Forecasting With Many predictors

The Econometrics of Predictability

This version: June 9, 2014

June 11, 2014

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

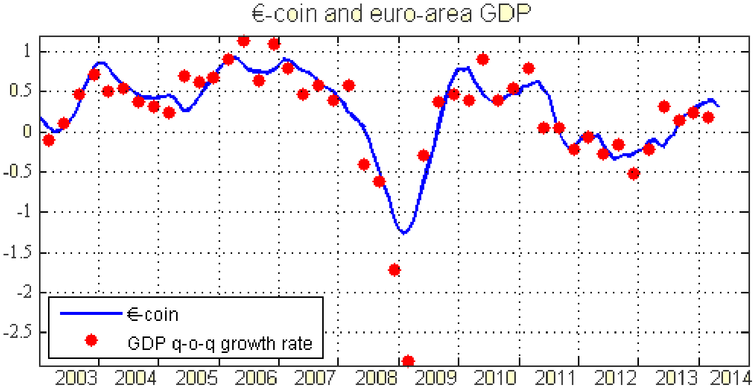
Why factor models

- 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

Eurocoin

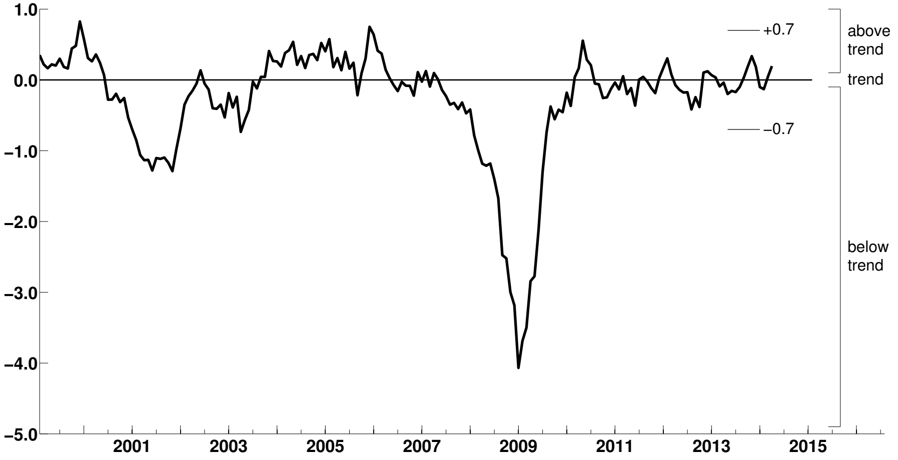
- European Coincident Indicator
- First factor in a Europe-wide model

€-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

- 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

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

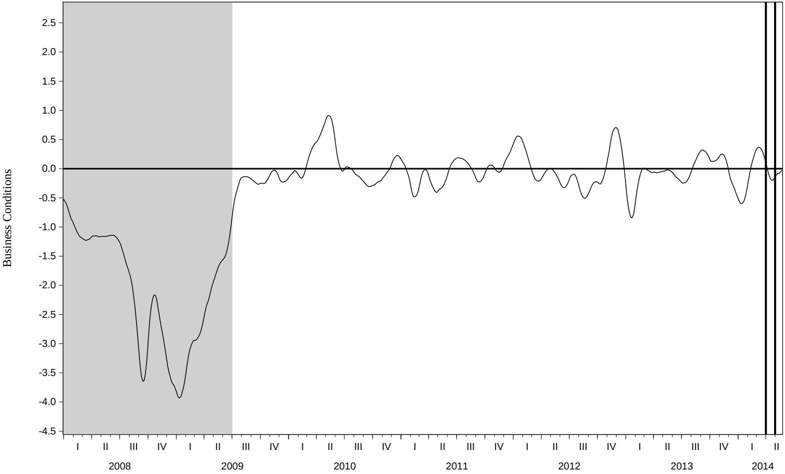
- Indicators

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

  - Δ_i allows series to have different observational frequencies

ADS Business Conditions Index

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



Notation

- 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)

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

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

$$\text{Cov}\left[\begin{array}{c}\epsilon_t\\ \mathbf{x}_t\end{array}\right] = \left[\begin{array}{cc}\sigma_\epsilon^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_k\end{array}\right]$$

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

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

- Degenerate normal - no error since $\boldsymbol{\beta}$ is effectively *known*
- What about the case when k is large
- Use *diagonal* asymptotics, $k/T \rightarrow c$, $0 < \underline{\kappa} < c < \bar{\kappa} < \infty$
- 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 $T \rightarrow \infty$
- True even if all $\boldsymbol{\beta} = \mathbf{0}$!

(Really) Big models don't make sense

- 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

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

- If $k \rightarrow \infty$, $\inf_i |\beta_i| > \underline{\kappa} > 0$, then $\mathrm{V}\left[y_t\right] \rightarrow \infty$
- 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 $O\left({}^1\!/_k\right)$
    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

$$x_{it} = \sum_{j=1}^{r} \lambda_{ij} f_{jt} + \epsilon_{it}$$

- λ_{ij} are the factor loadings for series i, factor j
- ϵ_{it} is the idiosyncratic error for series i
- In vector notation,

$$\underset{k\times 1}{\mathbf{x}_t} = \underset{k\times r}{\mathbf{\Lambda}}\,\underset{r\times 1}{\mathbf{f}_t} + \underset{r\times 1}{\boldsymbol{\epsilon}_t}$$

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

Static Factor Models

- In matrix notation,

$$\underset{T\times k}{\mathbf{X}} = \underset{T\times r}{\mathbf{F}}\underset{r\times k}{\mathbf{\Lambda}'} + \underset{T\times k}{\boldsymbol{\epsilon}}$$

  - $\mathbf{X}$ is T by k
  - $\mathbf{F}$ 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 $\mathrm{E}\left[\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right] = \mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_m^2\right)$
- Covariance of $\mathbf{x}_t$ is then

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

  - $\mathbf{\Omega} = \mathrm{Cov}\left[\mathbf{f_t}\right]$, 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

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

- $\boldsymbol{\beta}$ 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 $\mathbf{B} = \mathbf{Q}\mathbf{A}$ and $\mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}$.

- $\mathbf{f}_t$ 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

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

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

  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

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

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

Eigenvalues and Eigenvectors

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 $\mathbf{A}$, $\lambda_i > 0, i = 1, \ldots, r$ ($\lambda_i \geq 0$)
  - Rank
    - 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

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

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

- $\mathbf{F} = \mathbf{X}\mathbf{V}$ 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

$$x_{it} = 1 \times f_t + \epsilon_{it}$$

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

$$\left[\begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array}\right]$$

- First eigenvector is

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

  - Form is due to normalization

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

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

Estimated Factors

- Estimated factor is then

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

- 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

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

Practical Considerations when Estimating Factors

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

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

- 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

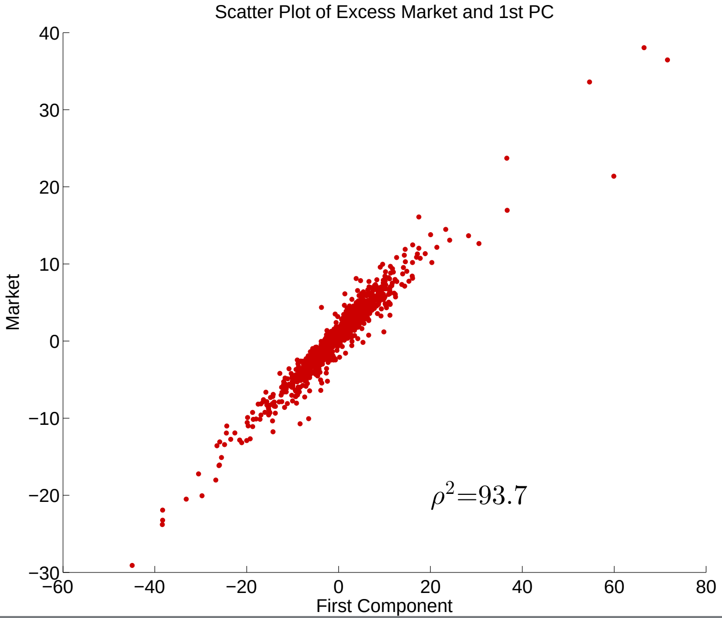
$$\mathbf{R}_{\mathbf{x}} = T^{-1} \sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t'$$

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

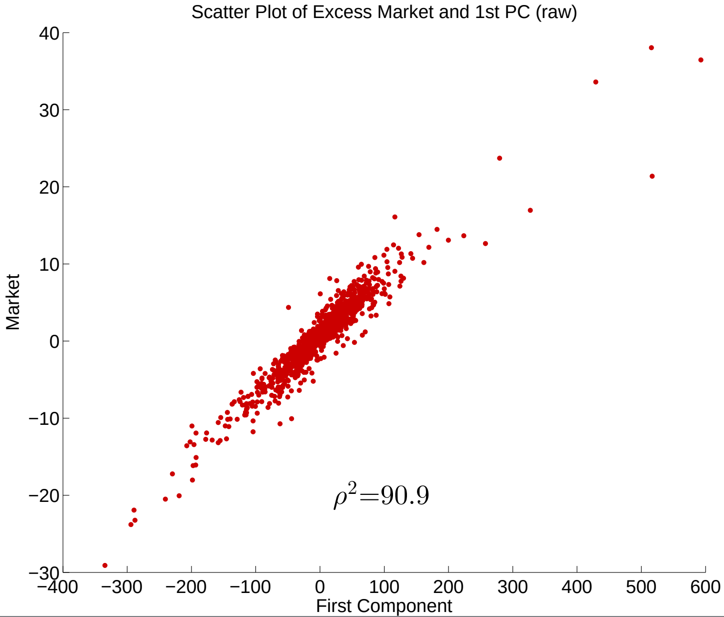
Fama-French Data

- Initial exploration based on Fama-French data
  - 100 portfolios
    - Sorted on size and boot-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)



Selecting the Number of Factors (r)

Choosing the number of factors

- 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?)
    - Closely related to system R^2,

