Forecasting With Many predictors

The Econometrics of Predictability

This version: June 15, 2014

June 15, 2014

UNIVERSITY OF
OXFORD

Dynamic Factor Models

Dynamic Factor Models

- Dynamic factors model specify dynamics in the factors
- Basic DFM is

$$
\begin{aligned}
\mathbf{x}_t &= \sum_{i=0}^{s} \boldsymbol{\Phi}_i \mathbf{f}_t + \boldsymbol{\epsilon}_t \\
\mathbf{f}_t &= \sum_{j=1}^{q} \boldsymbol{\Psi} \mathbf{f}_{t-j} + \boldsymbol{\eta}_t
\end{aligned}
$$

3 x 1

I(0)

- Observed data depend on contemporaneous and lagged factors
- Factors have VAR-like dynamics
- Assumed that $\mathbf{f}_t$ and $\boldsymbol{\epsilon}_t$ are stationary, so $\mathbf{x}_t$ is also stationary
  - Important: must transform series appropriately when applying to data
- $\boldsymbol{\epsilon}_t$ can have weak dependence in both the cross-section and time-series
- $\mathrm{E}\left[\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_s\right] = \mathbf{0}$ for all t, s

Optimal Forecast from DFM

MA(q)

$$\mathbf{x}_t = \sum_{i=0}^{s} \boldsymbol{\Phi}_i \mathbf{f}_{t-i} + \epsilon_t, \quad \mathbf{f}_t = \sum_{j=1}^{q} \boldsymbol{\Psi} \mathbf{f}_{t-j} + \boldsymbol{\eta}_t$$

- Optimal forecast can be derived

$$\begin{aligned}
\mathrm{E}\left[x_{it+1} | \mathbf{x}_t, \mathbf{f}_t, \mathbf{x}_{t-1}, \mathbf{f}_{t-1}, \ldots\right] &= \mathrm{E}\left[\sum_{i=0}^{s} \boldsymbol{\phi}_i \mathbf{f}_{t+1-i} + \epsilon_{it+1} \Big| \mathbf{x}_t, \mathbf{f}_t, \mathbf{x}_{t-1}, \mathbf{f}_{t-1}, \ldots\right] \\
&= \mathrm{E}_t\left[\sum_{i=0}^{s} \boldsymbol{\phi}_i \mathbf{f}_{t+1-i}\right] + \mathrm{E}_t\left[\epsilon_{it+1}\right] \\
&= \sum_{i=1}^{s'} \mathbf{A}_i f_{t-i+1} + \sum_{j=1}^{n} \mathbf{B}_j x_{it-j+1}
\end{aligned}$$

- Predictability in both components
    - Lagged factors predict factors
    - Lagged x_{it} predict ϵ_{it}

Invertibility and MA processes

UNIVERSITY OF OXFORD

- DFM is really factors plus moving average
- Moving average processes can be replaced with AR processes when invertible

y_{t-1}

$$
\begin{aligned}
y_t &= \epsilon_t + \theta \epsilon_{t-1} \\
y_t - \theta y_{t-1} &= \epsilon_t + \cancel{\theta \epsilon_{t-1}} - \cancel{\theta}(\theta \epsilon_{t-2} + \cancel{\epsilon_{t-1}}) \\
&= \epsilon_t - \theta^2 \epsilon_{t-2} \\
y_t - \theta y_{t-1} + \theta^2 y_{t-2} &= \epsilon_t - \cancel{\theta^2 \epsilon_{t-2}} + \theta^2 (\theta \epsilon_{t-3} + \cancel{\epsilon_{t-2}}) \\
&= \epsilon_t + \theta^2 (\theta \epsilon_{t-3} + \cancel{\epsilon_{t-2}}) \\
\sum_{i=0}^{\infty} (-\theta)^i y_{t-i} &= \epsilon_t \\
y_t &= \sum_{i=1}^{\infty} -(-\theta)^i y_{t-i} + \epsilon_t
\end{aligned}
$$

y_{t-2}

$\epsilon_t + \theta^3 \epsilon_{t-3}$

AR(∞)

- Can approximate finite MA with finite AR
- Quality will depend on the persistence of the MA component

Dynamic as Static Factor Models

- Superficially dynamic factor models appear to be more complicated than static factor models
- Dynamic Factor models can be directly estimated using Kalman Filter or spectral estimators that account for serial correlation in factors
  - Latter are not useful for forecasting since 2-sided
- (Big) However, DFM can be converted to Static model by relabeling
- In DFM, factors are

$$[\mathbf{f}_t, \mathbf{f}_{t-1}, \ldots, \mathbf{f}_{t-s}]$$

3

  - Total of $r\,(s+1)$ factors in model
- Equivalent to static model with *at most* $r\,(s+1)$ factors
  - Redundant factors will not appear in static version

Dynamic as Static Factor Models

UNIVERSITY OF OXFORD

- Consider basic DFM

$$
\begin{aligned}
x_{it} &= \phi_{i1} f_t + \phi_{i2} f_{t-1} + \epsilon_{it} \\
f_t &= \psi f_{t-1} + \eta_t
\end{aligned}
$$

- Model can be expressed as

$$
\begin{aligned}
x_{it} &= \phi_{i1} \left(\psi f_{t-1} + \eta_t \right) + \phi_{i2} f_{t-1} + \epsilon_{it} \\
&= \phi_{i1} \eta_t + \phi_{i2} \left(1 + \left(\phi_{i1}/\phi_{i2}\right) \psi\right) f_{t-1} + \epsilon_{it}
\end{aligned}
$$

- One version of static factors are η_t and f_{t-1}
  - In this particular version, η_t is not "dynamic" since it is WN
  - f_{t-1} follows an AR(1) process
- Other *rotations* will have different dynamics

Dynamic as Static Factor Models

UNIVERSITY OF OXFORD

$z_t = (x_t - \mu) \oslash \sigma$

- Basic simulation

$$
\begin{aligned}
x_{it} &= \phi_{i1} f_t + \phi_{i2} f_{t-1} + \epsilon_{it} \\
f_t &= \psi f_{t-1} + \eta_t
\end{aligned}
$$

