Forecasting With Many predictors

The Econometrics of Predictability

This version: June 3, 2014

June 3, 2014

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\ldots)$$

UNIVERSITY OF OXFORD

Forecasting with many predictors

- Dynamic Factor Models
- The 3-Pass Regression Filter
- Regularized Reduced Rank Regression
- Time permitting
  - Bagging
  - Filters and decompositions

How Many is Many?

- Many here means 25 or more
- Often many more, 100s of series

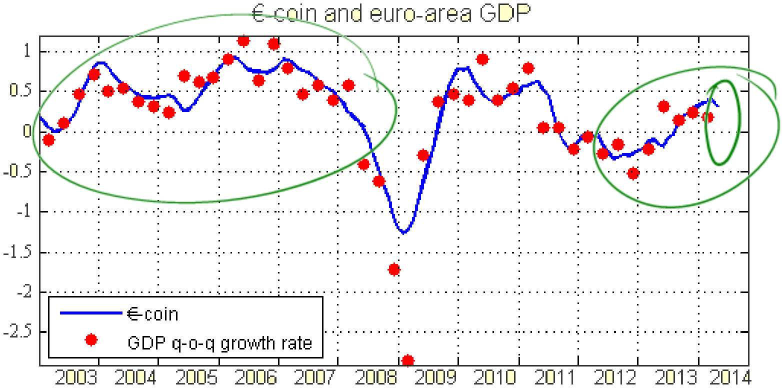
New challenges

Why factor models

- Are parsimonious while effectively including many regressors
- Can remove measurement error or other useless information from predictors
- Factor may be of interest
  - Leading indicators:
    - €-coin
    - Chicago Fed National Activity Index
    - Aruoba-Diebold-Scotti Business Conditions Index
  - Real and Nominal factors
  - Global and Local factors

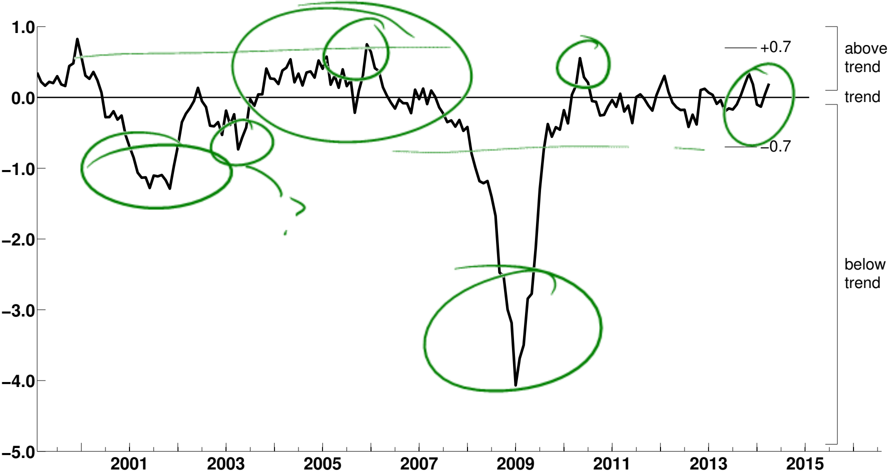
- European Coincident Indicator
- First factor in a Europe-wide model

€-coin: the Euro Area Economy in One Figure – May 2014



Chicago Fed National Activity Index

- Factor extracted from 85 series
- Based on research in forecasting inflation



ADS Business Conditions Index

UNIVERSITY OF OXFORD

- Based on factor model in Aruoba, Diebold & Scotti
- Extracts common factor in:
  - weekly initial jobless claims
  - monthly payroll employment
  - industrial production
  - personal income less transfer payments, manufacturing and trade sales
  - quarterly real GDP

The Model

- Scalar *latent* factor

$$x_t = \sum_{i=1}^{q} \rho_i x_{t-i} + \eta_i$$

- Indicators

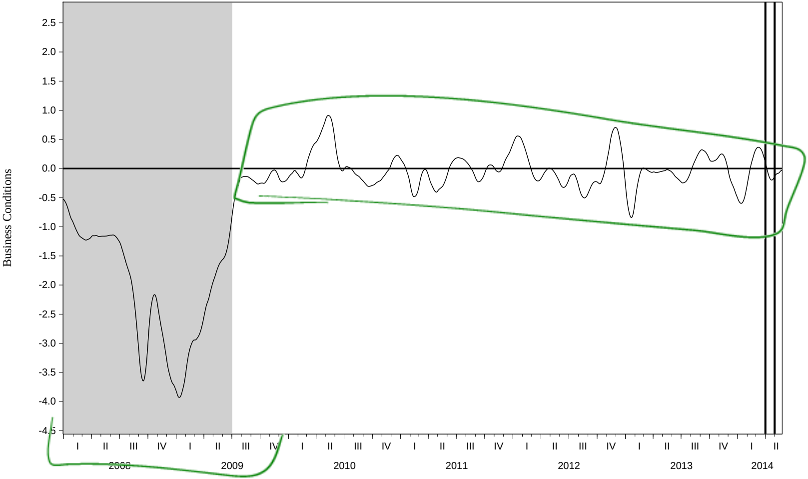
$$y_{it} = c_i + \beta_i x_t + \sum_{j=1}^{p_i} \gamma y_{it-\Delta_i} + \epsilon_i$$

  - Δ_i allows series to have different observational frequencies

$$Y_t = \gamma_0 + \gamma_1 x_t + \varepsilon_t \qquad \varepsilon_t \sim WN$$

ADS Business Conditions Index

Aruoba-Diebold-Scotti Business Conditions Index (12/31/2007- 05/24/2014)



Notation

UNIVERSITY OF OXFORD

- T number of time series observations
- k number of series available to forecast
- $\mathbf{y}_t$ series to be forecast, m by 1
  - m will often be 1
- $\mathbf{x}_t$ series used to forecast, k by 1
  - Usually assume $\mathrm{E}[\mathbf{x}_t] = \mathbf{0}$ and $\mathrm{Cov}[\mathbf{x}_t] = \mathbf{I}_k$
  - Demeaned and standardized
  - Suppose $\mathbf{x}_t = \mathbf{\Sigma}_{\mathbf{x}}^{-1/2} \left(\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_X\right)$
- $\mathbf{f}_t$ factors, r by 1
- $\mathbf{x}_t$ *may be* $\mathbf{y}_t$, but not necessarily
  - $\mathbf{y}_t$ could be subset of $\mathbf{x}_t$ (common)
  - $\mathbf{y}_t$ could be excluded from factor estimation (uncommon)

$T, k \to \infty$

$r \ll k$ $\quad 1 \le r \le 6$

Why factor models?