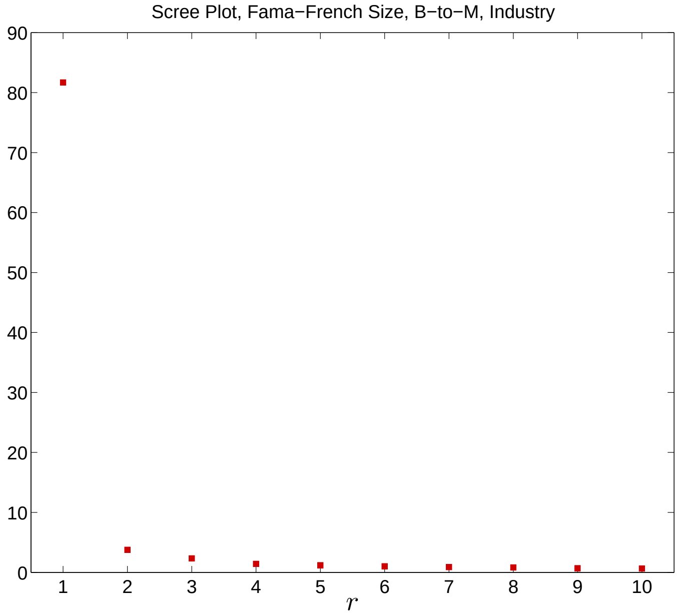
$$R^2\left(r\right) = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{j=1}^{k} \lambda_j}$$

  - Information criteria-based
    - Similar to AIC/BIC, only need to account for both k and T

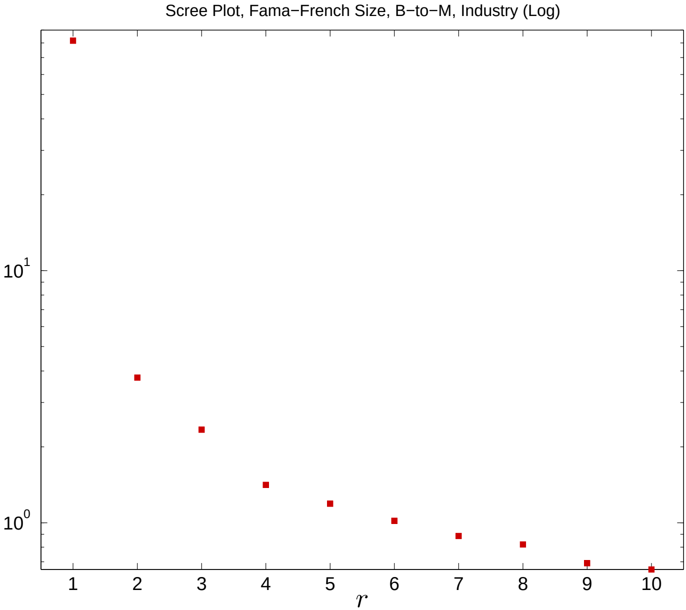
Stylized Fact(ors)

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

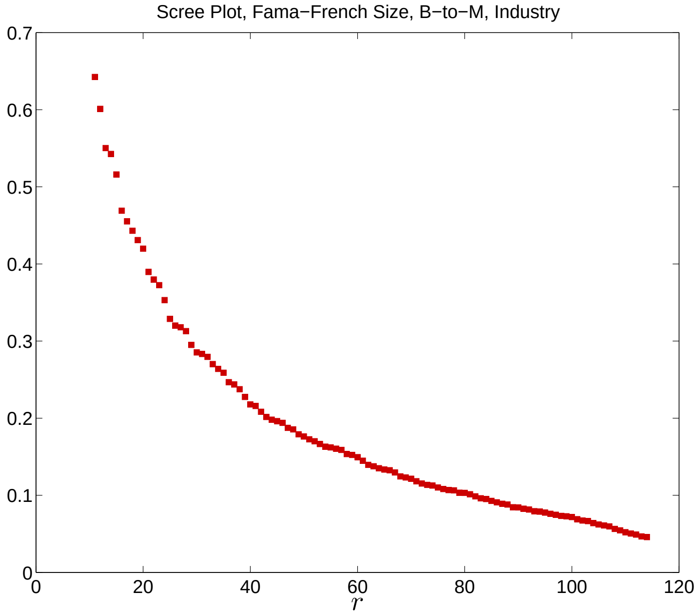
Scree Plot: Fama-French



Scree Plot: Fama-French



Scree Plot: Fama-French (Non-Factors)



Information Criteria

- 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

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

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

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

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

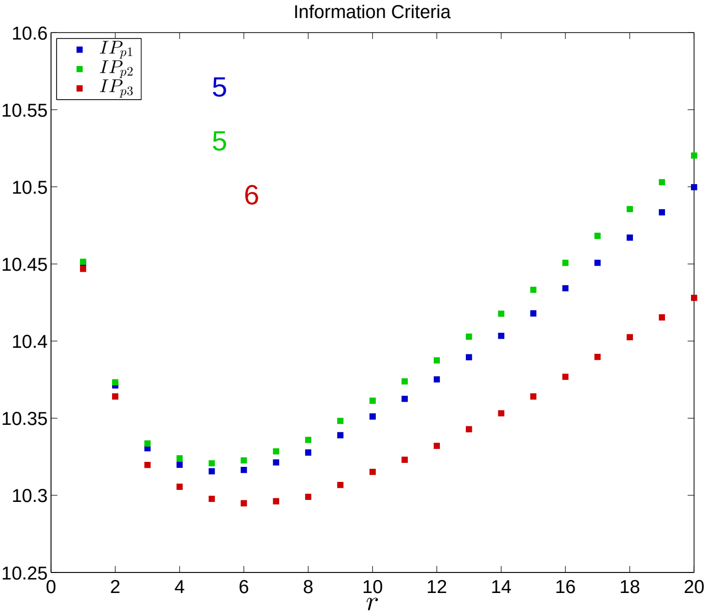
$$IC_{p_2} = \ln \widehat{V(r)} + r\left(\frac{k+T}{kT}\right)\ln\left(\min\left(k, T\right)\right)$$

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

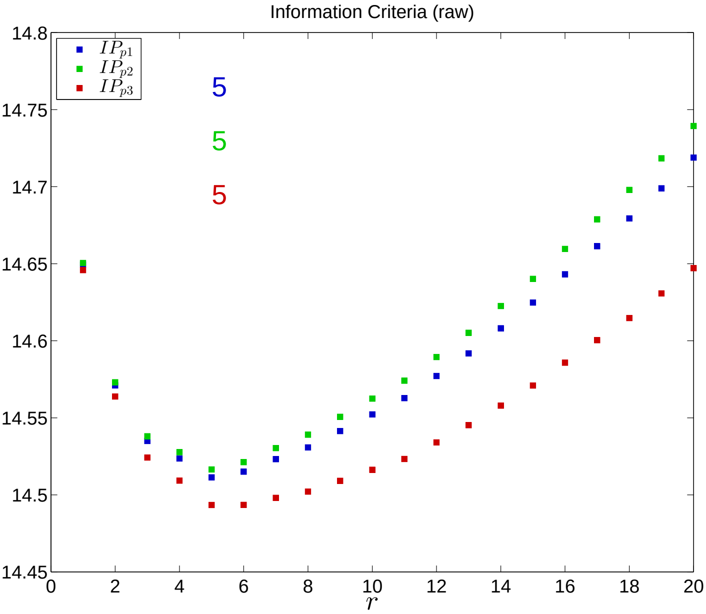
- Suppose $k \approx T$, IC_{p_2} is BIC-like

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

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

$$\begin{aligned} R^2_{\text{global}}(r) &= 1 - \frac{\operatorname{tr}\left(\mathbf{X} - \hat{\boldsymbol{\beta}}(r)\mathbf{F}(r)\right)'\left(\mathbf{X} - \hat{\boldsymbol{\beta}}(r)\mathbf{F}(r)\right)}{\operatorname{tr}(\mathbf{X}'\mathbf{X})} \\ &= \frac{\sum_{i=1}^{r}\lambda_i}{\sum_{j=1}^{k}\lambda_j} \end{aligned}$$

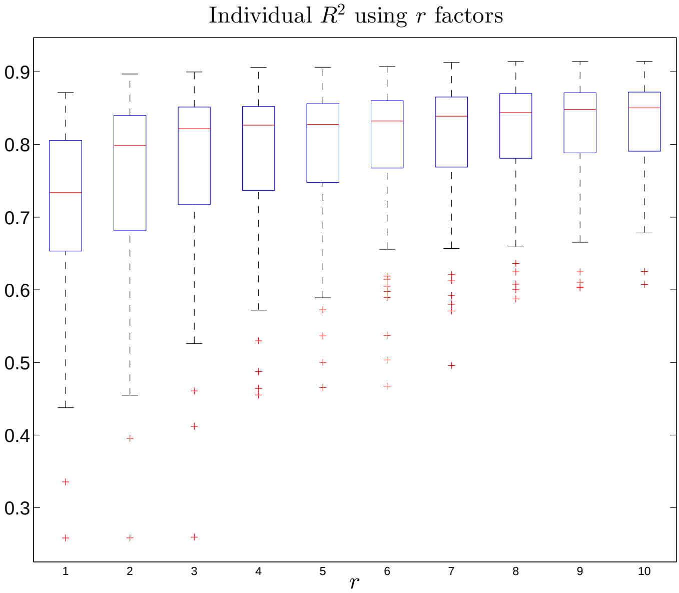
- Numerator is just $\widehat{V(r)} = \sum_{i=1}^{k}\sum_{t=1}^{T}\left(x_{it} - \sum_{j=1}^{r}\hat{\beta}_{ij}f_{jt}\right)^2$
- When $\mathbf{x}$ has been studentized, $\operatorname{tr}\left(\mathbf{X}'\mathbf{X}\right) = \sum_{j=1}^{k}\lambda_j = Tk$
- Individual fit

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

  - Useful for assessing series not well described by factor model

Individual Fit

UNIVERSITY OF OXFORD



Dynamic Factor Models

Dynamic Factor Models

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

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

- 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

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

- Optimal forecast can be derived

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

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

Invertibility and MA processes

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

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

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

Dynamic as Static Factor Models

- 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

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

  - 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
- $k = 100$, $T = 100$
  - 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