- $\phi_{i1} \sim N(1,1), \phi_{i2} \sim N(.2,1)$
  - Smaller signal makes it harder to find second factor
- $\psi = 0.5$
  - Higher persistence makes it harder since $\mathrm{Corr}\left[f_t, f_{t-1}\right]$ is larger
- Everything else standard normal
- $k = 100$, $T = 100$
  - Also $k = 200$ and $T = 200$ (separately)
- All estimation using PCA on correlation

Number of Factors for Forecasting

Better to have r above $r^\star$ than below

Measuring Closeness of Estimate

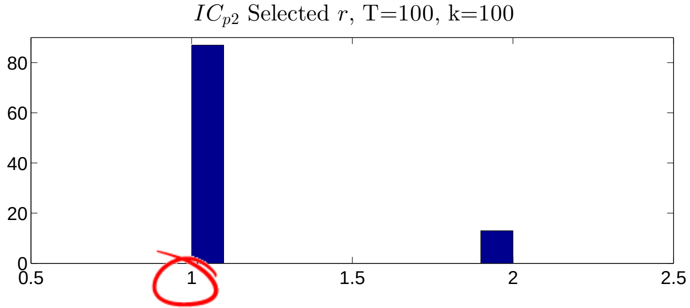
- Factors are not point identified
  - Can use an arbitrary rotation and model is equivalent
- Natural measure of similarity between original (GDP) factors and estimated factors is global R^2

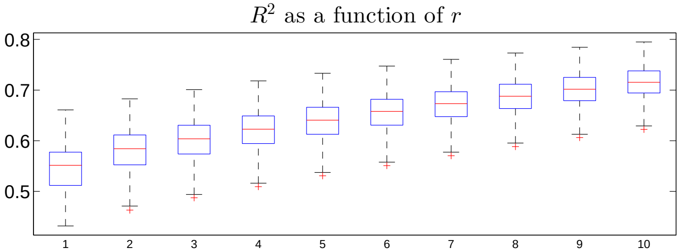
Pc Sim

$$
\begin{aligned}
\hat{\mathbf{f}}_t &= \mathbf{A}\mathbf{f}_t + \boldsymbol{\eta}_t \\
R^2 &= 1 - \frac{\sum_{t=1}^{T} \hat{\boldsymbol{\eta}}_t' \hat{\boldsymbol{\eta}}_t}{\sum_{t=1}^{T} \mathbf{f}_t' \mathbf{f}_t}
\end{aligned}
$$