- Factor models help avoid issues with large, kitchen-sink models
- Consider issue of parameter estimation error when forecasting
- Suppose correct model is linear

$$y_{t+1} = \boldsymbol{\beta} \mathbf{x}_t + \epsilon_t$$

- Forecast using OLS estimates is then

$$\begin{aligned}
\hat{y}_{t+1|t} &= \hat{\boldsymbol{\beta}} \mathbf{x}_t \\
&= \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} + \boldsymbol{\beta}\right) \mathbf{x}_t \\
&= \underbrace{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \mathbf{x}_t}_{\text{estimation error}} + \underbrace{\boldsymbol{\beta} \mathbf{x}_t}_{\text{correct forecast}}
\end{aligned}$$

OLS when there are many regressors

UNIVERSITY OF OXFORD

- Suppose $\epsilon_t, \mathbf{x}_t$ are independent and jointly normally distributed

$$\text{Cov}\begin{bmatrix} \epsilon_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \sigma_\epsilon^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \end{bmatrix}$$

mean 0

- Standard assumptions have k fixed, so as $T \to \infty$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{p} 0$

$$\hat{y}_{t+1|t} \sim N\left(\boldsymbol{\beta}\mathbf{x}_t, 0\right)$$

- Degenerate normal - no error since $\boldsymbol{\beta}$ is effectively *known*
- What about the case when k is large
- Use *diagonal* asymptotics, $k/T \to c$, $0 < \underline{\kappa} < c < \bar{\kappa} < \infty$
- In this case

$$\hat{y}_{t+1|t} \sim N\left(\boldsymbol{\beta}\mathbf{x}_t, {}^{k}/_{T} \times \sigma_\epsilon^2\right)$$

  - Is still random, even when $T \to \infty$
- True even if all $\boldsymbol{\beta} = \mathbf{0}$!

(Really) Big models don't make sense

UNIVERSITY OF OXFORD

- When the number of parameters is large, then almost all coefficients must be 0

$$y_t = \sum_{i=1}^{k} \beta_i x_{t,i} + \epsilon_i$$

- Variance of the LHS is the same as the RHS

$$\mathrm{V}[y_t] = \sum_{i=1}^{k} \beta_i^2 + \sigma_\epsilon^2$$

- If $k \to \infty$, $\inf_i |\beta_i| > \underline{\kappa} > 0$, then $\mathrm{V}[y_t] \to \infty$
- Even when T is very large, it will not usually make sense to have k extremely large
- Factor models will effectively have small β_i coefficient, only using two steps
    1. Construct average-like estimators of factors from $\mathbf{x}_t$ – coefficients are $O(1/k)$
    2. Weight these using a small number of relatively large coefficients

Static Factor Models

Static Factor Models

- Consider the cross-section of asset returns
- Model uses factors as RHS variables

$$x_{jt} = \sum_{i=1}^{q} \lambda_{ji} f_{it} + \epsilon_{jt}$$

- λ_i are the factor loadings
- ϵ_{jt} is the idiosyncratic error for series j
- In vector notation,

$$\mathbf{x}_t = \mathbf{\Lambda} \mathbf{f}_t + \boldsymbol{\epsilon}_t$$

  - $\mathbf{\Lambda}$ is k by r
  - $\mathbf{f}_t$ is r by 1

UNIVERSITY OF OXFORD

- In matrix notation,

$$\mathbf{X} = \mathbf{\Lambda F} + \boldsymbol{\epsilon}$$

  - $\mathbf{X}$ is k by T
  - $\mathbf{F}$ is T by r
  - $\boldsymbol{\epsilon}$ is k by 1
- When model is a strict (as opposed to approximate), $\mathrm{E}\left[\boldsymbol{\epsilon}_t\right] = \mathbf{0}$ and $\mathrm{E}\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'\right] = \mathbf{\Sigma}_{\epsilon} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)$
- Covariance of $\mathbf{x}_t$ is then

$$\mathbf{\Lambda \Omega \Lambda}' + \mathbf{\Sigma}_{\epsilon}$$

  - $\mathbf{\Omega} = \mathrm{Cov}\left[\mathbf{f_t}\right]$, r by r
  - Covariance will play a crucial role in estimation of factors

Estimation using Principal Components

- Principal components can be used to estimate factors
- Formally, problem is

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

- $\boldsymbol{\beta}$ is k by r
  - $\boldsymbol{\beta}$ is related to but different from $\boldsymbol{\Lambda}$
  - $\boldsymbol{\Lambda}$ is the DGP parameter
  - $\boldsymbol{\beta}$ is a normalized and *rotated* version of $\boldsymbol{\Lambda}$

Definition (Rotation)

A square matrix $\mathbf{B}$ is said to be a rotation of a square matrix $\mathbf{A}$ if $\mathbf{B} = \mathbf{QA}$ and $\mathbf{QQ}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}$.

- $\mathbf{f}_t$ is r by 1
- $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$ is a *normalization*, and is required
  - $\boldsymbol{\beta}\mathbf{f}_t = ((\boldsymbol{\beta}/2)(2\mathbf{f}_t))$
  - Generally, for full rank $\mathbf{Q}$, $(\boldsymbol{\beta}\mathbf{Q})\left(\mathbf{Q}^{-1}\mathbf{f}_t\right) = \tilde{\boldsymbol{\beta}}\tilde{\mathbf{f}}_t$

The Objective Function

UNIVERSITY OF OXFORD

- If $\boldsymbol{\beta}$ was observable, solution would be OLS

$$\hat{\mathbf{f}}_t = \left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\mathbf{y}_t$$

  This can be substituted into the objective function

$$\sum_{t=1}^{T}\left(\mathbf{x}_t - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\mathbf{y}_t\right)'\left(\mathbf{x}_t - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\mathbf{x}_t\right) = \sum_{t=1}^{T}\mathbf{x}_t'\left(\mathbf{I} - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right)\mathbf{x}_t$$

- This works since $\mathbf{I} - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'$ is *idempotent*
  - $\mathbf{AA} = \mathbf{A}$
- Some additional manipulation using the trace operator on a scalar leads to two equivalent expressions

$$\min_{\boldsymbol{\beta}}\sum_{t=1}^{T}\mathbf{x}_t'\left(\mathbf{I} - \boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1}\boldsymbol{\beta}'\right)\mathbf{x}_t = \max_{\boldsymbol{\beta}}\operatorname{tr}\left(\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1/2}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)^{-1/2}\right)$$

$$= \max_{\boldsymbol{\beta}}\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta}$$

  - All subject to $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$
- Solution to last problem sets $\boldsymbol{\beta}$ to the *eigenvectors* of $\boldsymbol{\Sigma}_{\mathbf{x}}$

UNIVERSITY OF OXFORD

Definition (Eigenvalue)

The eigenvalues of a real, symmetric matrix k by k matrix $\mathbf{A}$ are the k solutions to

$$|\lambda \mathbf{I}_k - \mathbf{A}| = 0$$

where $|\cdot|$ is the determinant.

