigma_\epsilon^{-1}e_t$$

where e_t is GLS residual

$$e_t = y_t - (\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1}\Lambda'\Sigma_\epsilon^{-1}y_t$$

- Importantly, the evaluation of the likelihood requires: (i) the likelihood of the low-dimensional model for $\tilde{y}_t$ and (ii) the residuals from the GLS transformation
- Note that $\ell(\tilde{y};\psi)$ is computed using the prediction error decomposition that is obtained from the Kalman filter for the low-dimensional model

Some comments

- The log likelihood can be optimized with respect to the parameter vector ψ using numerical methods
- In practice
  - Pick starting values for parameters
  - Do GLS transformation
  - Run Kalman filter for low-dimensional model
  - Evaluate the log likelihood

Pros and Cons

- Pro maximum likelihood
  - Flexible way of incorporating rich dynamics in the factor process
  - Easy to handle missing values using the Kalman filter
- Con maximum likelihood
  - Requires more specification choices
  - Maximizing the likelihood wrt to parameters is computationally demanding

Non-parametric averaging

- An alternative class of estimators is based on taking weighted averages of the observations
- Such estimators are computationally less expensive and require less specification choices

Model details

$$y_t = \Lambda f_t + \epsilon_t$$

- Both Λ and F are treated as deterministic model parameters
- ϵ_t is the idiosyncratic component

Assumptions

1. $N^{-1}\Lambda'\Lambda \overset{N\to\infty}{\to} D_\lambda$, with D_λ full rank
2. $\text{MaxEval}(\Sigma_\epsilon) \leq c < \infty$ for some constant $c > 0$ and all N[1]

[1] $\text{MaxEval}(A)$ denotes the largest eigenvalue of the matrix A.

Discussion assumptions

- $N^{-1}\Lambda'\Lambda \overset{N\to\infty}{\to} D_\lambda$, with D_λ full rank: implies that the influence of the factors is not vanishing when N get large. In other words, the assumption ensures that we can find the factors when we take averages of the data.
- $\text{MaxEval}(\Sigma_\epsilon) \leq c < \infty$ requires the largest eigenvalue of Σ_ϵ to be bounded. This limits the correlation between $\epsilon_{i,t}$ and $\epsilon_{j,t}$ and ensures that there are no "factors" in the idiosyncratic components

Intuition for why averaging works

- Let W be a deterministic $N \times r$ weight matrix that is normalized such that $N^{-1}W'W = I_r$

Consider the estimator for the factors (as a function of the weights)

$$\hat{f}_t(N^{-1}W) = N^{-1}W'y_t$$

- This is like taking r weighted averages of the data
- Claim: if $N^{-1}W'\Lambda \overset{N\to\infty}{\to} H$ with H full rank we can consistently estimate the space spanned by the factors

Intuition for why averaging works

We have

$$\hat{f}_t(N^{-1}W) = N^{-1}W'y_t = N^{-1}W'\Lambda f_t + N^{-1}W'\epsilon_t$$

and

- $N^{-1}W'\Lambda f_t \to Hf_t$
- Consider the variance of the jth element of $N^{-1}W'\epsilon_t$

$$\begin{aligned}
\mathrm{Var}(N^{-1}W_j'\epsilon_t) &= N^{-2}W_j'\mathrm{Var}(\epsilon_t)W_j \\
&= N^{-2}W_j'\Sigma_\epsilon W_j \\
&\leq N^{-1}W_j'W_j N^{-1}\mathrm{MaxEval}(\Sigma_\epsilon) \\
&\leq N^{-1}c \to 0
\end{aligned}$$

hence as the variance tends to zero we have $N^{-1}W'\epsilon_t \xrightarrow{p} 0$.