- Note that $\mathbf{A}$ is a 2 by 2 matrix of regression coefficients

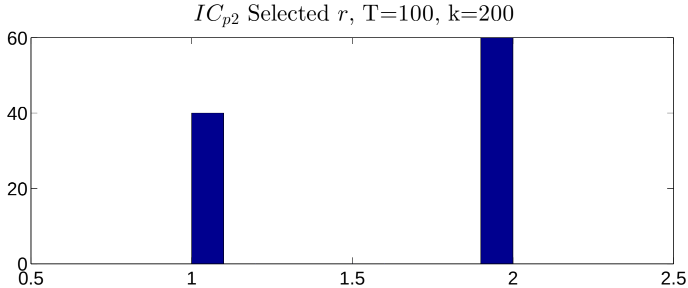
Dynamic as Static Factor Models

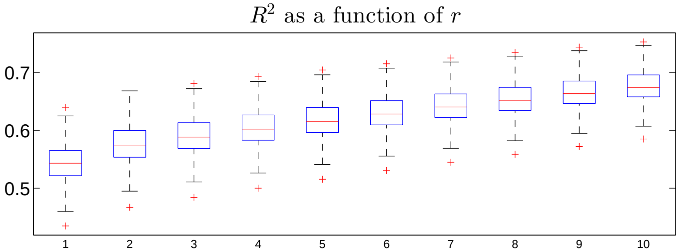




Dynamic as Static Factor Models

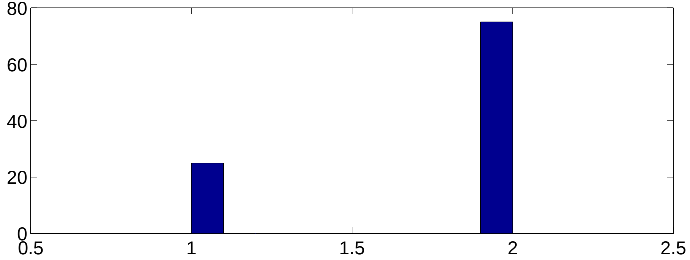
UNIVERSITY OF OXFORD



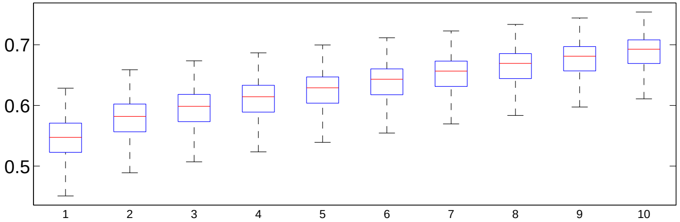


Dynamic as Static Factor Models

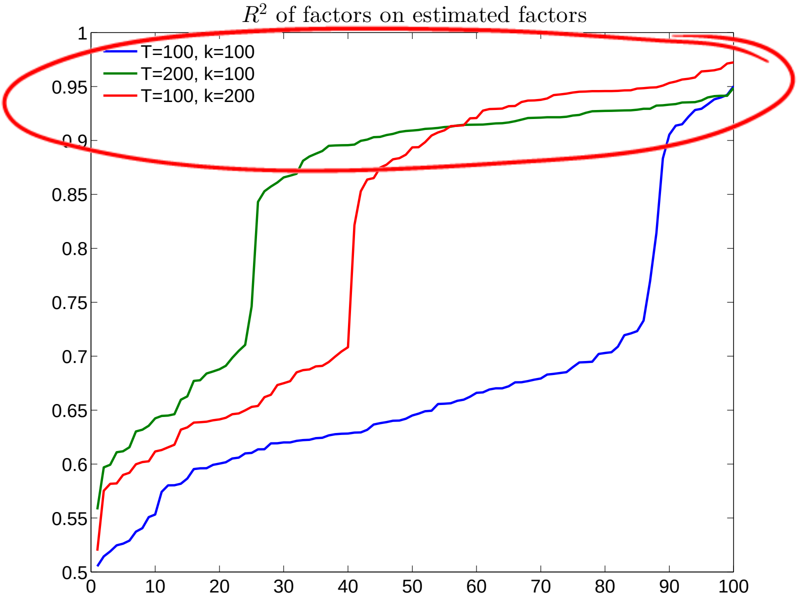
IC_{p2} Selected r, T=200, k=100



R^2 as a function of r



Dynamic as Static Factor Models



Stock and Watson's DFM Data

Stock & Watson (2012) Data

- Stock & Watson have been at the forefront of factor model development
- Data is from 2012 paper "Disentangling the Channels of the 2007-2009 Recession"
- Dataset consists of 137 monthly and 74 quarterly series
  - Not all used for factor estimation
  - Aggregates not used if disaggregated series available
- Monthly series are aggregated to quarterly, which is frequency of data
- Series with missing observations are dropped for simplicity
  - Before dropping those with missing values data set has 132 series
  - After 107 series remain

$$\underline{X} = \underline{A} + \underline{B} + \underline{C}$$

The series

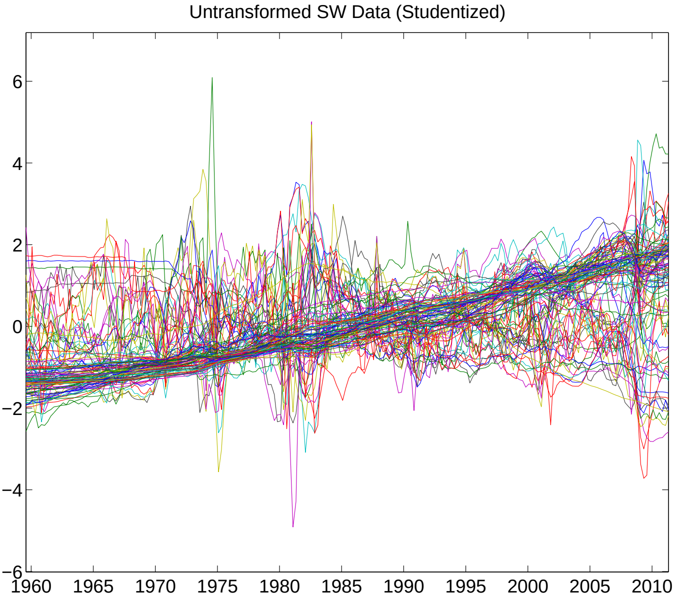
National Income and Product Accounts (NIPA)	12
Industrial Production	9
Employment and Unemployment	30
Housing Starts	6
Inventories, Orders, and Sales	7
Prices	25
Earnings and Productivity	8
Interest Rates	10
Money and Credit	6
Stock Prices, Wealth, Household Balance Sheets	8
Housing Prices	3
Exchange Rates	6
Other	2

Data Transformation

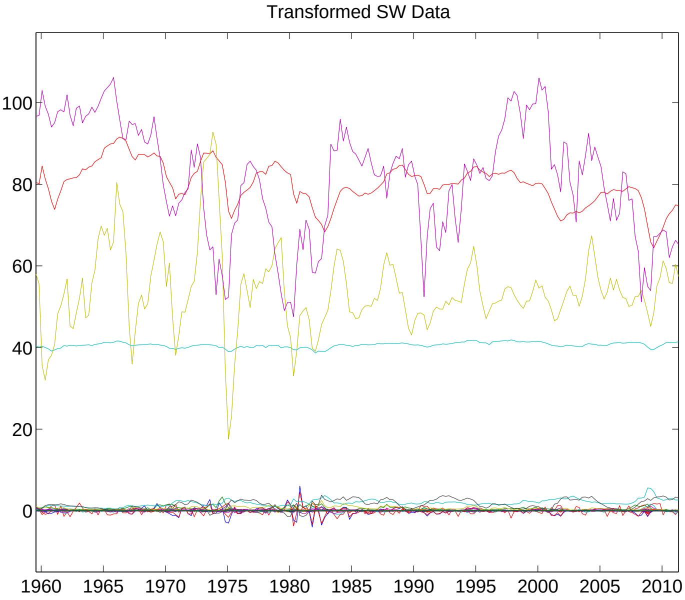
UNIVERSITY OF OXFORD

- Monthly series were aggregated to quarterly using
  - Average
  - End-of-quarter
- All series were transformed to be stationary using one of:
  - No transform
  - Difference
  - Double-difference
  - Log
  - Log-difference
  - Double-log-difference $\Delta^2 \ln X$
- Most series checked for outliers relative to IQR (rare)
- Final series were Studentized in estimation of PC