- Properties of eigenvalues
  - $\det \mathbf{A} = \prod_{i=1}^{r} \lambda_i$
  - $\text{tr}\mathbf{A} = \sum_{i=1}^{r} \lambda_i$
  - For positive (semi) definite $\mathbf{A}$, $\lambda_i > 0, i = 1, \ldots, r$ ($\lambda_i \geq 0$)
  - Full-rank $\mathbf{A}$ implies $\lambda_i \neq 0, i = 1, \ldots, r$
  - Rank $q < r$ matrix $\mathbf{A}$ implies $\lambda_i \neq 0, i = 1, \ldots, q$ and $\lambda_j = 0, j = q+1, \ldots, r$

Properties of Eigenvalues and Eigenvectors

Definition (Eigenvector)

An a k by 1 vector $\mathbf{u}$ is an eigenvector corresponding to an eigenvalue λ of a real, symmetric matrix k by k matrix $\mathbf{A}$ if

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{u}$$

- Properties of eigenvectors
  - If $\mathbf{A}$ is positive definite, then

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$$

where $\boldsymbol{\Lambda}$ is diagonal and $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = \mathbf{I}$

Definition (Orthonormal Matrix)

A k-dimensional orthonormal matrix $\mathbf{U}$ satisfies $\mathbf{U}'\mathbf{U} = \mathbf{I}_k$, and so $\mathbf{U}' = \mathbf{U}^{-1}$.

- Implication is

$$\mathbf{V}'\mathbf{A}\mathbf{V} = \mathbf{V}'\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'\mathbf{V} = \boldsymbol{\Lambda}$$

Computing Factors using PCA

UNIVERSITY OF OXFORD

- $\mathbf{X}$ is T by k
- $\mathbf{X}'\mathbf{X}$ is real and symmetric with eigenvalues $\mathbf{\Lambda} = \text{diag}(\lambda_i)_{i=1,\ldots,k}$
- Factors are estimated

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$$
$$\mathbf{V}'\mathbf{X}'\mathbf{X}\mathbf{V} = \mathbf{V}'\mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{V}$$
$$(\mathbf{X}\mathbf{V})'(\mathbf{X}\mathbf{V}) = \mathbf{\Lambda} \text{ since } \mathbf{V}' = \mathbf{V}^{-1}$$
$$\mathbf{F}'\mathbf{F} = \mathbf{\Lambda}.$$

- $\mathbf{F} = \mathbf{X}\mathbf{V}$ is the T by k matrix of factors
- $\boldsymbol{\beta} = \mathbf{V}'$ is the k by k matrix of factor loadings.
- All factors exactly reconstruct $\mathbf{Y}$

$$\mathbf{F}\boldsymbol{\beta} = \mathbf{F}\mathbf{V}' = \mathbf{Y}\mathbf{V}\mathbf{V}' = \mathbf{Y}$$

  - Assumes k is large
- Note that both factors *and* loadings are orthogonal since

$$\mathbf{F}'\mathbf{F} = \mathbf{\Lambda} \text{ and } \boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}$$

- Only loadings are normalized

Large k and factor analysis

- Consider simple example where

$$x_{it} = f_t + \epsilon_{it}$$

- f_t and ϵ_{it} are all independent, standard normal
- First eigenvector of $\mathbf{\Sigma}_\mathbf{x} = \mathbf{1} + I_k$ is

$$\left(k^{-1/2}, k^{-1/2}, \ldots, k^{-1/2}\right)$$

  - Form is due to normalization

$$\sum_{i=1}^{k} v_{ij}^2 = 1, \quad \sum_{i=1}^{k} v_{ij} v_{in} = 0$$

  - $\sum_{i=1}^{k} \left(k^{-1/2}\right)^2 = \sum_{i=1}^{k} k^{-1} = kk^{-1} = 1$

Estimated Factors

- Estimated factor is then

$$\hat{f}_t = \sum_{i=1}^{k} k^{-1/2} x_{it} = k^{1/2} \left({}^{1}\!/_{k} \sum x_{it} \right) = k^{1/2} \bar{x}$$

- What about $\bar{x}$

$$\begin{aligned} \bar{x} &= k^{-1} \left(\sum_{i=1}^{k} f_t + \epsilon_{it} \right) \\ &= f_t + \bar{\epsilon}_t \\ &\approx f_t \end{aligned}$$

- Normalization means factor is $O_p\left(k^{1/2}\right)$
  - Can always re-normalize factor to be $O_p(1)$ using $\hat{f}_t / k^{1/2}$
- Key assumption is that $\bar{\epsilon}_t$ follows some form of LLN *in k*
- In strict factor model, no correlation so simple

Approximate Factor Models

- Strict factor models require strong assumptions

$$\text{Cov}\left(\epsilon_{it}, \epsilon_{js}\right) = 0 \quad i \neq j,\ s \neq t$$

- These are easily rejectable in practice
- Approximate Factor Models relax these assumptions and allow:
  - (*Weak*) Serial correlation in $\boldsymbol{\epsilon}_t$

$$\sum_{s=0}^{\infty} |\gamma_s| < \infty$$

  - (*Weak*) Cross-sectional correlation between ϵ_{it} and ϵ_{jt}

$$\lim_{k\to\infty} \sum_{i\neq j}^{k} \mathrm{E}\left|\epsilon_{it}\epsilon_{jt}\right| < \infty$$

  - Heteroskedasticity in ϵ
- Requires pervasive factors

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\Lambda}\mathbf{f}_t + \epsilon_t \\ \lim_{k\to\infty} \text{rank}\left(k^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right) &= r \end{aligned}$$

Practical Considerations when Estimating Factors

- Key input for factor estimation is $\Sigma_{\mathbf{x}}$
- In most theoretical discussions of PCA, this is the covariance

$f_t = X_t V'$

$$\Sigma_{\mathbf{x}} = T^{-1} \sum_{t=1}^{T} (\mathbf{x}_t - \hat{\boldsymbol{\mu}})(\mathbf{x}_t - \hat{\boldsymbol{\mu}})$$

- Two other simple versions are used
  - Outer-product

$$T^{-1}\mathbf{X}'\mathbf{X} = T^{-1} \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'$$

    - Similar to fitting OLS *without* a constant
  - Correlation matrix

$$\mathbf{R}_x = T^{-1} \sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t'$$

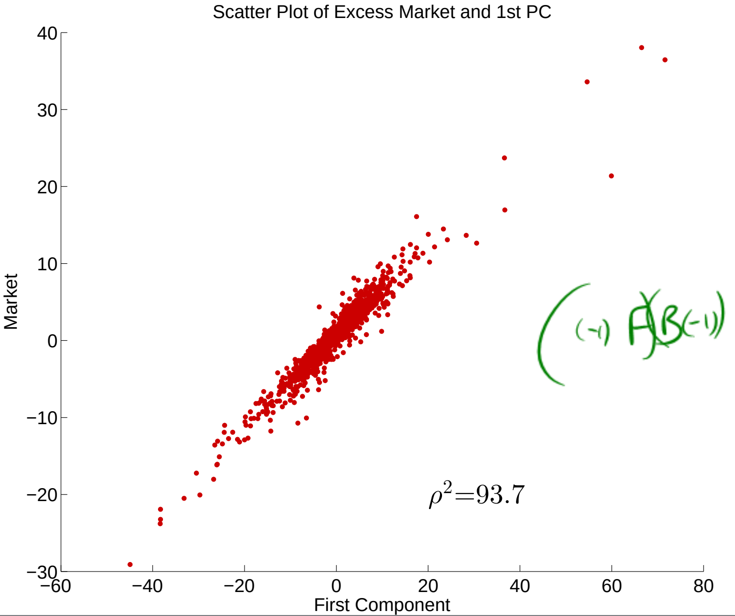
    - $\mathbf{z}_t = (\mathbf{x}_t - \hat{\boldsymbol{\mu}}) \oslash \hat{\sigma}$ are the original data series, only studentized
    - Makes sense for most economic data since scale is often not well defined (e.g. an index)