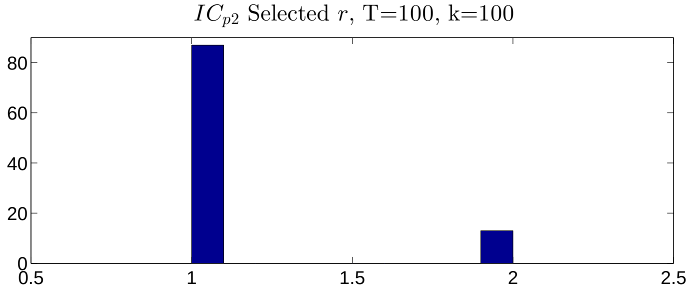
Measuring Closeness of Estimate

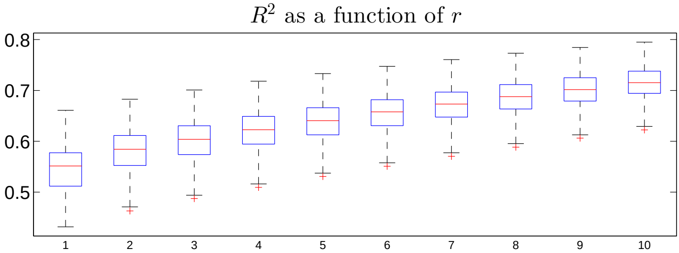
- Factors are not point identified
  - Can use an arbitrary rotation and model is equivalent
- Natural measure of similarity between original (GDP) factors and estimated factors is global R^2

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

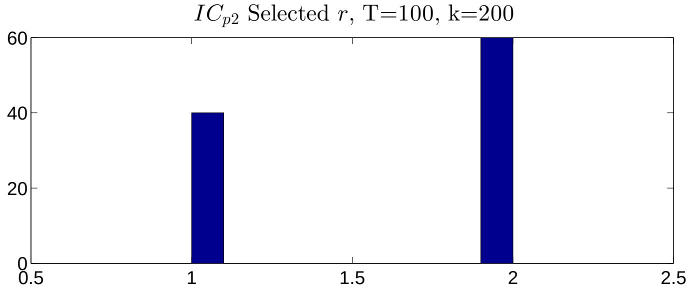
- Note that $\mathbf{A}$ is a 2 by 2 matrix of regression coefficients

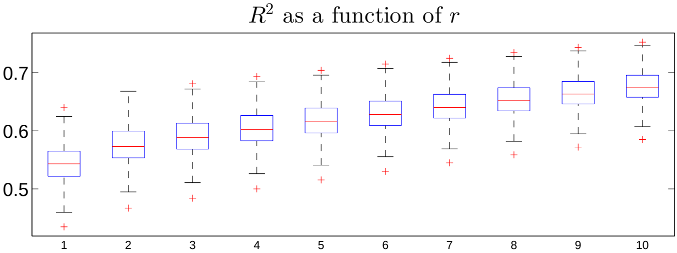
Dynamic as Static Factor Models



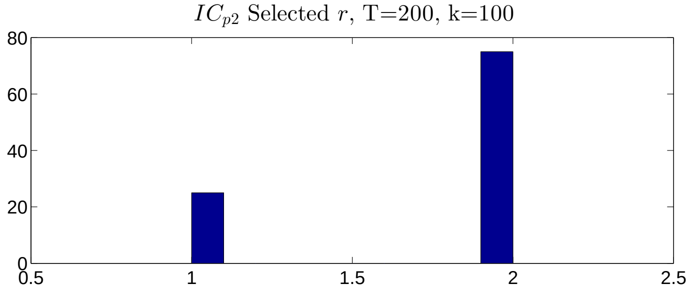


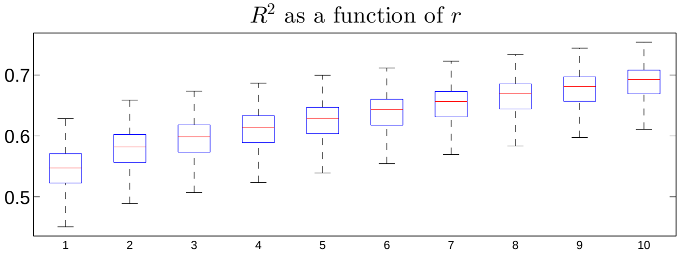
Dynamic as Static Factor Models





Dynamic as Static Factor Models

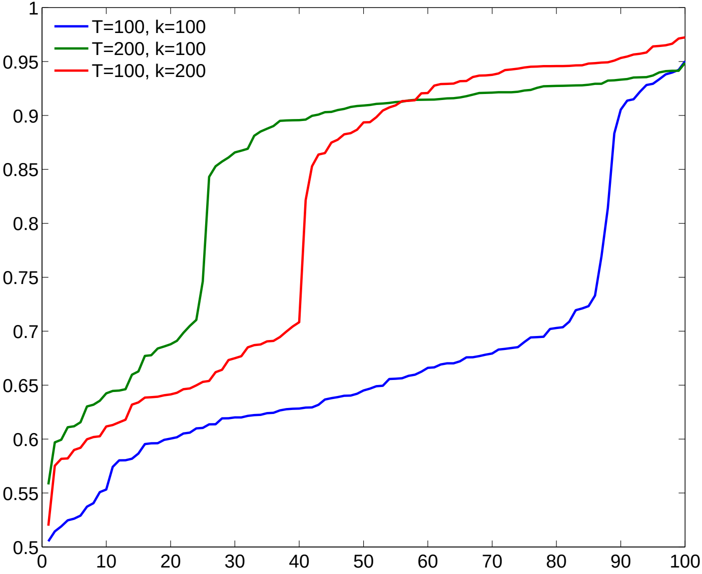




Dynamic as Static Factor Models

UNIVERSITY OF OXFORD

R^2 of factors on estimated factors



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

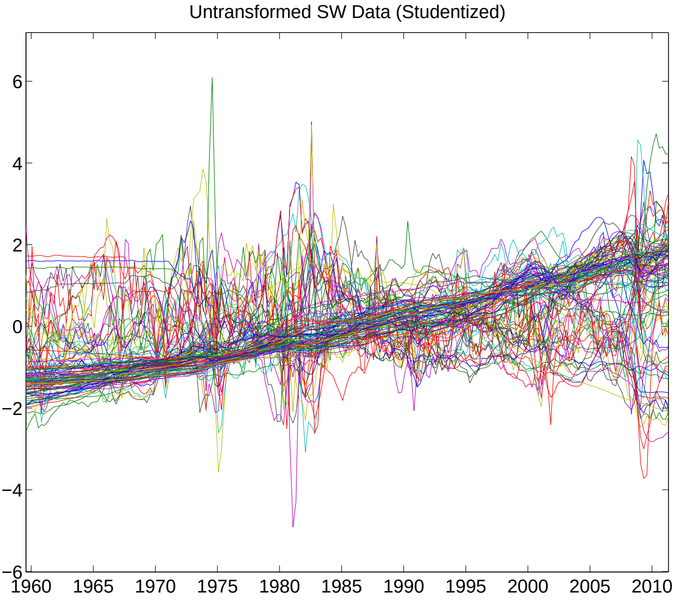
The series

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

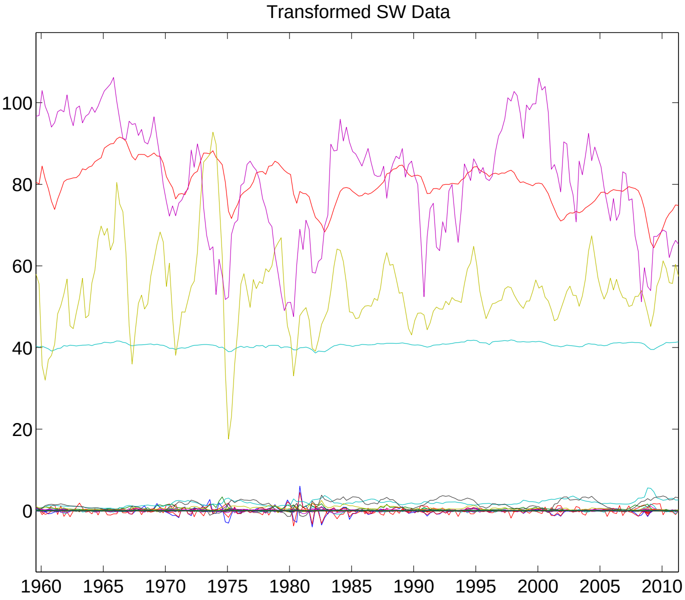
Data Transformation

- 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
- Most series checked for outliers relative to IQR (rare)
- Final series were Studentized in estimation of PC