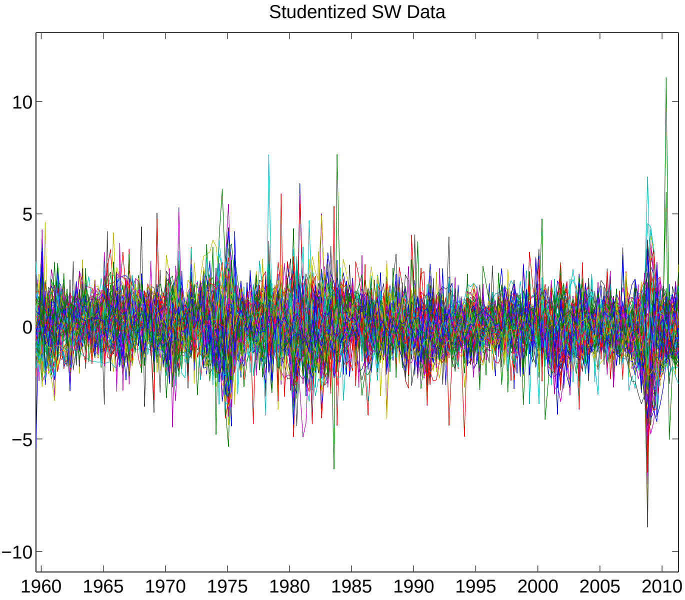
Raw Data Before Transform



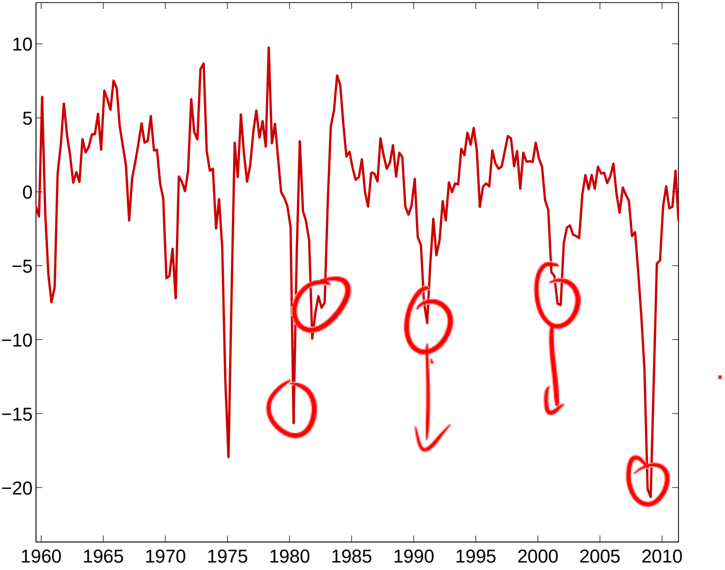
Raw Data after Transform



Studentized Data

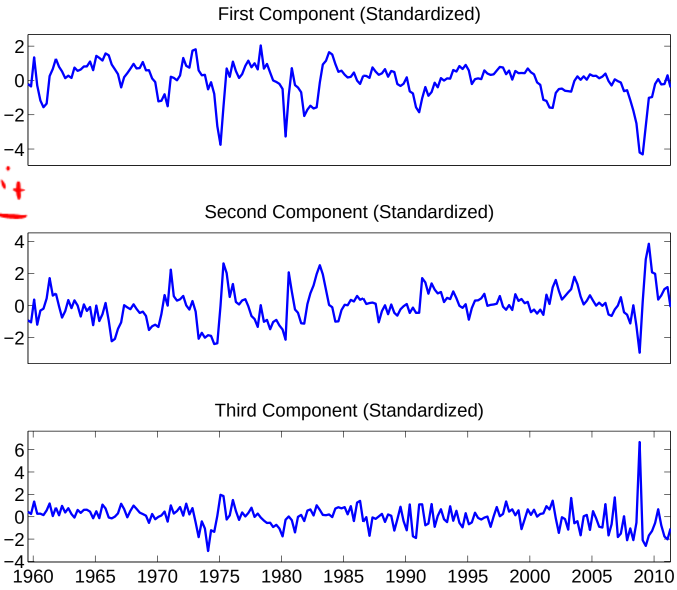


First Component

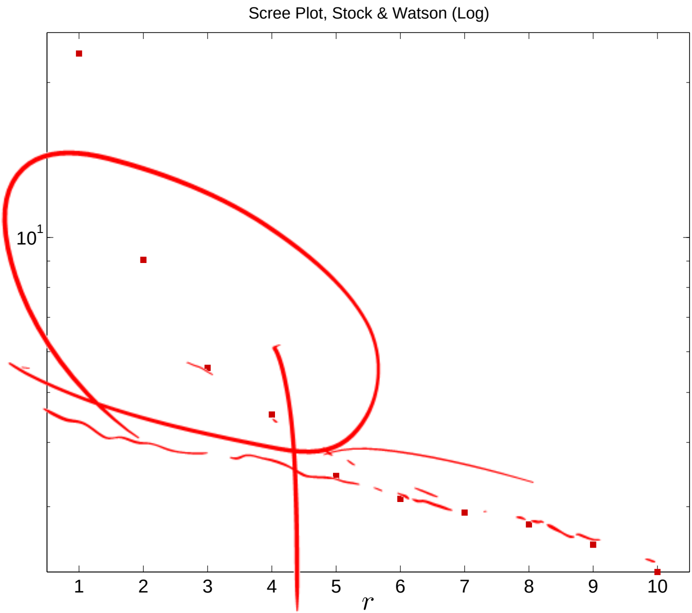


First Three Components

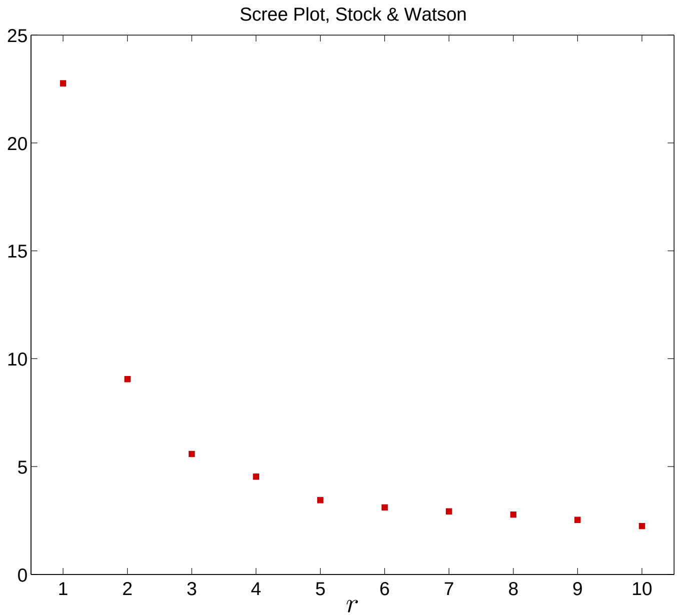
$f_t = \sum w_i x_{it}$



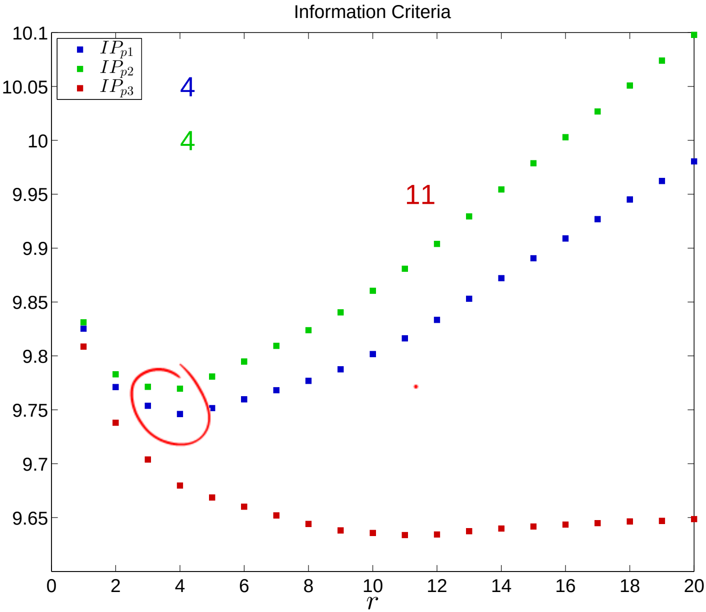
Scree Plot (Log)



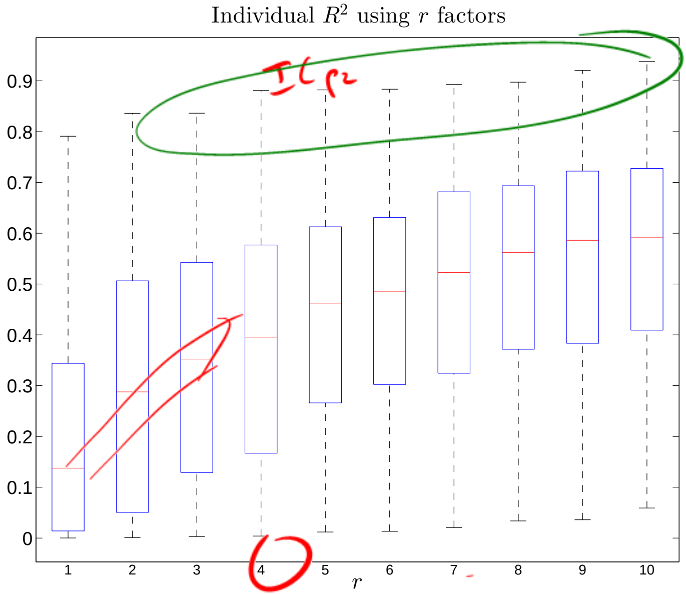
Scree Plot



Information Criteria



Individual Fit against r



Forecasting

Forecast Methods

UNIVERSITY OF OXFORD

- Forecast problem is not meaningfully different from standard problem
- Interest is now in $\mathbf{y}_t$, which may or may not be in $\mathbf{x}_t$
  - Note that stationary version of $\mathbf{y}_t$ should be forecast, e.g. $\Delta\mathbf{y}_t$ or $\Delta^2\mathbf{y}_t$
- Two methods to forecast

Unrestricted

$$y_{t+1} = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i+1} + \underbrace{\boldsymbol{\theta}'\hat{\mathbf{f}}_t}_{r \times 1} + \epsilon_{it}$$

X_t

AR-X

- Treats factors as observed data, only makes sense if k is large
  - Uses an AR(P) to model residual dependence
  - Choice of number of factors to use, may be different from r
  - Can also use lags of $\mathbf{f}_t$ (uncommon)