Intuition for why averaging works

Hence we can show that

$$\hat{f}_t(N^{-1}W) = N^{-1}W'y_t \xrightarrow{p} Hf_t$$

- Since H is full rank we can consistently estimate a linear combination of the factors
- In most applications this is sufficient
- So far we have not discussed which weights W can be used (apart from that we must have $N^{-1}W'\Lambda \to H$ with H full rank)
- In fact many choices for the weights exist: we discuss the choice that principal components analysis makes

Principal components analysis

$$W = \hat{\Lambda}$$

where $\hat{\Lambda}$ are the eigenvectors that correspond to the r largest eigenvalues of $\hat{\Sigma}_y = \frac{1}{T}\sum_{t=1}^{T} y_t y_t'$

- This choice for the weights can also be viewed as the solution of the following minimization problem

$$\min_{\Lambda, f_1, \ldots, f_T} V_r(\Lambda, F) \qquad V_r(\Lambda, F) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \Lambda f_t)'(y_t - \Lambda f_t)$$

  such that $\frac{1}{N}\Lambda'\Lambda = I_r$.

Rewriting the minimization problem

$$\min_{\Lambda, f_1, \ldots, f_T} V_r(\Lambda, F) \qquad V_r(\Lambda, F) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \Lambda f_t)'(y_t - \Lambda f_t)$$

when Λ is observed we have

$$\hat{f}_t = (\Lambda'\Lambda)^{-1}\Lambda' y_t$$

and we use this to concentrate the minimization problem

$$\min_{\Lambda} V_r^c(\Lambda) \qquad V_r^c(\Lambda) = \frac{1}{T} \sum_{t=1}^{T} y_t'(I_T - \Lambda(\Lambda'\Lambda)^{-1}\Lambda')y_t$$

which becomes

$$\max_{\Lambda} \mathrm{Tr} \left((\Lambda'\Lambda)^{-1/2'} \Lambda' \frac{1}{T} \sum_{t=1}^{T} y_t y_t' \Lambda (\Lambda'\Lambda)^{-1/2} \right)$$

Rewriting the minimization problem

Finally, the solution for Λ in

$$\max_{\Lambda} \text{Tr}\left((\Lambda'\Lambda)^{-1/2'} \Lambda' \frac{1}{T} \sum_{t=1}^{T} y_t y_t' \Lambda (\Lambda'\Lambda)^{-1/2} \right)$$

is equivalent to

$$\max_{\Lambda} \Lambda' \hat{\Sigma}_y \Lambda \qquad \text{s.t.} \quad N^{-1}\Lambda'\Lambda = I_r$$