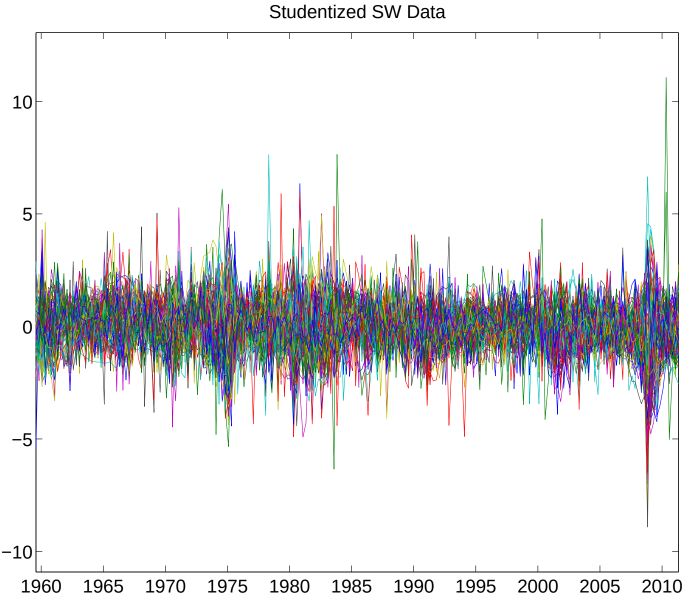
Raw Data Before Transform



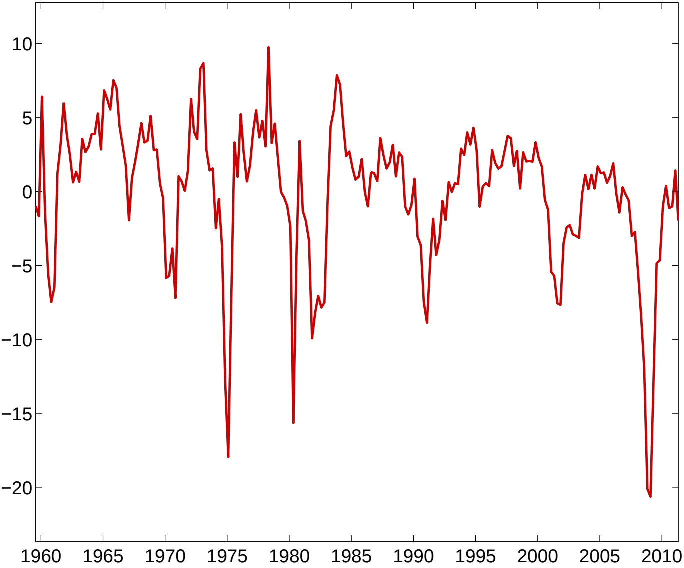
Raw Data after Transform



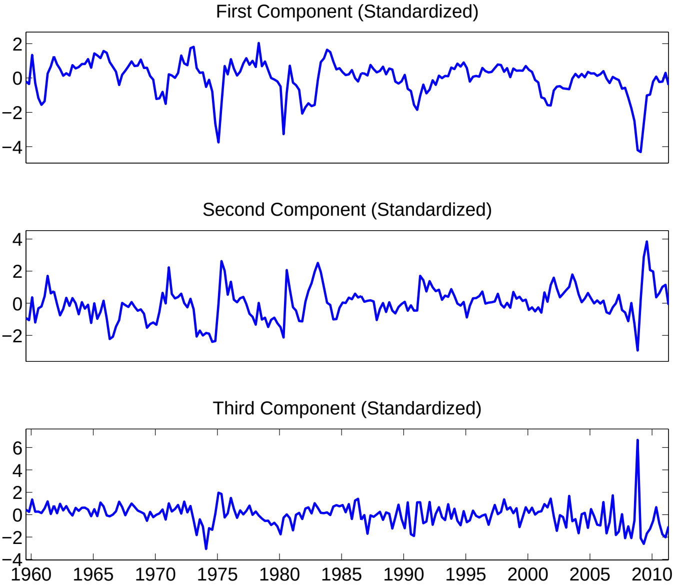
Studentized Data



First Component



First Three Components

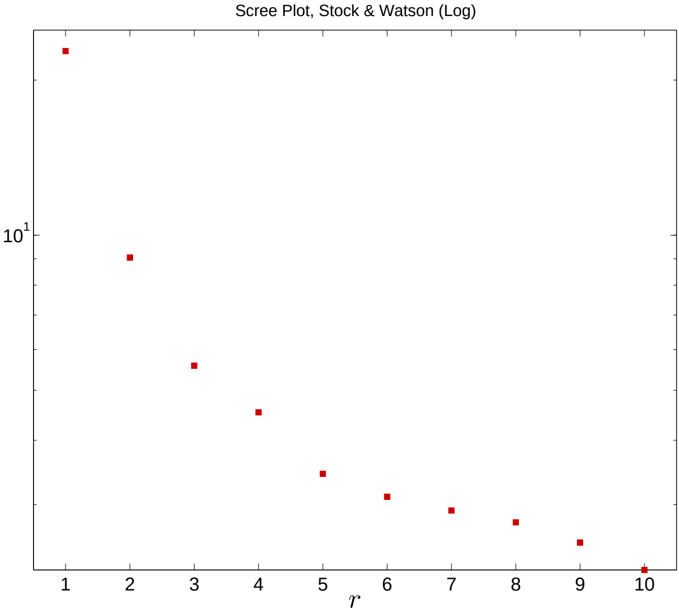


First Component (Standardized)

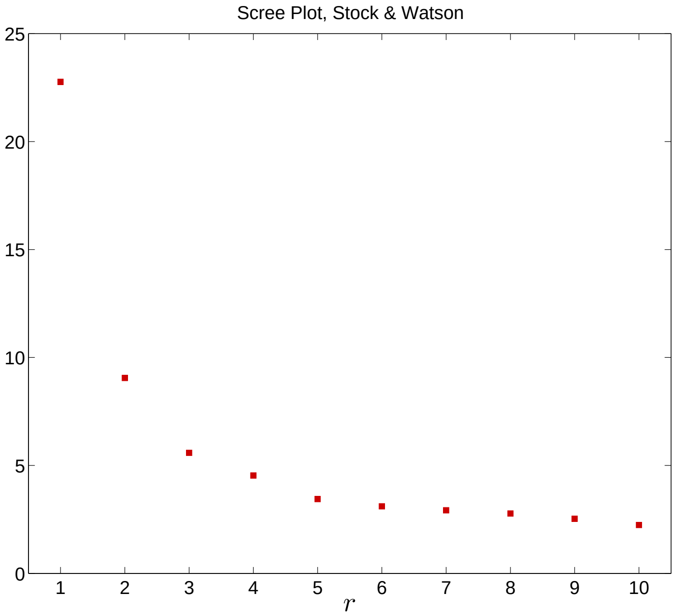
Second Component (Standardized)

Third Component (Standardized)

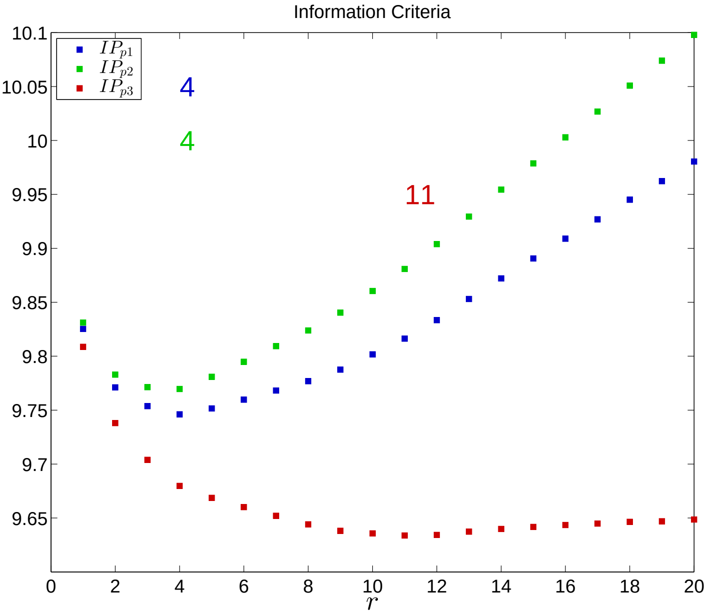
Scree Plot (Log)



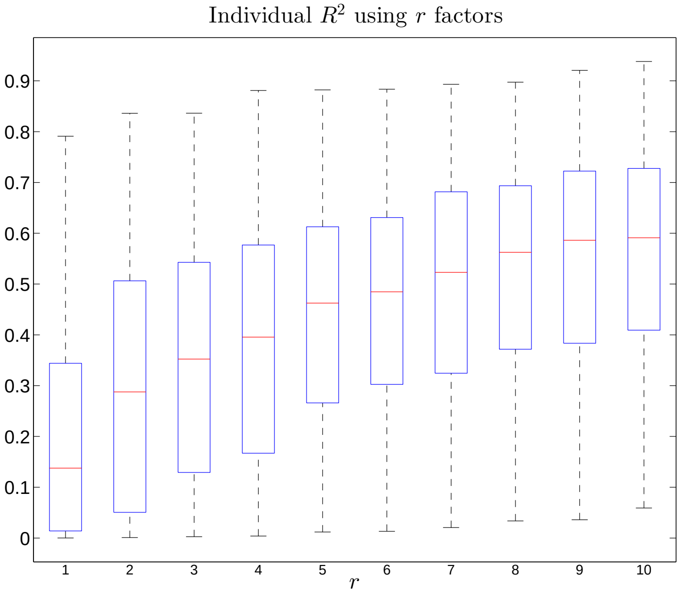
Scree Plot



Information Criteria



Individual Fit against r



Forecasting

Forecast Methods

- 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

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

- 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 also use lags of $\mathbf{f}_t$ (uncommon)
  - Model selection is applicable as usual, e.g. BIC

Forecast Methods

Restricted

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

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

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

- 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

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

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

- At longer horizons,

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

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

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

    - 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

Multistep Forecasting

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

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

  - (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

“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

- Basic PCA makes use of the covariance or more commonly correlation
- Correlation is technically a special case of *generalized PCA*

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

- 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* $\mathbf{W} = \operatorname{diag}\left(w_1, \ldots, w_k\right)$ *where*

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

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

Other Generalized PCA Estimators

- Absolute covariance weighting
  1. Compute complete residual covariance $\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}$ from residuals
  2. Replace $\hat{\sigma}^2_{\epsilon i}$ in step 2 with $\hat{\sigma}^2_{\epsilon i} = \sum_{j=1}^{k} \left|\hat{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}\left(i,j\right)\right|$
- 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

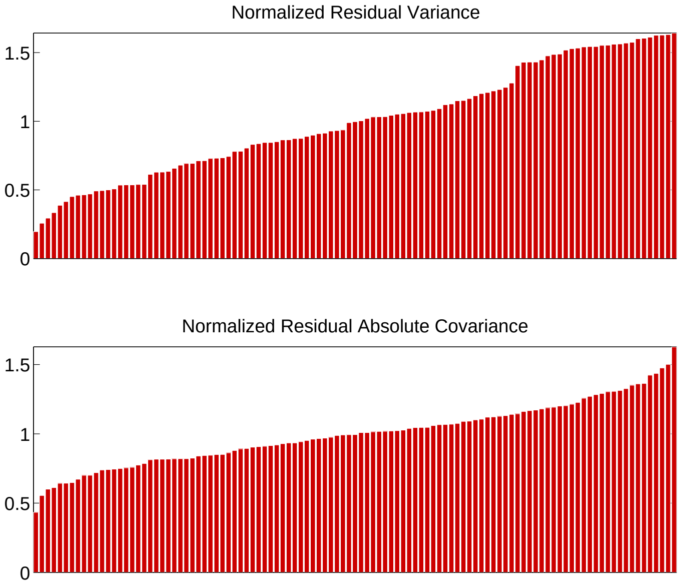
$$\hat{\epsilon}_{it} = \sum_{j=1}^{p_i} \phi_j \epsilon_{it-j} + \xi_{it}$$