  - Model selection is applicable as usual, e.g. BIC

Restricted

- When $\mathbf{y}_t$ is in $\mathbf{x}_t$, $\mathbf{y}_t = \boldsymbol{\beta}\hat{\mathbf{f}}_t + \epsilon_t$

$$\epsilon_t = \mathbf{y}_t - \boldsymbol{\beta}\hat{\mathbf{f}}_t$$

VAR +

$$\begin{aligned} \hat{y}_{t+1|t} &= \boldsymbol{\beta}\hat{\mathbf{f}}_{t+1|t} + \sum_{i=1}^{p} \phi_i \left(y_{t-i+1} - \boldsymbol{\beta}\hat{\mathbf{f}}_{t-i+1} \right) \\ &= \boldsymbol{\beta}\hat{\mathbf{f}}_{t+1|t} + \sum_{i=1}^{p} \phi_i \hat{\epsilon}_{t-i+1} \end{aligned}$$

ϵ_{t-i+1}

- VAR to forecast $\hat{\mathbf{f}}_{t+1}$ using lags of $\hat{\mathbf{f}}_t$
- Univariate AR for $\hat{\epsilon}_t$
- Usually found to be less successful than unrestricted
- Care is needed when using studentized data since forecasting recentered, rescaled version of y

Re-integrating forecasts

UNIVERSITY OF OXFORD

- When forecasting $\Delta \mathbf{y}_t$,

$$
\begin{aligned}
\mathrm{E}_t[\mathbf{y}_{t+1}] &= \mathrm{E}_t[\mathbf{y}_{t+1} - \mathbf{y}_t + \mathbf{y}_t] \\
&= \underline{\mathrm{E}_t[\Delta \mathbf{y}_{t+1}] + \mathbf{y}_t}
\end{aligned}
$$

- At longer horizons,

$$
\mathrm{E}_t[\mathbf{y}_{t+h}] = \sum_{i=1}^{h} \mathrm{E}_t[\Delta \mathbf{y}_{t+i}] + \mathbf{y}_t
$$