Hence based on this choice for the weights[2] we have

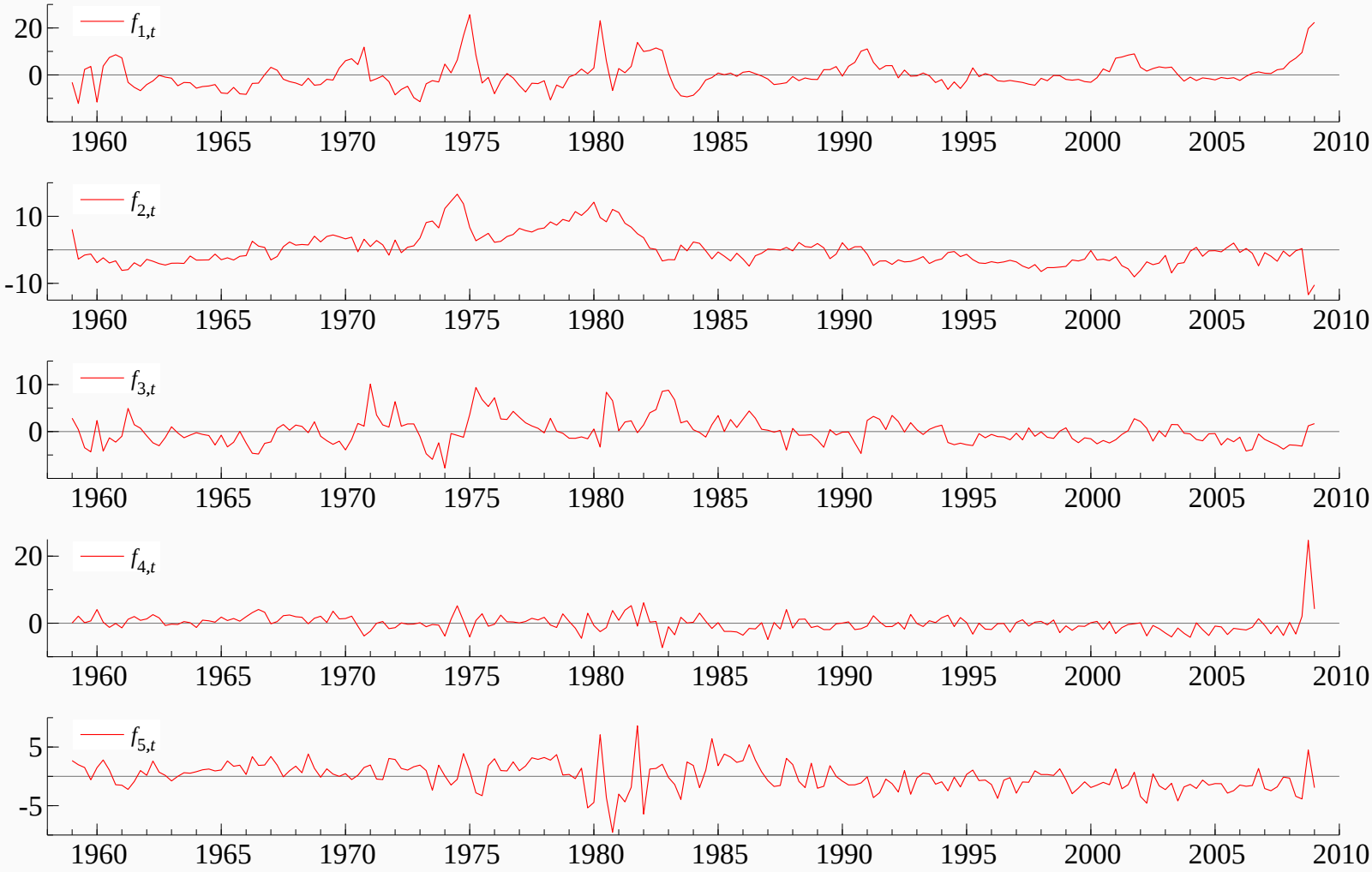
$$\hat{f}_t^{pca} = \hat{\Lambda}' y_t \qquad t = 1, \ldots, T.$$

[2] see stock watson 2011 on box for other examples of weights

Pros and Cons non-parametric averaging

- Pro averaging
  - Computationally inexpensive; can handle very very high dimensional data
  - Consistent under mild assumptions (considered very robust)
- Con averaging
  - No dynamics are exploited
  - And needs large N cannot work for N small

US macroeconomic factors



Correlations

Hybrid methods

- It is natural to ask whether some of the pros of maximum likelihood and principal components analysis can be combined
- This has led to the development of so-called hybrid methods
- They try to combining the computational advantage of principal components with the dynamic approach of state space methods

Hybrid methods

- Step 1: Compute $\hat{\Lambda}$ as the eigenvectors that correspond to the r largest eigenvalues of $\hat{\Sigma}_y$
- Step 2: Consider

$$\begin{aligned} y_t &= \hat{\Lambda} f_t + \epsilon_t & \epsilon_t &\sim NID(0, \Sigma_\epsilon) \\ f_t &= \Phi f_{t-1} + \epsilon_t & \epsilon_t &\sim NID(0, \Sigma_\eta) \end{aligned}$$

  This model can be studied using likelihood based methods. But it is easy to see that the parameter dimension now is much smaller. Hence it is computationally easier and can be combined with the dimension reduction discussed before.

Determining the number of factors

Determining r

1. So far we have assumed that r is known
2. In practice we need to choose r
    - sometimes is obvious, e.g. business cycle example
    - sometimes its hard, e.g. when forecasting
3. Goal is to estimate r consistently

Determining r

1. Information criteria
2. Eigenvalue criteria and the scree-plot
3. Out-of-sample forecasting based

Information criteria

- In a seminal contribution Bai & Ng (2002) proposed information criteria to select the number of common factors
- They show that standard criteria like AIC and BIC do not work for both N and T large
- Instead they propose modified penalty functions that depend on both N and T which can be used to consistently determine the number of common factors

Information criteria

- Recall the criterion function from principal components analysis[3]

$$\min_{\Lambda, f_1, \ldots, f_T} V_r(\Lambda, F) \qquad V_r(\Lambda, F) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \Lambda f_t)'(y_t - \Lambda f_t)$$

- This function depends on r in the sense that $V_r(\Lambda, F)$ becomes smaller when we increase the number of factors.
- The idea is to penalize this function for larger r

[3] For maximum likelihood one can replace $V_r(\Lambda, F)$ by the likelihood.