Fama-French Data

UNIVERSITY OF OXFORD

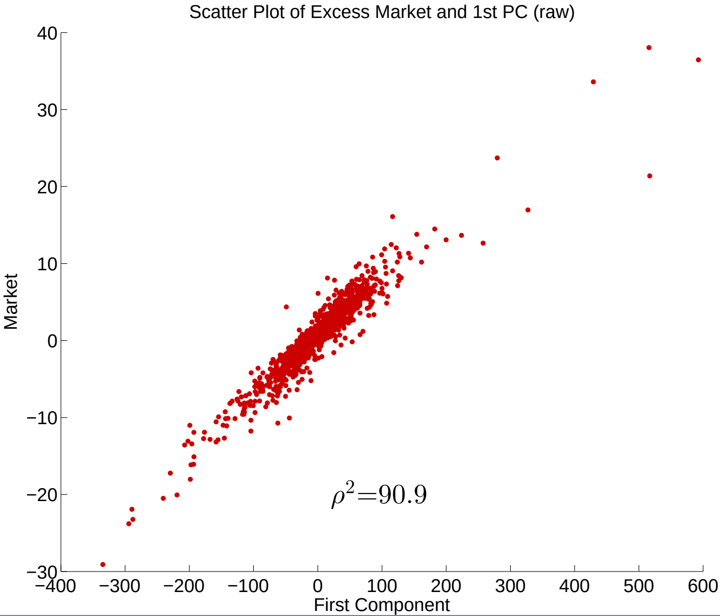
- Initial exploration based on Fama-French data
  - 100 portfolios
    - Sorted on size and book-to-market
  - 49 portfolios
    - Sorted on industry
- Equities are known to follow a strong factor model
  - Series missing more than 24 missing observations were dropped
    - 73 for 10 by 10 sort remaining
    - 41 of 49 industry portfolios
  - First 24 data points dropped for all series
  - July 1928 – December 2013
- $T = 1,026$
- $k = 114$
- Two versions, studentized and *raw*

First Factor from FF Data



First Factor from FF Data (Raw)

UNIVERSITY OF OXFORD



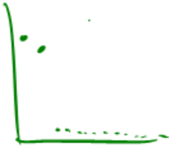
Selecting the Number of Factors (r)

Choosing the number of factors

UNIVERSITY OF OXFORD

- So far have assumed r is known
- In practice r has to be estimated
- Two methods
  - Graphical using Scree plots
    - Plot of ordered eigenvalues, usually standardized by sum of all
    - Interpret this as the R^2 of including r factors
    - Recall $\sum_{i=1}^{l} \lambda_i = k$ for correlation matrix (Why?)
  - Information criteria-based
    - Similar to AIC/BIC, only need to account for both k and T

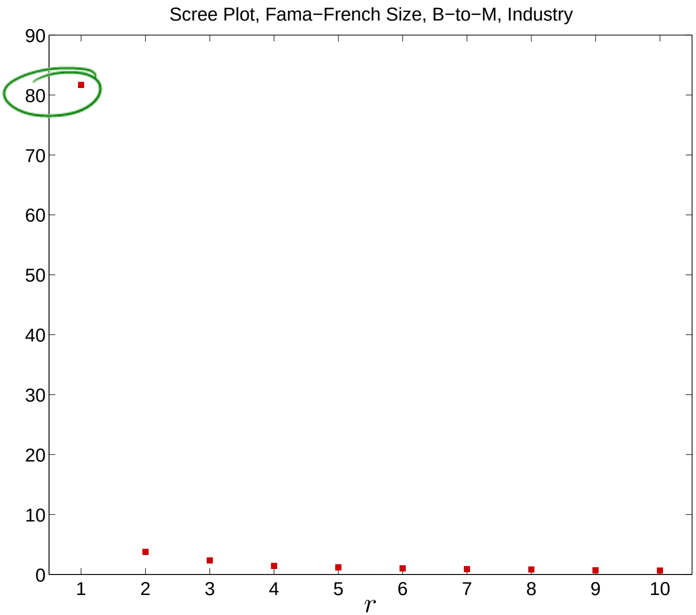
$\lambda_i / \sum \lambda_i \approx R^2$



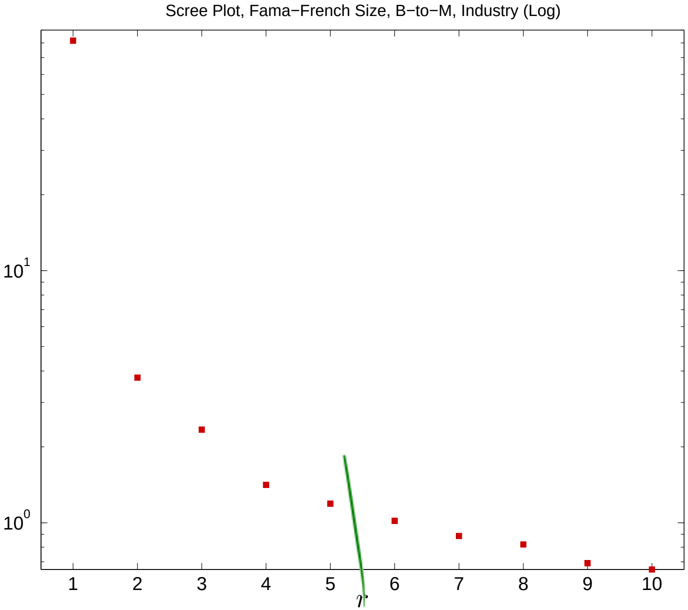
Stylized Fact(ors)

If in doubt, all known economic panels have between 1 and 6 factors

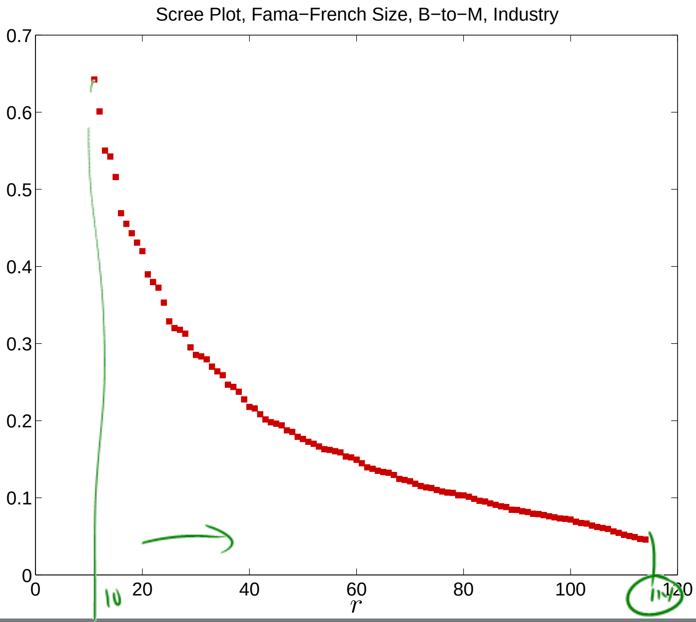
Scree Plot: Fama-French



Scree Plot: Fama-French



Scree Plot: Fama-French (Non-Factors)