- When forecasting $\Delta^2 \mathbf{y}_t$

$$
\begin{aligned}
\mathrm{E}_t[\mathbf{y}_{t+1}] &= \mathrm{E}_t[\mathbf{y}_{t+1} - \mathbf{y}_t - \mathbf{y}_t + \mathbf{y}_{t-1} + 2\mathbf{y}_t - \mathbf{y}_{t-1}] \\
&= \mathrm{E}_t\left[\Delta^2 \mathbf{y}_{t+1}\right] + 2\mathbf{y}_t - \mathbf{y}_{t-1}
\end{aligned}
$$

  - In many cases interest is in $\Delta \mathbf{y}_t$ when forecasting $\Delta^2 \mathbf{y}_t$
    - For example CPI, inflation and change in inflation
    - Same as reintegrating Δy_t to y_t

UNIVERSITY OF OXFORD

- Multistep can be constructed using either method
- Unrestricted requires additional VAR for $\hat{\mathbf{f}}_t$
- Alternative use direct forecasting

t+4   t, t-1, . . .

$$y_{t+h|t} = \hat{\phi}_{(h)0} + \sum_{i=1}^{p^h} \hat{\phi}_{(h)i} y_{t-i+1} + \hat{\boldsymbol{\theta}}'_{(h)} \hat{\mathbf{f}}_t$$

$$\begin{pmatrix} y_{t+1} \\ y_{t+2} \\ y_{t+3} \\ y_{t+4} \end{pmatrix}$$

  - (h) used to denote explicit parameter dependence on horizon
  - $y_{t+h|t}$ can be either the period-h value, or the h-period cumulative forecast (more common)
- Direct has been documented to be better than iterative in DFMs
  - Problem dependent

$$\sum_{j=1}^{h} y_{t+j}$$

“Forecasting”

- Used BIC search across models
- 3 setups
  - GDP lags only (4), Components Only (6), Both

$$\sum_{j=1}^{h} \Delta g_{t+j} = \phi_0 + \sum_{s=1}^{4} \gamma_s \Delta g_{t-s+1} + \sum_{n=1}^{6} \psi_n f_{jt} + \epsilon_{ht}$$

	GDP Only	R^2	Components Only	R^2	Both: GDP	Both: Components	Both: R^2
$h=1$	1, 2, 4	.517	1, 2, 3, 4, 6	.662	1	1, 2, 3, 4, 6	.686
$h=2$	1, 4	.597	1, 2, 3, 4, 6	.763	1	1, 2, 3, 4, 6	.771
$h=3$	1, 4	.628	1, 2, 3, 4, 6	.785	1	1, 2, 3, 4, 6	.792
$h=4$	1, 4	.661	1, 2, 3, 4, 6	.805	–	1, 2, 3, 4, 6	.805

Improving Estimated Components

Generalized Principal Components

UNIVERSITY OF OXFORD

- Basic PCA makes use of the covariance or more commonly correlation
- Correlation is technically a special case of *generalized PCA* $X_{it} = \theta f_t + \epsilon_{it}$

$$\min_{\boldsymbol{\beta},\mathbf{f}_t,\ldots\mathbf{f}_t} \sum_{t=1}^{T} \left(\mathbf{x}_t - \boldsymbol{\beta}\mathbf{f}_t\right)' \overset{I}{\boldsymbol{\Sigma}_{\epsilon}^{-1}} \left(\mathbf{x}_t - \boldsymbol{\beta}\mathbf{f}_t\right) \text{ subject to } \boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$$

- Clever choices of $\boldsymbol{\Sigma}_{\epsilon}$ lead to difference estimators
  - Using $\operatorname{diag}\left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2\right)$ where $\hat{\sigma}_j^2$ is variance of x_j leads to correlation
  - Tempting to use GLS version based on r principal components

Algorithm (Principal Component Analysis using GLS)

1. *Estimate* $\hat{\epsilon}_{it} = x_{it} - \hat{\boldsymbol{\beta}}_i \hat{\mathbf{f}}_t$ *using r factors*
2. *Estimate* $\hat{\sigma}_{\epsilon i}^2 = T^{-1} \sum \hat{\epsilon}_{it}^2$ *and* $\mathbf{W} = \operatorname{diag}(w_1, \ldots, w_k)$ *where*

Std

$$w_i = \frac{1/\hat{\sigma}_{\epsilon i}}{\sum_{j=1}^{k} 1/\hat{\sigma}_{\epsilon j}}$$

3. *Compute PCA-GLS using* $\mathbf{WX}$

Other Generalized PCA Estimators

UNIVERSITY OF OXFORD

- Absolute covariance weighting
  1. Compute complete residual covariance $\hat{\boldsymbol{\Sigma}}_\epsilon$ from residuals
  2. Replace $\hat{\sigma}^2_{\epsilon i}$ in step 2 with $\hat{\sigma}^2_{\epsilon i} = \sum_{j=1}^{k} \left|\hat{\boldsymbol{\Sigma}}_\epsilon\left(i,j\right)\right|$
- Down-weights series which have both large idiosyncratic variance *and* strong residual covariance
- Stock & Watson (2005) use more sophisticated method
  1. Estimate AR(P) on $\hat{\epsilon}_{it}$ for all series

$$\hat{\epsilon}_{it} = \sum_{j=1}^{p_i} \phi_j \epsilon_{it-j} + \xi_{it}$$

  2. Construct quasi-differenced x_{it} using coefficients

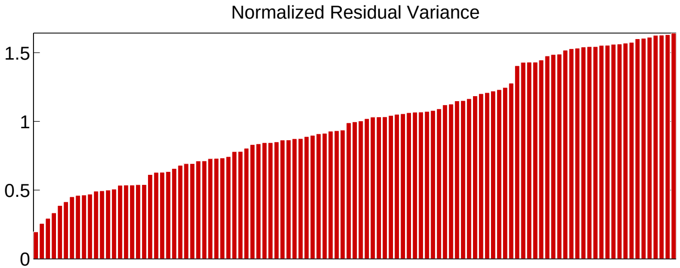
$$\tilde{x}_{it} = x_{it} - \sum_{j=1}^{p_i} \hat{\phi}_j x_{it-j}$$