Information criteria

- Bai & Ng (2002) propose different criteria
- The following works well in practice

$$IC(r) = \ln V_r(\hat{\Lambda}, \hat{F}) + rg(N, T)$$

where

- $\hat{\Lambda}, \hat{F}$ are the principal components estimates
- $g(N, T)$ is a penalty function that satisfies $g(N, T) \to 0$ but $\min(N, T)g(N, T) \to \infty$

Information criteria

- Any $g(N, T)$ such that $g(N, T) \to 0$ but $\min(N, T)g(N, T) \to \infty$ will lead to a consistent estimate for the number of factors
- The following works reasonably well in practice

$$g(N, T) = (N + T)\ln(\min(N, T))/(NT)$$

- And we may select r from

$$\begin{aligned} \hat{r} &= \min_r IC_2(r) \\ & IC_2(r) = \ln V_r(\hat{\Lambda}, \hat{F}) + r(N + T)\ln(\min(N, T))/(NT) \end{aligned}$$

Number of factors US macroeconomics

r	IC2
1.0000	4.1033
2.0000	3.2928
3.0000	2.6746
4.0000	2.6661
5.0000	2.7194
6.0000	2.7201
7.0000	2.6250
8.0000	2.6346
9.0000	2.6652

- Unfortunately, results are sensitive to the choice of $g(N, T)$

Eigenvalue ratios

- An alternative for determining the number of factors is studying the eigenvalues of $\Sigma_y = \mathrm{Var}(y_t)$
- Note that $\Sigma_y = \Lambda \Sigma_f \Lambda' + \Sigma_\epsilon$ and
  - $\Lambda \Sigma_f \Lambda'$ has r non-zero eigenvalues which tend to infinity for N large (follows from $N^{-1}\Lambda'\Lambda \overset{N\to\infty}{\to} D_\lambda$)
  - Σ_ϵ has bounded eigenvalues, e.g. $\mathrm{MaxEval}(\Sigma_\epsilon) \leq c < \infty$
- In total r eigenvalues of Σ_y should diverge if there are r factors
- This forms the main idea that underlies the eigenvalue ratio statistics of Onatski (2011) and Ahn & Horenstein (2013)