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

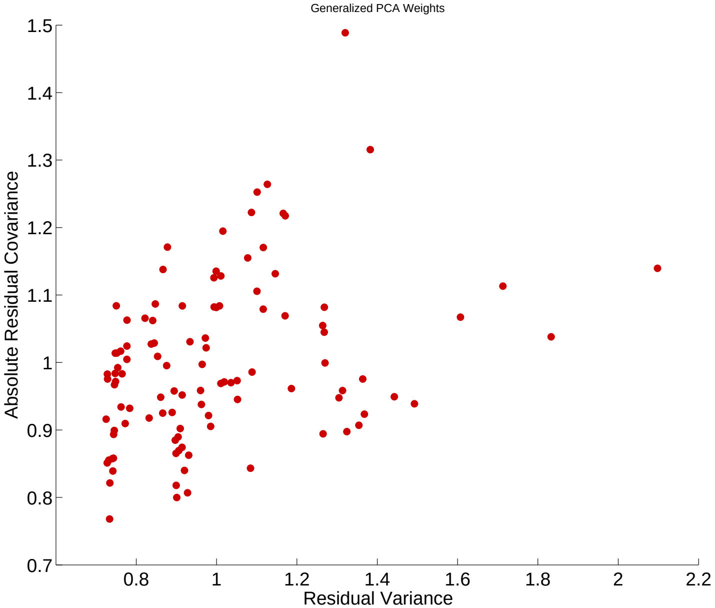
$$\tilde{x}_{it} = x_{it} - \sum_{j=1}^{p_i} \hat{\phi}_j x_{it-j}$$

  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



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

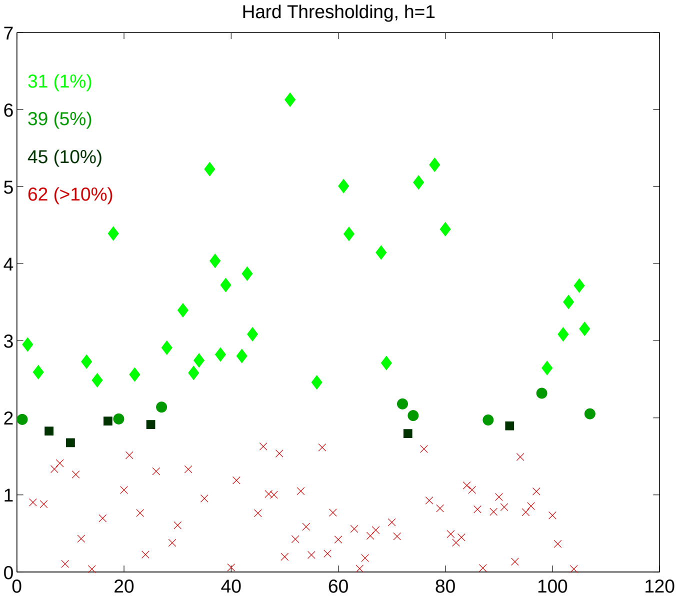
- Bai & Ng (2008) consider problem of selecting *forecasting relevant* factors
- Well known issue for PCA is that factors are selected only using $\mathbf{x}_t$
- Can this be improved using information about y_t?

Algorithm (Hard Thresholding for Variable Selection)

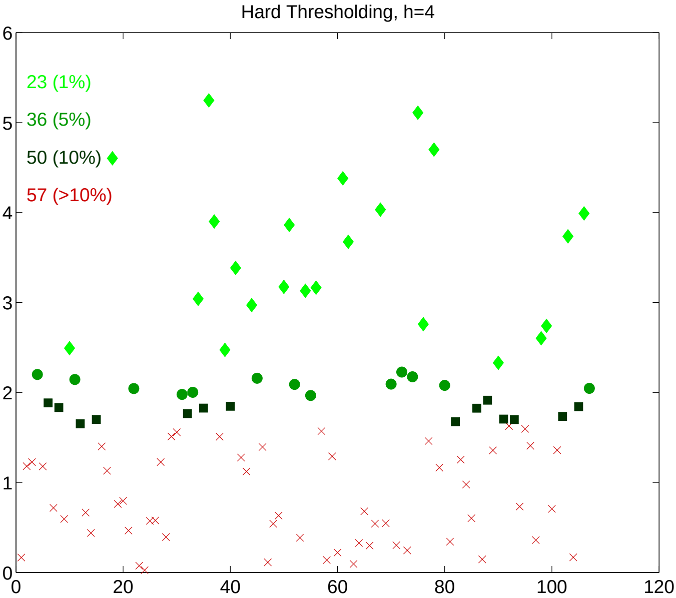
1. *Regress* $y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \gamma x_{t-1} + \epsilon_t$
2. *Compute White heteroskedasticity robust standard errors and t-stat*
3. *Retain any* x_t *where* $|t| > C_\alpha$ *for some choice of* α. *Common choices are 10%, 5% or 1%.*

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

Hard Thresholding for GDP, $h = 1$

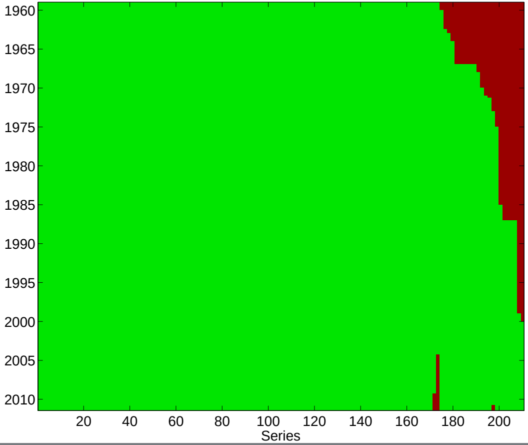


Hard Thresholding for GDP, $h = 4$



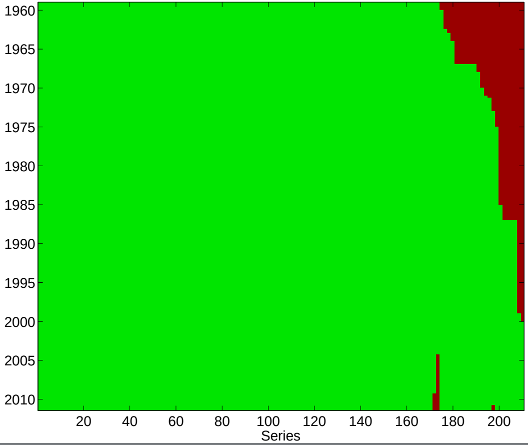
Prinicpal Component Analysis with Missing Data

- 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

- 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* $w_{ij} = I\left[y_{ij} \text{ observed}\right]$ *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

Beyond DFM

- DFMs are an important innovation – both supported by economic theory and statistical evidence
- From a forecasting point of view, they have some limitations
- Alternatives
    - Partial Least Squares Regression
        - Focuses attention on forecasting problem
    - Three-pass Regression Filter
        - Allows focus on factors through *proxies*
    - Regularized Reduced Rank Regression
        - Improve DFM factor selection for forecasting problem
        - Theoretically more sound than using variable selection using BIC

Partial Least Squares

Partial Least Squares

- Partial Least Squares uses the predicted variable when selecting factors
- PCA/DFM only look at $\mathbf{x}_t$ when selecting factors
- The difference means that PLS may have advantages
  - If the factors predicting $\mathbf{y}_t$ are not excessively pervasive
  - If the rotation implied by PCA requires many factors to extract the ideal factor

$$y_{t+1} \quad = \quad \beta f_{1t} + \epsilon_t$$

  - Suppose two estimated factors with the form

$$\begin{bmatrix} \tilde{f}_{1t} \\ \tilde{f}_{2t} \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ \sqrt{1/2} & -\sqrt{1/2} \end{bmatrix} \begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix}$$

  - Correct forecasting model for y_{t+1} requires both $\tilde{f}_{t1}$ and $\tilde{f}_{2t}$

$$\begin{aligned} y_{t+1} &= \gamma_1 \tilde{f}_{1t} + \gamma_2 \tilde{f}_{2t} + \epsilon_t \\ &= \sqrt{1/2}\gamma_1 f_{1t} + \sqrt{1/2}\gamma_2 f_{1t} + \sqrt{1/2}\gamma_1 f_{2t} - \sqrt{2}\gamma_2 f_{2t} + \epsilon_t \\ &= (\gamma_1 + \gamma_2)\sqrt{1/2} f_{1t} + (\gamma_1 - \gamma_2)\sqrt{1/2} f_{2t} + \epsilon_t \end{aligned}$$

  - Implies $\sqrt{1/2}(\gamma_1 + \gamma_2) = \beta$ and $\sqrt{1/2}(\gamma_1 - \gamma_2) = 0$ $(\gamma_1 = \gamma_2, \gamma_1 = \beta / \left(2\sqrt{1/2}\right))$
  - Without this knowledge, 2 parameters to estimate, not 1

Partial Least Squares

- Partial least squares (PLS) uses only bivariate building blocks
- Never requires inverting k by k covariance matrix
  - Computationally very simple
  - Technically no more difficult than PCA
- Uses a basic property of linear regression

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \epsilon_t$$

- Define $\hat{\eta}_t = y_t - \hat{\gamma}_1 x_{1t}$ where $\hat{\gamma}_1$ is from OLS of y on x_1
  - Immediate implication is $\sum \hat{\eta}_t x_{1t} = 0$
- Define $\hat{\xi}_t = \hat{\eta}_t - \hat{\gamma}_2 x_{2t}$ where $\hat{\gamma}_2$ is from OLS of $\hat{\eta}$ on x_2
  - Will have $\sum \hat{\xi}_t x_{2t} = 0$ but also $\sum \hat{\xi}_t x_{1t} = 0$

Partial Least Squares

- Ingredients to PLS are different from PCA
- Assumed model

$$
\begin{aligned}
\mathbf{y}_t &= \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Gamma}\mathbf{f}_{1t} + \boldsymbol{\epsilon}_t \\
\mathbf{x}_t &= \boldsymbol{\Lambda}_1\mathbf{f}_{1t} + \boldsymbol{\Lambda}_2\mathbf{f}_{2t} + \boldsymbol{\xi}_t \\
\mathbf{f}_t &= \Psi\mathbf{f}_{t-1} + \boldsymbol{\eta}_t
\end{aligned}
$$