Information Criteria

UNIVERSITY OF OXFORD

- Bai & Ng (2002) studied the problem of selecting the correct number of factors in an approximate factor model
- Proposed a number of information criteria with the form

$$\ln \widehat{V(r)} + r \times g(k, T)$$

$$\widehat{V(r)} = \sum_{t=1}^{T} \left(\mathbf{x}_t - \hat{\boldsymbol{\beta}}(r)\,\mathbf{f}_t(r)\right)' \left(\mathbf{x}_t - \hat{\boldsymbol{\beta}}(r)\,\mathbf{f}_t(r)\right)$$

  - $\widehat{V(r)}$ is the value of the objective function with r factors
- Three versions

$$IC_{p_1} = \ln \widehat{V(r)} + r\left(\frac{k+T}{kT}\right)\ln\left(\frac{kT}{k+T}\right)$$

$$IC_{p_2} = \ln \widehat{V(r)} + r\left(\frac{k+T}{kT}\right)\ln\left(\min(k, T)\right)$$

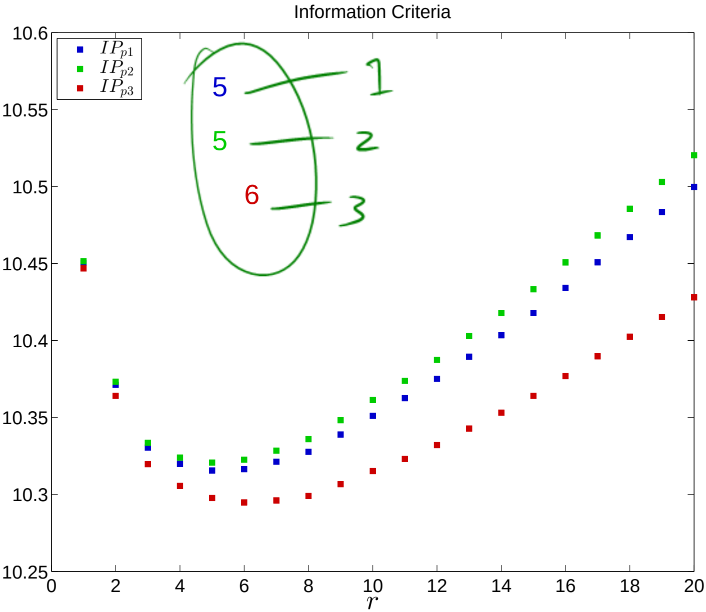
$$IC_{p_3} = \ln \widehat{V(r)} + r\left(\frac{\ln\left(\min(k, T)\right)}{\min(k, T)}\right)$$

- Suppose $k \approx T$, IC_{p_2} is BIC-like

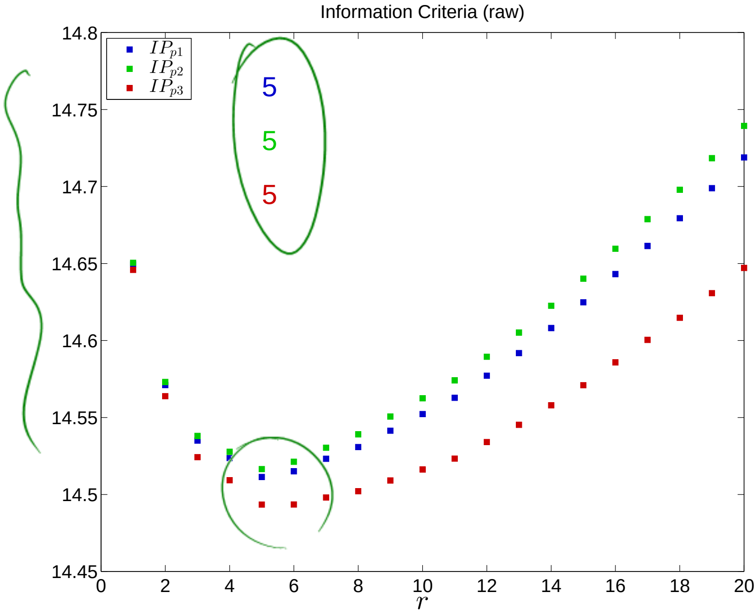
$$IC_{p2} = \ln \widehat{V(r)} + 2r\left(\frac{\ln T}{T}\right)$$

Sum on k

Information Criteria: Fama-French



Information Criteria: Fama-French (Raw)



Assessing Fit

- Fit can be assessed both globally and for individual series
- Least squares objective leads to natural R^2 measurement of fit
- Global fit

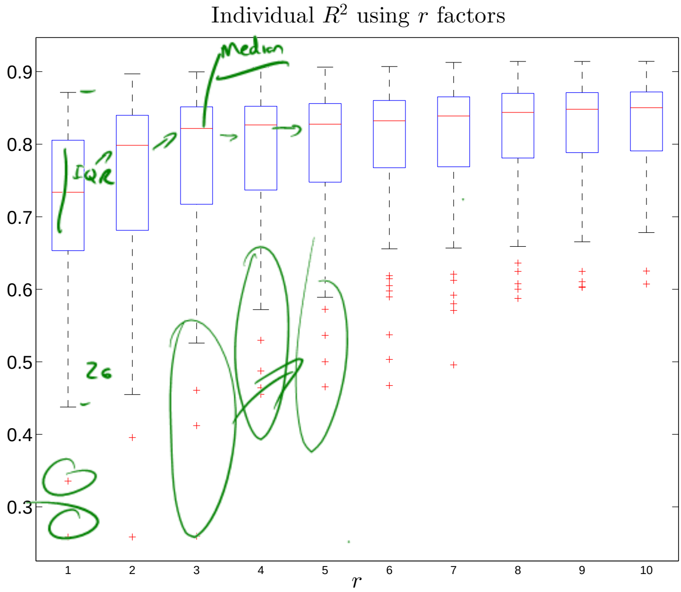
$$R^2_{\text{global}}(k) = 1 - \frac{\text{tr}\left(\mathbf{X} - \hat{\boldsymbol{\beta}}(k)\,\mathbf{F}(k)\right)'\left(\mathbf{X} - \hat{\boldsymbol{\beta}}(k)\,\mathbf{F}(k)\right)}{\text{tr}\left(\mathbf{X}'\mathbf{X}\right)}$$

$$= 1 - \frac{\sum_{i=1}^{k}\sum_{t=1}^{T}\left(x_{it} - \sum_{j=1}^{k}\hat{\beta}_{ij}f_{jt}\right)^2}{Tk}$$

  - Assumes $\mathbf{X}$ is standardized
- Individual fit

$$R_i^2(k) = 1 - \frac{\sum_{t=1}^{T}\left(x_{it} - \sum_{j=1}^{k}\hat{\beta}_{ij}f_{jt}\right)^2}{\sum_{t=1}^{T}x_{it}^2}$$

  - Useful for assessing series not well described by factor model


Individual R^2 using r factors
0.9
0.8
0.7
0.6
0.5
0.4
0.3
1
2
3
4
5
6
7
8
9
10
r
Median
IQR
2σ