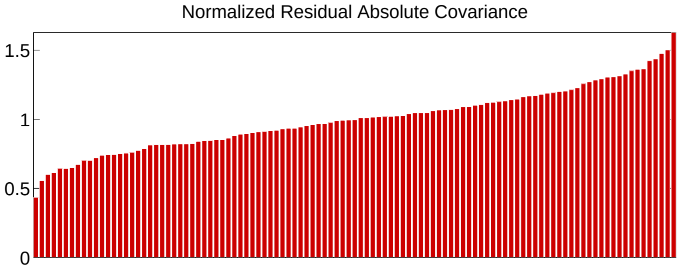
  3. Estimate $\hat{\sigma}^2_{\epsilon i}$ using $\hat{\xi}_{it}$
  4. Re-estimate factors using quasi-differenced data and weighting, iterate if needed

Generalized Principal Components Inputs

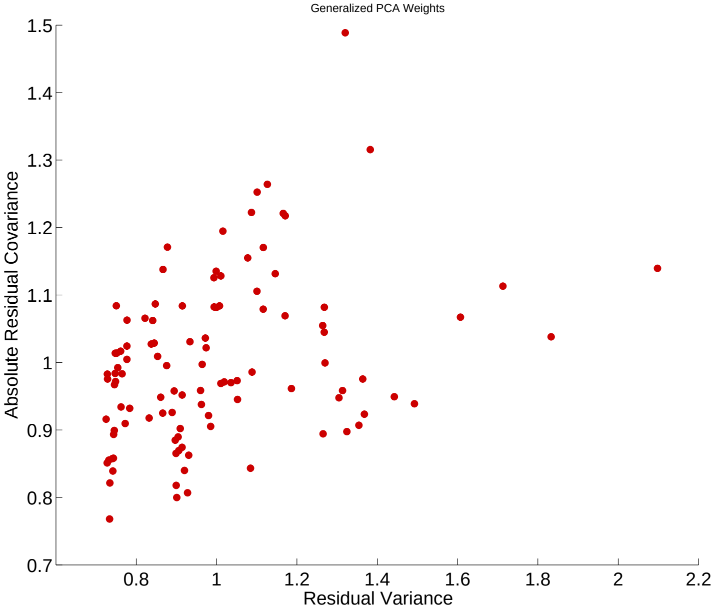
UNIVERSITY OF OXFORD

$\frac{1}{G_i}$





Generalized Principal Components Weights



Redundant and repeated factors

- Redundant factors can have adverse effects on common components
- Exactly redundant factors are identical to increasing the variance of a studentized data series
  - Including x_{it} m-times is the same as using mx_{it}
- Some evidence that excluding highly correlated factors is useful (Boivin & Ng 2006)

Algorithm (Removal of Redundant Factors)

1. *For each series i find series with maximally correlated error, call index j_i*
2. *Drop series in $\{j_i\}$ that are maximally correlated with more than 1 series*
3. *For series which are each other's j_i, drop series with lower R^2*

- Can increase step 1 to two or even three series

Thresholding to Select Forecasting Relevant Factors

UNIVERSITY OF OXFORD

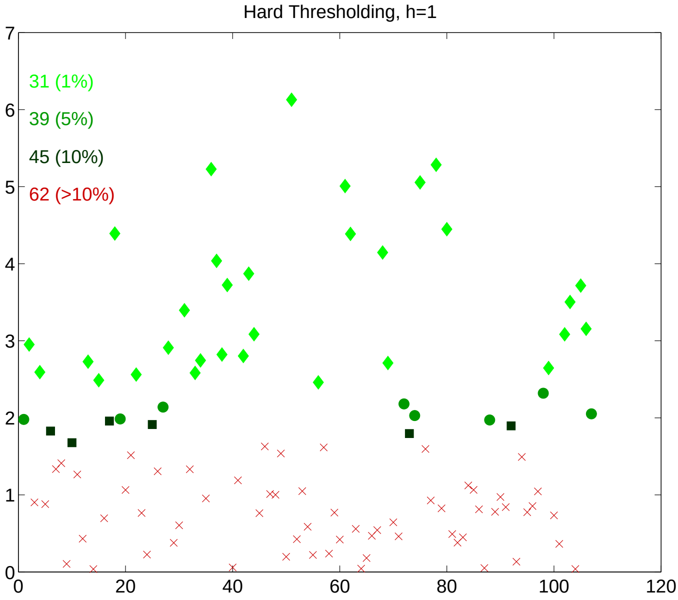
- Bai & Ng (2008) consider problem of selecting *forecasting relevant* factors
- Well known issue for PCA is that factors are selected only using $\mathbf{x}_t$
- Can this be improved using information about y_t?

Algorithm (Hard Thresholding for Variable Selection)

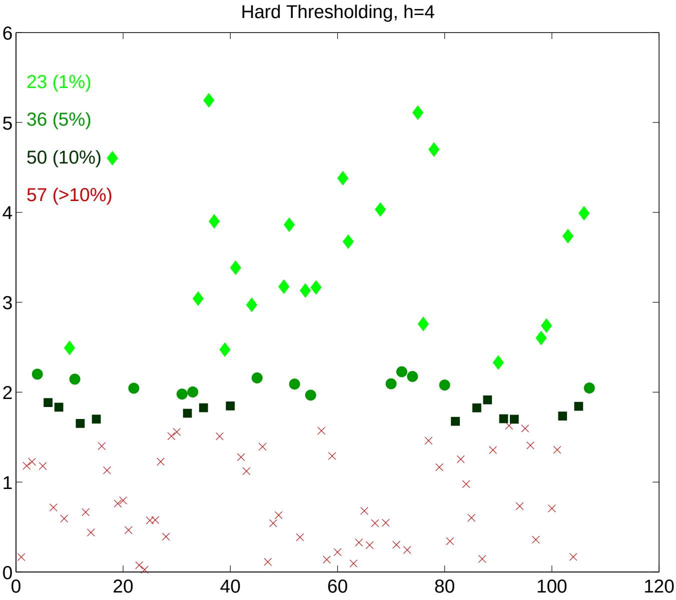
1. *Regress* $y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \gamma x_{t-1} + \epsilon_t$
2. *Compute White heteroskedasticity robust standard errors and t-stat*
3. *Retain any* x_t *where* $|t| > C_\alpha$ *for some choice of* α*. Common choices are 10%, 5% or 1%.*