- Variable to predict is now a key component
  - $\mathbf{y}_t$, m by 1
  - Often $m = 1$
  - Not studentized (important if $m > 1$)
- Same set of predictors
  - $\mathbf{x}_t$, k by 1
  - Assumed studentized
  - $\mathbf{y}_t$ can be in $\mathbf{x}_t$ *if* $\mathbf{y}_t$ is really in the future, so that the values in $\mathbf{x}_t$ are lags
    - Normally $\mathbf{y}_t$ is excluded

Partial Least Squares

UNIVERSITY OF OXFORD

Algorithm (r-Factor Partial Least Squares Regression)

1. *Studentize* $\mathbf{x}_j$ *, set* $\tilde{\mathbf{x}}_j^{(0)} = \mathbf{x}_j$ *and* $\mathbf{f}_{0t} = \boldsymbol{\iota}$
2. *For* $i = 1, \ldots, r$
   a. *Set* $\mathbf{f}_{it} = \sum_j c_{ij} \tilde{\mathbf{x}}_t^{(i-1)}$ *where* $c_{ij} = \sum_t \tilde{\mathbf{x}}_{jt}^{(i-1)} \mathbf{y}_t$
   b. *Update* $\tilde{\mathbf{x}}_j^{(i)} = \tilde{\mathbf{x}}_j^{(i-1)} - \kappa_{ij} \mathbf{f}_t$ *where*

$$\kappa_{ij} = \frac{\mathbf{f}_i' \tilde{\mathbf{x}}_j^{(i-1)}}{\mathbf{f}_i' \mathbf{f}_i}$$

- Output is a set of uncorrelated factors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_r$
- Forecasting model is then $\mathbf{y}_t = \beta_0 + \boldsymbol{\beta}' \mathbf{f}_t + \boldsymbol{\epsilon}_t$
- Useful to save $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_r]$ and $\mathbf{K} = [\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_r]$ and $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}'\right)$
  - Numerical issues may produce some non-zero covariance for factors far apart
  - Normally only interested in a small number, so not important

Factors in PLS

- Factors are just linear combinations of original data
- Obvious for first factor, which is just $\mathbf{f_1} = \mathbf{X}\mathbf{c}_1 = \tilde{\mathbf{X}}^{(0)}\mathbf{c}_1$
- Second factors is $\mathbf{f}_2 = \tilde{\mathbf{X}}^{(1)}\mathbf{c}_2$

$$\begin{aligned}
\tilde{\mathbf{X}}^{(1)} &= \mathbf{X}\left(\mathbf{I}_k - \mathbf{c}_1\boldsymbol{\kappa}_1'\right) \\
&= \mathbf{X} - (\mathbf{X}\mathbf{c}_1)\,\boldsymbol{\kappa}_1' \\
&= \mathbf{X} - \mathbf{f}_1\boldsymbol{\kappa}_1' \\
\tilde{\mathbf{X}}^{(1)}\mathbf{c}_2 &= \tilde{\mathbf{X}}^{(0)}\left(\mathbf{I}_k - \mathbf{c}_1\boldsymbol{\kappa}_1\right)\mathbf{c}_2 \\
&= \mathbf{X}\boldsymbol{\beta}_2
\end{aligned}$$

- Same logic holds for any factor

$$\begin{aligned}
\tilde{\mathbf{X}}^{(j-1)}\mathbf{c}_j &= \tilde{\mathbf{X}}^{(j-2)}\left(\mathbf{I}_k - \mathbf{c}_{j-1}\boldsymbol{\kappa}_{j-1}'\right)\mathbf{c}_j \\
&= \tilde{\mathbf{X}}^{(j-3)}\left(\mathbf{I}_k - \mathbf{c}_{j-2}\boldsymbol{\kappa}_{j-2}'\right)\left(\mathbf{I}_k - \mathbf{c}_{j-1}\boldsymbol{\kappa}_{j-1}'\right)\mathbf{c}_j \\
&= \mathbf{X}\left(\mathbf{I}_k - \mathbf{c}_1\boldsymbol{\kappa}_1'\right)\ldots\left(\mathbf{I}_k - \mathbf{c}_{j-1}\boldsymbol{\kappa}_{j-1}'\right)\mathbf{c}_j \\
&= \mathbf{X}\boldsymbol{\beta}_j
\end{aligned}$$

Forecasting with Partial Least Squares

- When forecasting y_{t+h} , use

$$\mathbf{y} = \begin{bmatrix} y_{1+h} \\ \vdots \\ y_t \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_{t-h} \end{bmatrix}$$

- When studentizing $\mathbf{X}$ save $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$, the vectors of means and variance
  - Alternatively studentize all t observations of $\mathbf{X}$, but only use $1, \ldots, t-h$ in PLS
- Important inputs to preserve:
  - $\mathbf{c}_i$ and $\boldsymbol{\kappa}_i$, $i = 1, 2, \ldots, r$

Algorithm (Out-of-sample Factor Reconstruction)

```
1. Set $f_{0t} = 1$ and $\tilde{\mathbf{x}}_t^{(0)} = (\mathbf{x}_t - \hat{\boldsymbol{\mu}}) \oslash \hat{\boldsymbol{\sigma}}$
2. For $i = 1, \ldots, r$
     a. Compute $f_{it} = \mathbf{c}_i' \tilde{\mathbf{x}}_t^{(i-1)}$
     b. Set $\tilde{\mathbf{x}}_t^{(i)} = \tilde{\mathbf{x}}_t^{(i-1)} - f_{it} \boldsymbol{\kappa}_i'$
```

- Construct forecast from $\mathbf{f}_t$ and $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)$

Comparing PCA and PLS

- There is a non-trivial relationship between PCA and PLS
- PCA iteratively solves the following problem to find $\mathbf{f}_i = \mathbf{X}\boldsymbol{\beta}_i$

$$\max_{\boldsymbol{\beta}_i} \mathrm{V}\left[\mathbf{X}\boldsymbol{\beta}_i\right] \text{ subject to } \boldsymbol{\beta}_i'\boldsymbol{\beta}_i = 1 \text{ and } \mathbf{f}_i'\mathbf{f}_j = 0,\ \ j < i$$

- PLS solves a similar problem to find $\mathbf{f}_i$
  - Different in one important way

$$\max_{\boldsymbol{\beta}_i} \mathrm{Corr}^2\left[\mathbf{X}\boldsymbol{\beta}_i, \mathbf{y}\right] \mathrm{V}\left[\mathbf{X}\boldsymbol{\beta}_i\right] \text{ subject to } \mathbf{f}_i'\mathbf{f}_j = 0,\ \ j < i$$

  - Assumes single y ($m = 1$)
- Implications:
  - PLS can only find factors that are common to $\mathbf{x}_t$ and y_t due to Corr term
  - PLS also cares about the factor space in $\mathbf{x}_t$, so more repetition of one factor in $\mathbf{x}_t$ will affect factor selected
- When $\mathbf{x}_t = \mathbf{y}_t$, PLS is equivalent to PCA

The Three-pass Regression Filter

Three-pass Regression Filter

- Generalization of PLS to incorporate user forecast proxizes, $\mathbf{z}_t$
- When proxies are not specified, proxies can be automatically generated, very close to PLS
- Model structure

$$\begin{aligned}\mathbf{x}_t &= \boldsymbol{\lambda} + \boldsymbol{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t \\ y_{t+1} &= \beta_0 + \boldsymbol{\beta}'\mathbf{f}_t + \eta_t \\ \mathbf{z}_t &= \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\mathbf{f}_t + \boldsymbol{\xi}_t\end{aligned}$$

  - $\mathbf{f}_t = \left[\mathbf{f}'_{1t}, \mathbf{f}'_{2t}\right]'$
  - $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2]$, $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1, \mathbf{0}\right]$, $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2]$
- $\boldsymbol{\beta}$ can have 0's so that some factors are not important for y_{t+1}
- Most discussion is on a single scalar y, so $m = 1$
- $\mathbf{z}_t$ is l by 1, with $0 < l \ll \min\left(k, T\right)$
  - l is finite
  - Number of factors used in forecasting model

Three-pass Regression Filter

UNIVERSITY OF OXFORD

Algorithm (Three-pass Regression Filter)

1. *(Time series regression) Regress* $\mathbf{x}_i$ *on* $\mathbf{Z}$ *for* $i = 1, \ldots, k$, $x_{it} = \phi_{i0} + \mathbf{z}_t' \boldsymbol{\phi}_i + \nu_{it}$
2. *(Cross section regression) Regress* $\mathbf{x}_t$ *on* $\hat{\boldsymbol{\phi}}_i$ *for* $t = 1, \ldots, T$, $x_{it} = \gamma_{i0} + \hat{\boldsymbol{\phi}}_i \mathbf{f}_t + \upsilon_{it}$. *Estimate is* $\hat{\mathbf{f}}_t$.
3. *(Predictive regression) Regress* y_{t+1} *on* $\hat{\mathbf{f}}_t$, $y_{t+1} = \beta_0 + \boldsymbol{\beta}' \hat{\mathbf{f}}_t + \eta_t$

- Final forecast uses out-of-sample data but is otherwise identical
- Trivial to use with an *imbalanced* panel
  - Run step 1 when $\mathbf{x}_i$ is observed
  - Include x_{it} and $\hat{\boldsymbol{\phi}}_i$ whenever observed in step 2
- Imbalanced panel may nto produce accurate forecasts though

Forecasting with Three-pass Regression Filter

- Use data

$$\mathbf{y} = \begin{bmatrix} y_{1+h} \\ y_{2+h} \\ \vdots \\ y_t \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{t-h} \end{bmatrix}$$

  to estimate 3PRF

  - Retain $\hat{\boldsymbol{\phi}}_i$ for $i = 1, \ldots, k$
  - Retain $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$

- To forecast $y_{t+h|t}$

  - Compute $\hat{\mathbf{f}}_t$ by regressing $\mathbf{x}_t$ on $\hat{\boldsymbol{\phi}}_i$ and a constant
  - Construct $\hat{y}_{t+h|t}$ using $\hat{\beta}_0 + \hat{\boldsymbol{\beta}}\hat{\mathbf{f}}_t$