Dynamic Factor Models

Dynamic Factor Models

UNIVERSITY OF OXFORD

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

$$\mathbf{x}_t = \sum_{i=0}^{s} \mathbf{\Phi}_i \mathbf{f}_t + \boldsymbol{\epsilon}_t$$

$$\mathbf{f}_t = \sum_{j=1}^{q} \mathbf{\Psi} f_{t-j} + \boldsymbol{\eta}_t$$

- 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
- $\boldsymbol{\epsilon}_t$ 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

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

- Optimal forecast can be derived

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

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

Invertibility and MA processes

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

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

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

Dynamic as Static Factor Models

UNIVERSITY OF OXFORD

- 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 estimates
  - Latter are not useful for forecasting since 2-sided
- (Big) However, DFM can be converted to Static model by relabeling
- In DFM, factors are

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

  - 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

- 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

- Basic simulation

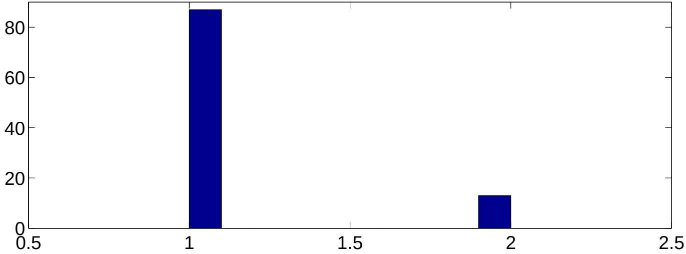
$$
\begin{aligned}
x_{it} &= \phi_{i1} f_t + \phi_{i2} f_{t-1} + \epsilon_{it} \\
f_t &= \psi f_{t-1} + \eta_t
\end{aligned}
$$

- $\phi_{i1} \sim N(1,1)$,$\phi_{i2} \sim N(.2,1)$
  - Smaller signal makes it harder to find second factor
- $\psi = 0.5$
  - Higher persistence makes it harder since Corr $\left[f_t, f_{t-1}\right]$ is larger
- Everything else standard normal
- $k = 200$, $T = 200$
  - Also $k = 400$ and $T = 400$ (separately)
- All estimation using PCA on correlation