- Bai & Ng also discuss methods for soft thresholding, but these require technology beyond this course (LASSO and Elastic Net)

Hard Thresholding for GDP, $h = 1$

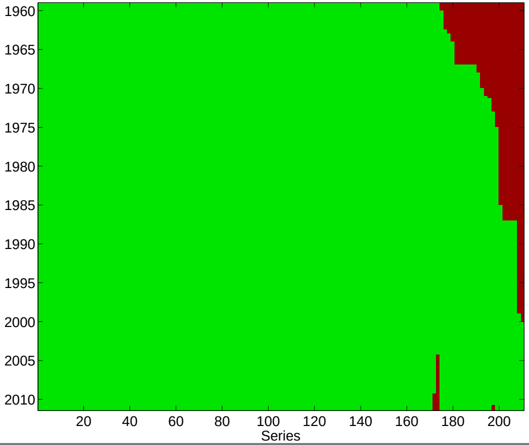


Hard Thresholding for GDP, $h = 4$



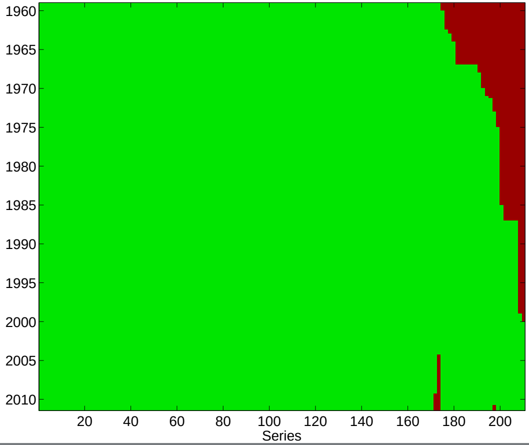
Prinicpal Component Analysis with Missing Data

- Two obvious solutions to missing data in PCA
  - Drop all series that have missing observations
  - Impute values for the missing values
- Missing data structure in SW 2012



Prinicpal Component Analysis with Missing Data

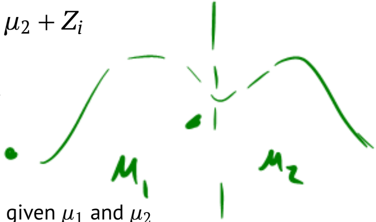
- Two obvious solutions to missing data in PCA
  - Drop all series that have missing observations
  - Impute values for the missing values
- Missing data structure in SW 2012



Expectations-Maximization (EM) Algorithm

- Some problem with unobserved states can be solved using the EM algorithm
- Consider problem of estimating means from an i.i.d. mixture

$$X_i = Y_i \mu_1 + (1 - Y_i)\,\mu_2 + Z_i$$



- Y_i is i.i.d. Bernoulli(p), Z_i is standard normal
- Y_i was observable, trivial problem (OLS)
- When Y_i is not observable, much harder
- EM algorithm will iterate across two steps:
  1. Construct "as-if" Y_i using expectations of Y_i given μ_1 and μ_2
  2. Compute

$$\hat{\mu}_1 = \frac{\sum \Pr(Y_i = 1)\, X_i}{\sum \Pr(Y_i = 1)} \qquad \hat{\mu}_2 = \frac{\sum \Pr(Y_i = 0)\, X_i}{n - \sum \Pr(Y_i = 1)}$$

  3. Return to 1, stopping if the means are not changing much
- Algorithm is initialized with "guesses" about μ_1 and μ_2
  - Example: Mean of data above median, mean of data below median
- Consider case where $\mu_1 = 10$, $\mu_2 = -10$

Imputing Missing Values in PCA

UNIVERSITY OF OXFORD

- Ideally would like to solve PCA problem only for observed data
- Difficult in practice, no know closed form estimator
- Expectation-Maximization (EM) algorithm can be used to simply impute missing data
  - Replace missing with r-factor expectation (E)
  - Maximize the likelihood (M), or minimize sum of squares

Algorithm (EM Algorithm for Imputing Missing Values in PCA)

1. *Define* $w_{ij} = I\left[x_{ij} \text{ observed}\right]$ *and set* $i = 0$
2. *Construct* $\mathbf{X}^{(0)} = \mathbf{W} \odot \mathbf{X} + (1 - \mathbf{W}) \odot \boldsymbol{\iota}\bar{\mathbf{X}}$ *where* $\boldsymbol{\iota}$ *is a* T *by 1 vector of 1s*
3. *Until* $\left|\left|\mathbf{X}^{(i+1)} - \mathbf{X}^{(i)}\right|\right| < c$*:*
   a. *Estimate r factors and factor loadings,* $\hat{\mathbf{F}}^{(i)}$ *and* $\hat{\boldsymbol{\beta}}^{(i)}$ *from* $\mathbf{X}^{(i)}$ *using PCA*
   b. *Construct* $\mathbf{X}^{(i+1)} = \mathbf{W} \odot \mathbf{X} + (1 - \mathbf{W}) \odot \left(\hat{\mathbf{F}}^{(i)}\hat{\boldsymbol{\beta}}^{(i)}\right)$
   c. *Set* $i = i + 1$

Hierarchical Factors

- Can use partitioning to construct hierarchical factors
- Global and Local
  1. Extract 1 or more factors from all series
  2. For each regions or country *j*, regress series from country *j* on Global Factors, and extract 1 or more factors from residuals

  - Country factors uncorrelated with Global, but not local from other regions/countries
- Nominal and Real
  1. Extract 1 or more general factors
  2. For each group real/nominal series, regress on general factors and then extract factors from residuals