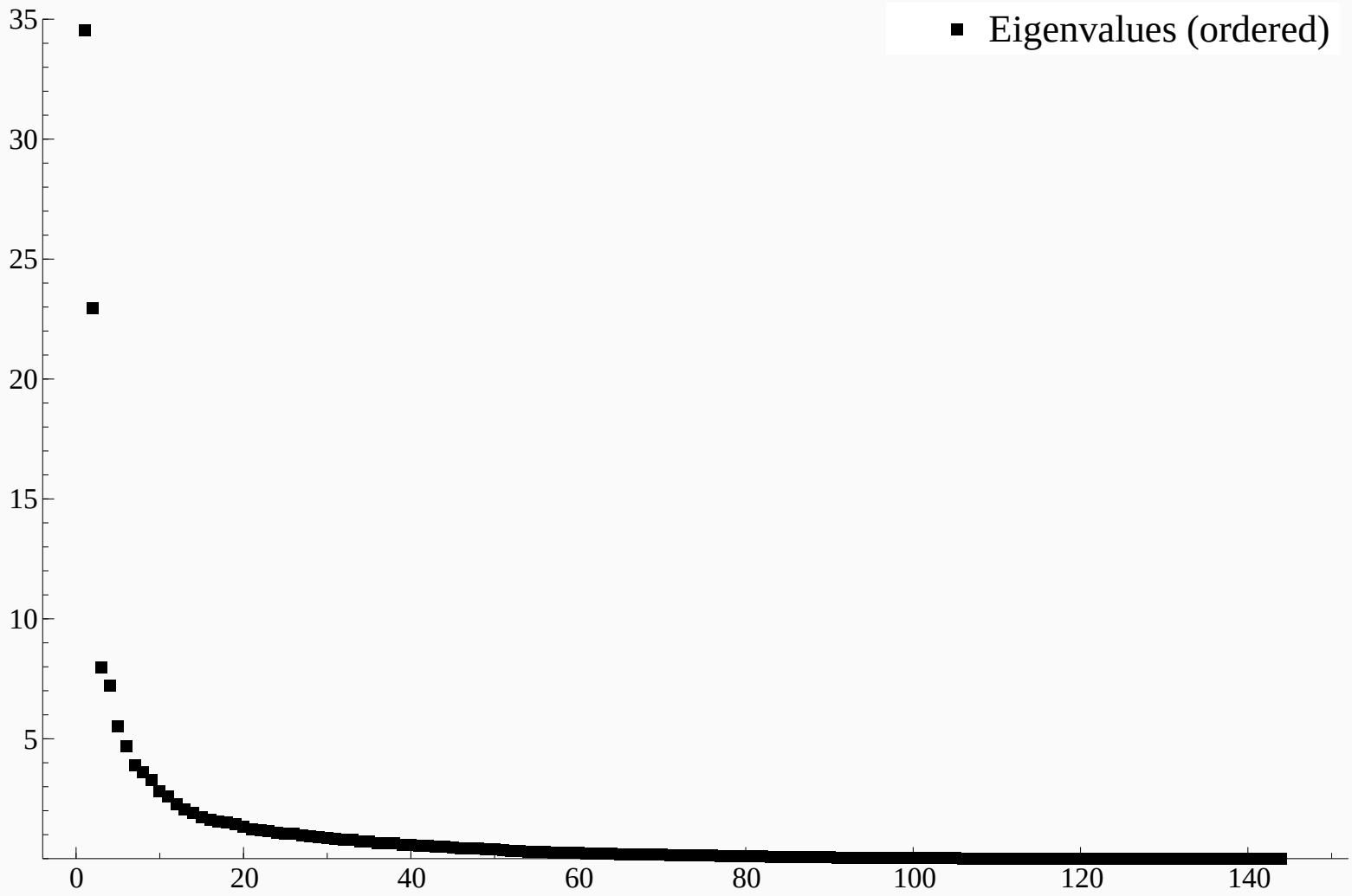
Ahn & Horenstein eigenvalue ratio

Let $\hat{\mu}_j(\hat{\Sigma}_y)$ denote the jth largest eigenvalue of $\hat{\Sigma}_y$, the eigenvalue ratio estimator is given by