Automatic Proxy Variables

- $\mathbf{z}_t$ are potentially useful but not required

Algorithm (Automatic Proxy Selection)

```
1. Initialize $\mathbf{w}^{(i)} = \mathbf{y}$
2. For $i = 1, 2, \ldots, L$
    a. Set $\mathbf{z}_i = \mathbf{w}^{(i)}$
    b. Compute 3PRF forecast $\hat{\mathbf{y}}^{(i)}$ using proxies $1, \ldots, i$
    c. Update $\mathbf{w}^{(i+1)} = \mathbf{y} - \hat{\mathbf{y}}^{(i)}$
```

- Proxies are natural since forecast errors
- Automatic algorithm finds factor most related to $\mathbf{y}$, then the 1-factor residual, then the 2-factor residual and so on
- Nearly identical to the steps in PLS
- Possibly easier to use 3PRF with missing data

Theory Motivated Proxies

- One of the strengths of 3PRF is the ability to include theory motivated proxies
- Kelly & Pruit show that money growth and output growth can be used to improve inflation proxies over automatic proxies
- The use of theory motivated proxies effectively favors some factors over others
- Potentially useful for removing factors that might be unstable, resulting in poor OOS performance
- When will theory motivated proxies help?
  - Proxies contain common, persistent components
  - Some components in y that are not in $\mathbf{z}$ have unstable relationship

Exact Relationship between 3PRF and PLS

- 3PRF and PLS are identical under the following conditions
  - $\mathbf{X}$ has been studentized
  - The 2-first stages do not include constants
- Factors that come from 3PRF and PLS differ by a rotation
- PLS factors are uncorrelated by design
- Equivalent factors can be constructed using

$$\boldsymbol{\Sigma}_{\mathbf{f}}^{-1/2}\mathbf{F}^{3PRF}$$

  - $\boldsymbol{\Sigma}_{\mathbf{f}}$ is the covariance matrix of $\mathbf{F}^{3PRF}$
  - Will stiff differ by scale and possibly factor of ± 1
  - Order may also differ

Forecasting from DFM and PLS/3PRF

- Forecast
  - GDP growth
  - Industrial Production
  - Equity Returns
  - Spread between Baa and 10 year rate
- All data from Stock & Watson 2012 dataset
- Dataset split in half
  - 1959:2 – 1984:1 for initial estimation
  - 1985:1 – 2011:2 for evaluation
- Consider horizons from 1 to 4 quarters
- Entire procedure is conducted out-of-sample

DFM Components

- Forecasts computed using different methods:
  - 3 components
  - 3 components and 4 lags with Global BIC search
  - IP_{p2} selected components only
- $\mathbf{X}$ recursively studentized
  - Only use series that have no missing data
- Cheating: some macro data-series are not available in real-time, but all forecasts benefit

PLS/3PRF Components and Benchmark

- Consider 1, 2 and 3 factor forecasts
- Automatic proxy selection only
- Always studentize $\mathbf{X}$
- Benchmark is AR(4)

Out-of-sample R^2

IP

PCA(3)	0.6038	0.4255	0.3125	0.2667
AR(4)	0.5521	0.3695	0.2699	0.2031
BIC	0.5671	0.3676	0.3047	0.2936
PCA-IC	0.5380	0.4089	0.3235	0.2773
3PRF-1	0.4653	0.3728	0.2999	0.2601
3PRF-2	0.5351	0.4081	0.3095	0.2494
3PRF-3	0.5230	0.3619	0.2294	0.1600

GDP

PCA(3)	0.6031	0.4204	0.2483	0.1485
AR(4)	0.5239	0.3578	0.2601	0.1860
BIC	0.6210	0.4573	0.2790	0.1669
PCA-IC	0.6010	0.435	0.3046	0.2246
3PRF-1	0.5385	0.4371	0.3444	0.2848
3PRF-2	0.5205	0.3759	0.2665	0.1922
3PRF-3	0.4637	0.2918	0.1796	0.1189

Out-of-sample R^2

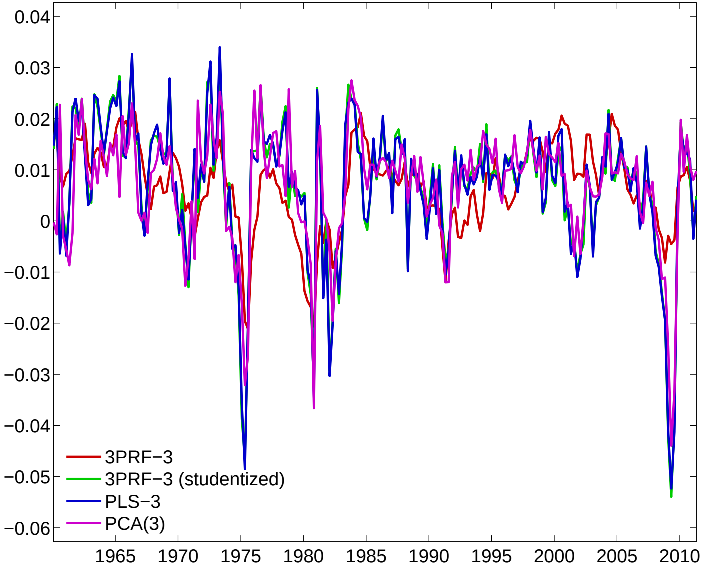
BAA-GS10 (Diff)

PCA(3)	-0.0754	-0.2065	-0.178	-0.0484
AR(4)	-0.0464	-0.0914	-0.0865	-0.0097
BIC	0.0232	-0.1253	-0.0036	-0.0380
PCA-IC	0.0390	-0.0698	-0.0711	0.0242
3PRF-1	-0.0072	-0.1735	-0.1367	-0.0240
3PRF-2	0.0303	-0.1887	-0.1283	-0.0564
3PRF-3	-0.1909	-0.4024	-0.3301	-0.1710

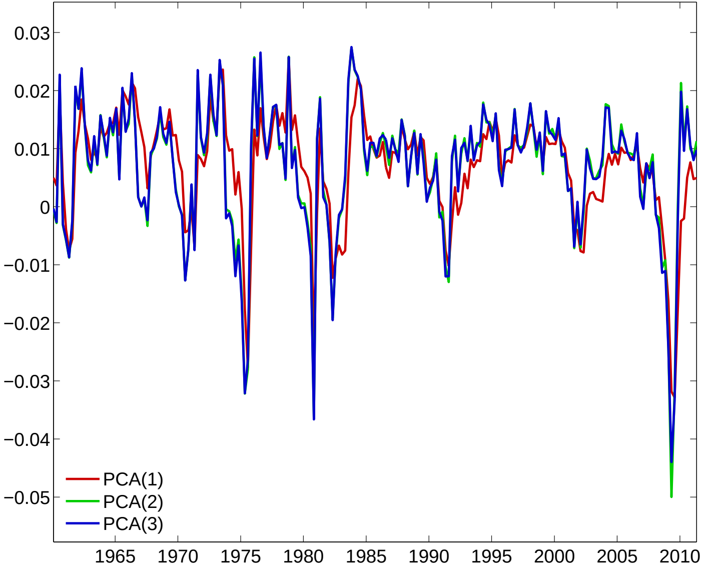
S&P 500 Return

PCA(3)	0.0442	-0.1133	-0.1870	-0.2149
AR(4)	0.0677	-0.0095	-0.0546	-0.0725
BIC	0.0232	-0.1281	-0.1895	-0.1950
PCA-IC	0.0070	-0.0929	-0.0949	-0.0982
3PRF-1	-0.0245	-0.1575	-0.1764	-0.1863
3PRF-2	0.0903	-0.1488	-0.2122	-0.2165
3PRF-3	0.0055	-0.2029	-0.3885	-0.3833