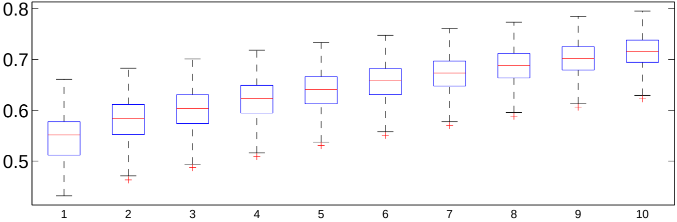
Number of Factors for Forecasting

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

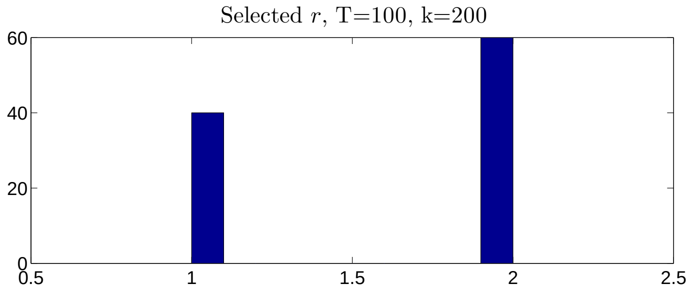
Selected r, T=100, k=100

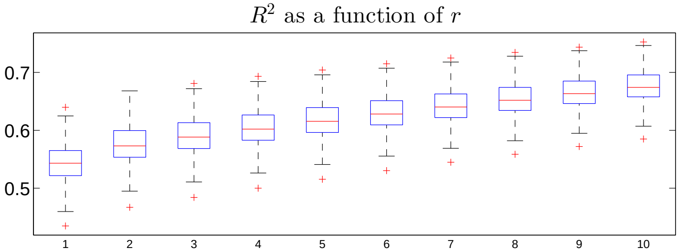


R^2 as a function of r

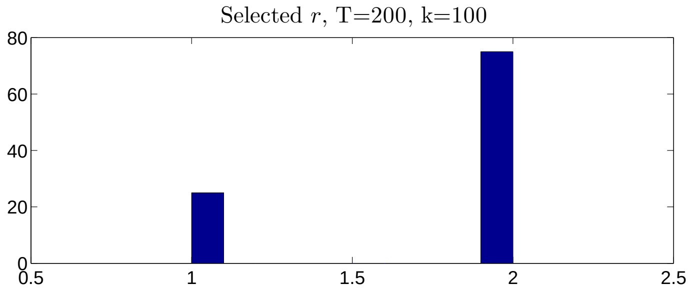


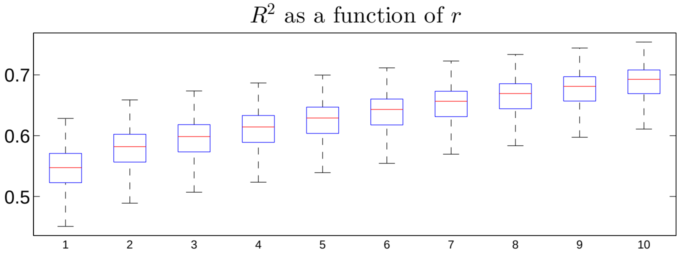
Dynamic as Static Factor Models



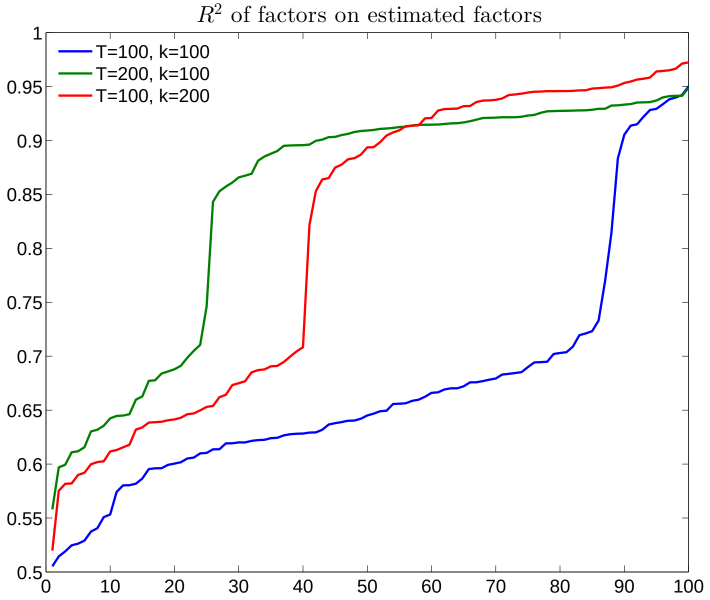


Dynamic as Static Factor Models





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

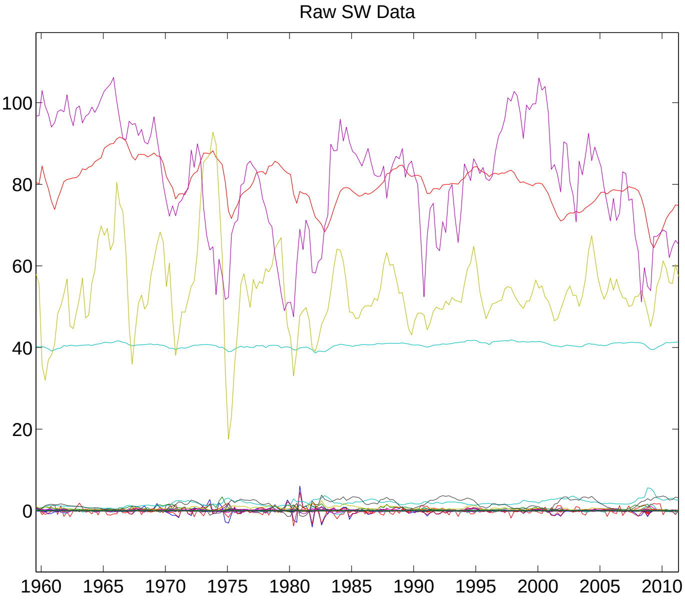
The series

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

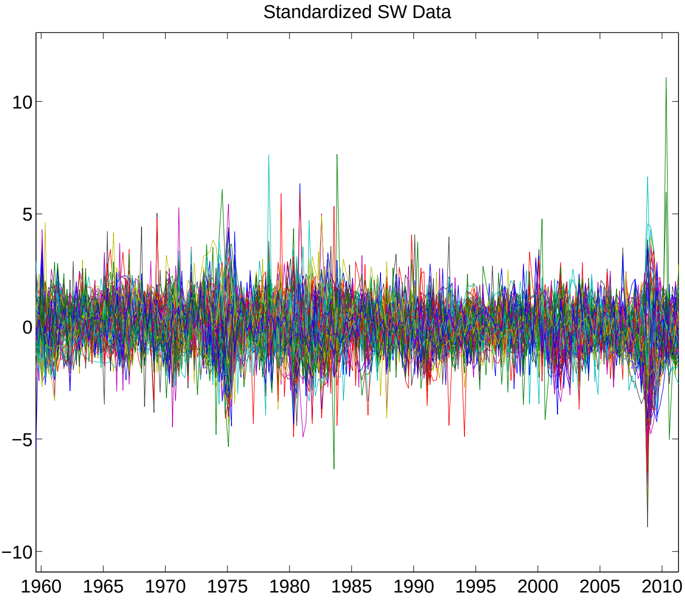
Data Transformation

- All series were transformed to be stationary using one of:
  - No transform
  - Difference
  - Double-difference
  - Log
  - Log-difference
  - Double-log-difference
- After transformation, monthly series were aggregated to quarterly using
  - Average
  - End-of-quarter
- Finally studentized