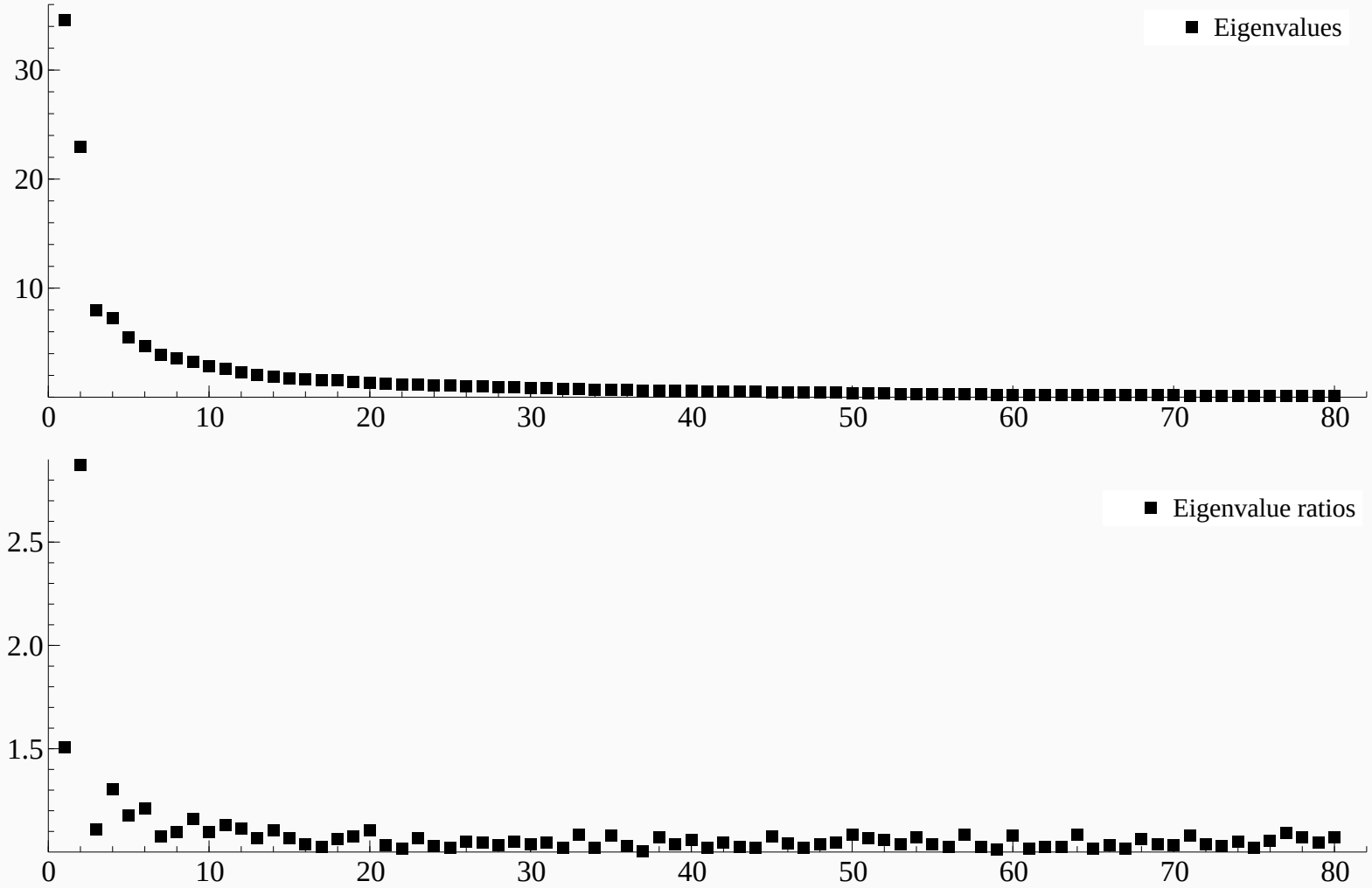
$$\hat{r} = \arg\max_{j \in \{1,2,\ldots,N\}} \hat{\mu}_j(\hat{\Sigma}_y) / \hat{\mu}_{j+1}(\hat{\Sigma}_y)$$

- Intuition: only the ratio between the rth largest eigenvalue and the $r+1$th largest eigenvalue should be large.

Eigenvalues US macro panel



Eigenvalues US macro panel



Out-of-sample forecasting

Forecasting

- Including common factors in standard time series models typically improves the out-of-sample forecasts

Let x_t be the series of interest, we consider direct forecasting based on

$$x_{t+h} = \beta^h f_t + \gamma^h w_t + \eta_{t+h}$$

where

$$y_t = \Lambda f_t + \epsilon_t$$

- Idea: obtain f_t from the large panel and use it to predict x_{t+h}

Out-of-sample forecasting

- Take y_t as US unemployment
- We split the sample in 1985 and take $h = 1$
- Forecast using rolling window approach
- Evaluate using mean-squared error loss
- Compare AR(2) to AR(2)+factors

Out-of-sample forecasting

r	MSE
0	1.23
1	0.99*
2	0.98*
3	0.88*
4	0.93*
5	1.03

- Factors 1 and 3 correlate the most with the real economy series
- It seems optimal to include several factors

References & Material

Reference & Material

- References:
  J. Stock & M. Watson (2016) Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics, in Chapter 8 of Handbook of Macroeconomics, Volume 2A