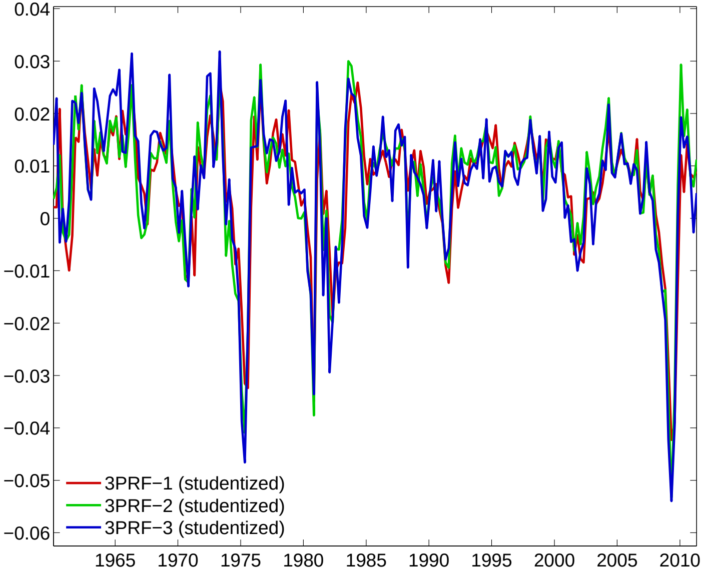
Alternative Fits of GDP



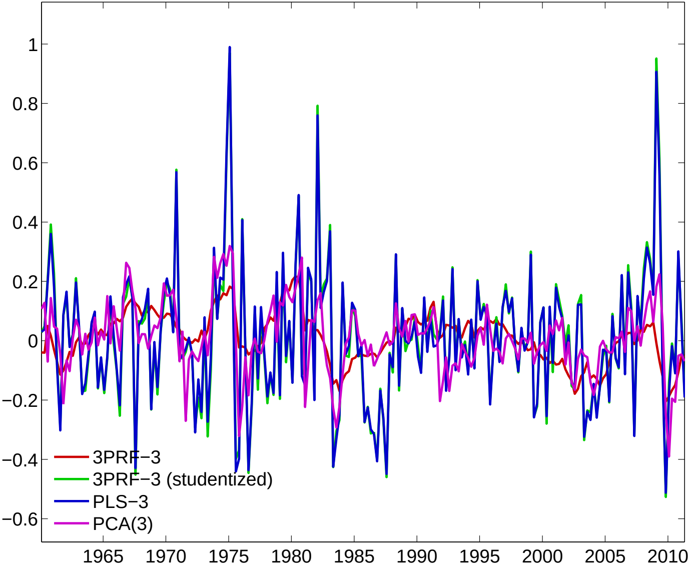
Number of PC and Fit of GDP



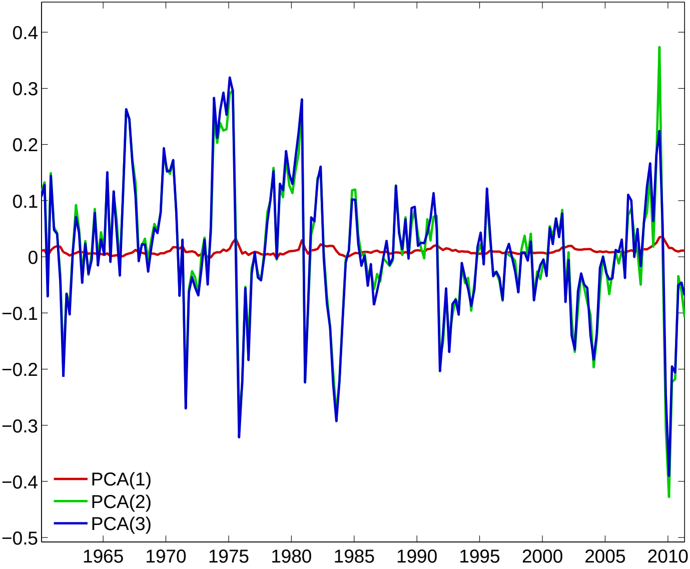
Number of 3PRF Factors and Fit of GDP



Alternative Fits of Baa-10 year spread

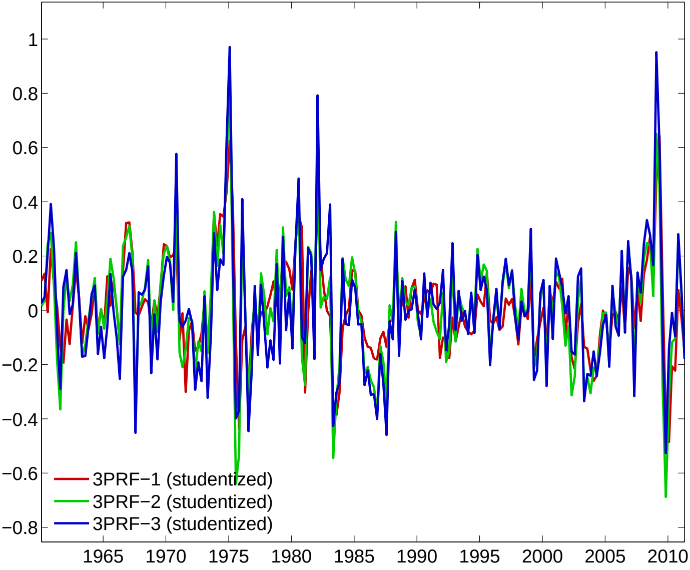


Number of PC and Fit of Spread



Number of 3PRF Factors and Fit of Spread

UNIVERSITY OF OXFORD



Regularized Reduced Rank Regression

Regularized Reduced Rank Regression

- When k is large, OLS will not produce useful forecasts
- Reduced rank regression places some restrictions on the coefficients on $\mathbf{x}_t$

$$\begin{aligned} y_{t+1} &= \gamma_0 + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_t + \epsilon_t \\ &= \gamma_0 + \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x}_t\right) + \epsilon_t \\ &= \gamma_0 + \boldsymbol{\alpha}\mathbf{f}_t + \epsilon_t \end{aligned}$$

  - $\boldsymbol{\alpha}$ is 1 by r – factor loadings
  - $\boldsymbol{\beta}$ is r by k – selects the factors
- When $k \approx T$, even this type of restriction does not produce well behaved forecasts

Regularizing Covariance Matrices

- Regularization is a common method to ensure that covariance matrices are invertible when $k \approx T$, or even when $k > T$
- Many regularization schemes
- Tikhonov

  $$\tilde{\mathbf{\Sigma}}_{\mathbf{x}} = \hat{\mathbf{\Sigma}}_{\mathbf{x}} + \rho \mathbf{Q}\mathbf{Q}'$$

  where $\mathbf{Q}\mathbf{Q}'$ has eigenvalues bounded from 0 for any k

  - Common choice of $\mathbf{Q}\mathbf{Q}'$ is $\mathbf{I}_k$, $\tilde{\mathbf{\Sigma}}_{\mathbf{x}} = \hat{\mathbf{\Sigma}}_{\mathbf{x}} + \rho \mathbf{I}_k$
  - Makes most sense when $\mathbf{x}_t$ has been studentized
- Eigenvalue cleaning

  $$\hat{\mathbf{\Sigma}}_{\mathbf{x}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$$

  - For $i \leq r$, $\tilde{\lambda}_i = \lambda_i$ is unchanged
  - For $i > r$, $\tilde{\lambda}_i = \left(k - r\right)^{-1} \sum_{i>c} \lambda_i$

  $$\tilde{\mathbf{\Sigma}}_{\mathbf{x}} = \mathbf{V}\tilde{\mathbf{\Lambda}}\mathbf{V}'$$

  - Effectively imposes a r-factor structure

Combining Reduced Rank and Regularization

- These two methods can be combined to produce RRRR
- In small k case,

$$y_{t+1} = \gamma_0 + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_t + \epsilon_t$$

normalized$\boldsymbol{\beta}$ can be computed as as solution to generalized eigenvalue problem

  - Normal eigenvalue problem

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

  - Generalized Eigenvalue Problem

$$|\mathbf{A} - \lambda\mathbf{B}| = 0$$

- Reduced Rank LS

$$\left|\underset{k\times m}{\boldsymbol{\Sigma}_{\mathbf{xy}}}\mathbf{W}\underset{m\times k}{\boldsymbol{\Sigma}'_{\mathbf{xy}}} - \lambda\underset{k\times k}{\boldsymbol{\Sigma}_{\mathbf{x}}}\right| = 0$$

$\boldsymbol{\beta}$ are the r generalized eigenvectors associated with the r largest generalized eigenvalues of this problem

  - $\mathbf{W}$ is a weighting matrix, either $\mathbf{I}_m$ or a diagonal GLS version using variance of y_{it} on i$^{\text{th}}$ diagonal

RRRR-Tikhonov

- $\boldsymbol{\beta}$ are the r generalized eigenvectors associated with the r largest generalized eigenvalues of

$$\left|\boldsymbol{\Sigma}_{\mathbf{xy}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xy}}' - \lambda\left(\boldsymbol{\Sigma}_{\mathbf{x}} + \rho\mathbf{QQ}'\right)\right| = 0$$

  - $\mathbf{X}$ is studentized
  - $\mathbf{QQ}'$ is typically set to $\mathbf{I}_k$
  - ρ is a tuning parameter, usually set using 5- or 10-fold cross validation
  - r also need to be selected
    - Cross validation
    - Model-based IC

RRRR-Spectral Cutoff

- $\boldsymbol{\beta}$ are the r generalized eigenvectors associated with the r largest generalized eigenvalues of

$$\left|\mathbf{\Sigma_{fy}W\Sigma_{fy}'} - \lambda\mathbf{\Sigma_f}\right| = 0$$

- $\mathbf{\Sigma_f}$ is the covariance of the first r_f principal components
  - r_f to distinguish from r (the number of columns in $\boldsymbol{\beta}$)
  - $\mathbf{\Sigma_{fy}}$ is the covariance between the PCs and the data to be predicted
  - r_f must be chosen using another criteria – Scree plot or Information Criteria
- The spectral cutoff method essentially chooses a set of r factors from the set of r_f PCs
- This is not a trivial exercise since factors are always identified only up to a rotation
- For example, allows a 1-factor model to be used for forecasting even when the factor can only be reconstructed from all r_f PCs
- Partially bridges the gap between PCA and PLS/3PRF

Forecasting in RRRR

- Once $\hat{\boldsymbol{\beta}}$ was been estimated using generalized eigenvalue problem, run regression

$$y_{t+1} = \phi_0 + \boldsymbol{\alpha}\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_t\right) + \epsilon_t$$

  to estimate $\hat{\boldsymbol{\alpha}}$

- Can also include lags of y

$$y_{t+1} = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i+1} + \boldsymbol{\alpha}\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_t\right) + \epsilon_t$$

- When using spectral cutoff, regressions use $\mathbf{f}_t$ in place of $\mathbf{x}_t$
- Forecasts are simple since $\mathbf{x}_t$, $\hat{\boldsymbol{\beta}}$ and other parameters are known at time t
  - When using spectral cutoff, $\mathbf{f}_t$ is also known at time t
- r can be chosen using a normal IC such as BIC or using t-stats in the forecasting regression

General Setup for Forecasting

- When forecasting with the models, it is useful to setup some matrices so that observations are aligned
- Assume interest in predicting $y_{t+1|t}, \ldots, y_{t+h|t}$
  - Can also easily use cumulative versions, $\mathrm{E}_t\left[\sum_{i=1}^{h} y_{t+i}\right]$
- All matrices will have t rows
- Leads (max h) and lags (max P)

$$\mathbf{Y}^{\text{leads}} = \begin{bmatrix} y_2 & y_3 & \cdots & y_{h+1} \\ y_3 & y_4 & \cdots & y_{h+2} \\ \vdots & \vdots & \vdots & \vdots \\ y_{t-h+1} & y_{t-h+2} & \cdots & y_t \\ y_{t-1} & y_t & \cdots & - \\ y_t & - & \cdots & - \end{bmatrix}, \quad \mathbf{Y}^{\text{lags}} = \begin{bmatrix} y_1 & - & \cdots & - \\ y_2 & y_1 & \cdots & - \\ \vdots & \vdots & \vdots & \vdots \\ y_P & y_{P-1} & \vdots & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ y_{t-1} & y_{t-2} & \vdots & y_{t-P} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \cdots \\ \mathbf{x}_t \end{bmatrix}$$

- $-$ denotes a missing observation (nan)
- When forecasting at horizon h, use column h of $\mathbf{Y}^{\text{leads}}$ and rows $1, \ldots t-h$ of $\mathbf{Y}^{\text{lags}}$ and $\mathbf{X}$
  - Remove any rows that have missing values
- When using PCA methods, extract PC ($\mathbf{C}$) from all of $\mathbf{X}$ and use rows $1, \ldots t-h$ of $\mathbf{C}$