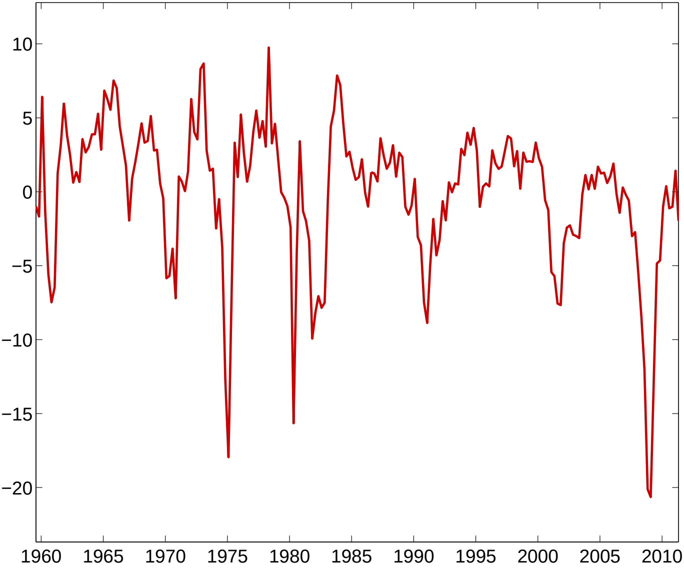
Raw Data after Transform



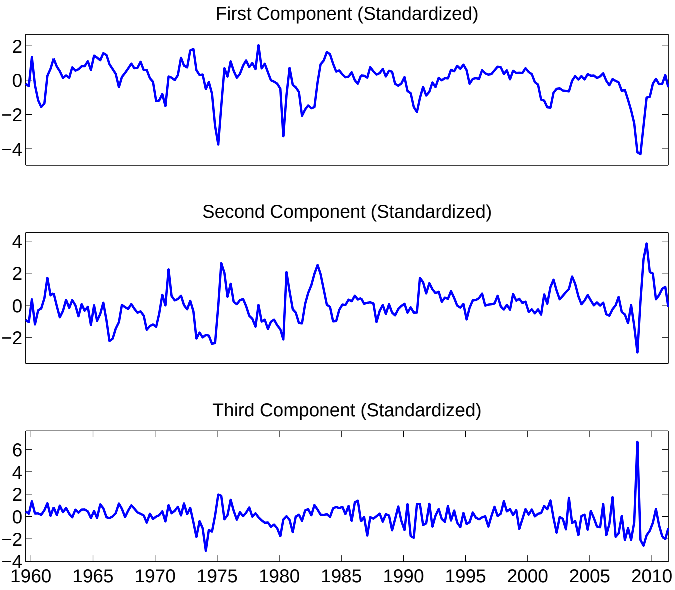
Studentized Data



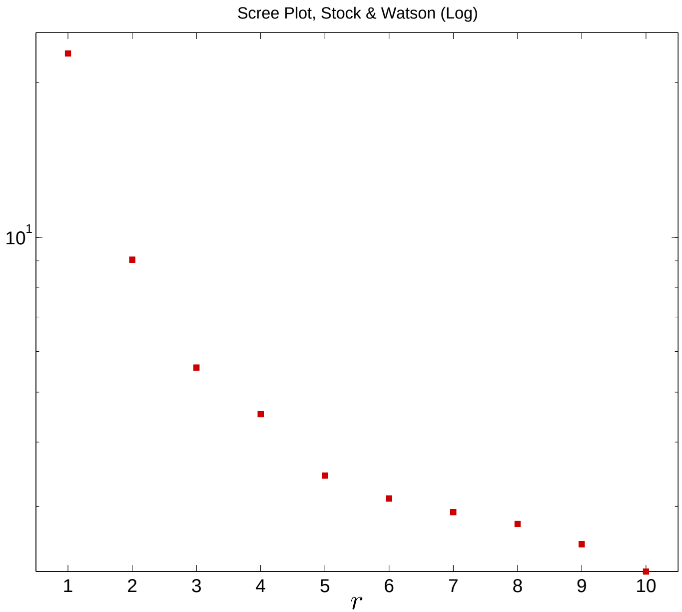
First Component



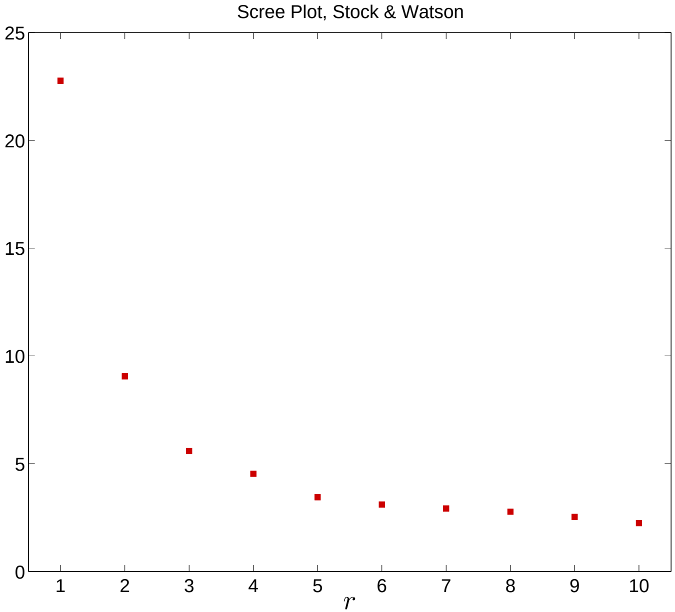
First Three Components



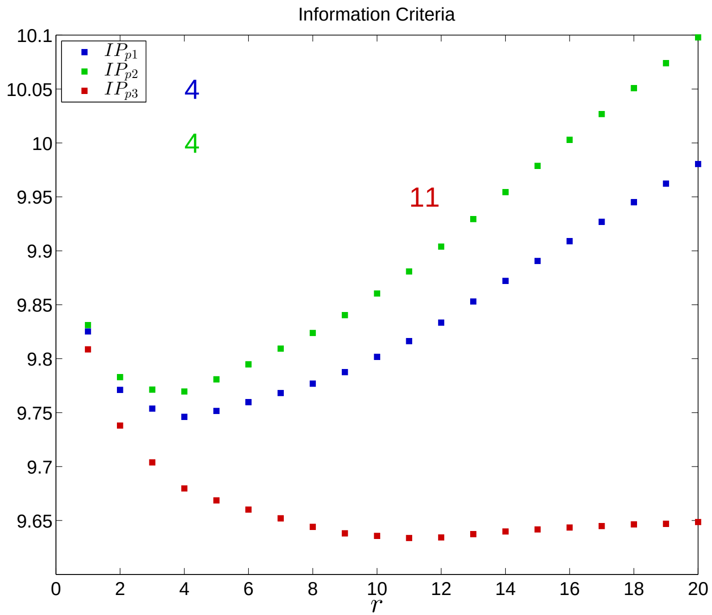
Scree Plot (Log)



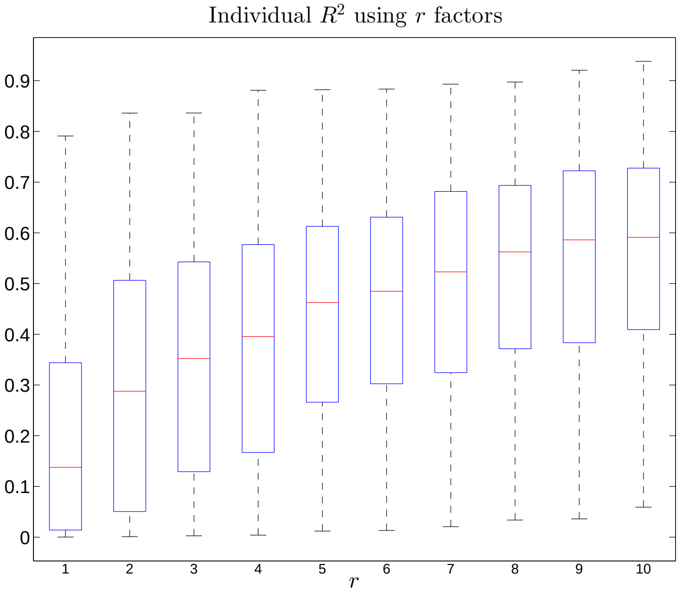
Scree Plot



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

1. Unrestricted

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

   - 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 use model selection as usual, e.g. BIC

2. Restricted - when $\mathbf{y}_t$ is in $\mathbf{x}_t$, $\mathbf{y}_t = \boldsymbol{\beta}\hat{\mathbf{f}}_t + \epsilon_t$
   - Use VAR to forecast $\hat{\mathbf{f}}_{t+1}$ using lags of $\hat{\mathbf{f}}_t$
   - Use univariate AR for $\hat{\epsilon}_t$

Re-integrating forecasts

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

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

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

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

  - Note in many cases interest in in $\Delta \mathbf{y}_t$ when forecasting $\Delta^2 \mathbf{y}_t$, e.g. CPI, inflation and change in inflation, no same as original problem

Multistep Forecasting

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

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

  - (h) used to denote explicit parameter dependence on horizon
- Direct has been documented to be better than iterative, but problem dependent

"Forecasting"

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

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

					Both		
	GDP Only	R^2	Components Only	R^2	GDP	Components	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*

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

- Clever choices of $\boldsymbol{\Sigma}_\epsilon$ lead to difference estimators
  - Using 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
    1. Estimate $\hat{\epsilon}_{it} = x_{it} - \hat{\boldsymbol{\beta}}_i \hat{\mathbf{f}}_t$ using r factors
    2. Estimate $\hat{\sigma}_{\epsilon j}^2 = T^{-1} \sum \hat{\epsilon}_{it}^2$
    3. Use $\hat{\boldsymbol{\Sigma}}_\epsilon = \text{diag}\left(\hat{\sigma}_{\epsilon 1}^2, \ldots, \hat{\sigma}_{\epsilon k}^2\right)$, which means dividing original x_{jt} by $\hat{\sigma}_{\epsilon j}$

Other Generalized PCA Estimators

- Alternatives to using basic GLS

1. Estimate $\hat{\mathbf{\Sigma}}_{\epsilon}$ using r factors
2. Compute $\hat{\sigma}^2_{\epsilon j} = \sum_{i=1}^{k} \left| \hat{\mathbf{\Sigma}}_{\epsilon} \left(i, j \right) \right|$

- 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 $\hat{\mathbf{\Sigma}}_{\epsilon}$ using ξ_{it}
4. Re-estimate factors using quasi-differenced data, iterate if needed

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)
- Method
  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

Other Applications of Factor Models

Factor Augmented VARs

UNIVERSITY OF OXFORD

- Large VARs are challenging since many parameters to estimate
- Small VARs might not have sufficient structure to pick up all shocks
- FAVAR is one solution to these problems

$$\begin{bmatrix} \mathbf{f}_t \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi} & 0 \\ \boldsymbol{\beta}\mathbf{\Phi} - \mathbf{\Xi}\boldsymbol{\beta} & \Xi \end{bmatrix} \begin{bmatrix} \mathbf{f}_{t-1} \\ \mathbf{y}_{t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \boldsymbol{\beta}\mathbf{G} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_t \\ \xi_t \end{bmatrix}$$

  - Ξ is diagonal
  - Can have additional lags
  - Achieves dimension reduction since off-diagonal is determined by diagonal

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