Partial Least Squares, Three-Pass Regression Filters and Reduced Rank Regularized Regression

The Econometrics of Predictability

This version: June 15, 2014

June 16, 2014

UNIVERSITY OF OXFORD

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 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

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

  - 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

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 $\mathbf{f}_2 = \tilde{\mathbf{X}}^{(1)}\mathbf{c}_2$

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

- 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

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

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

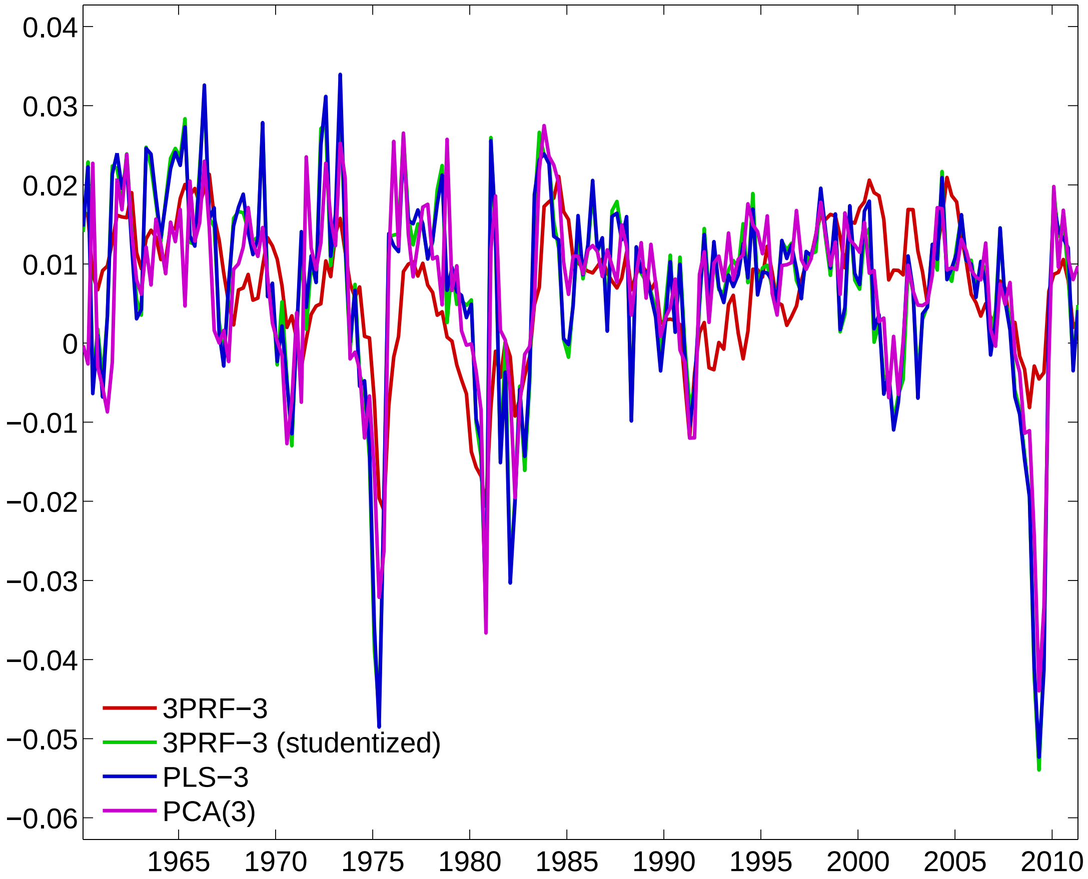
BAA-GS10 (Diff)

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

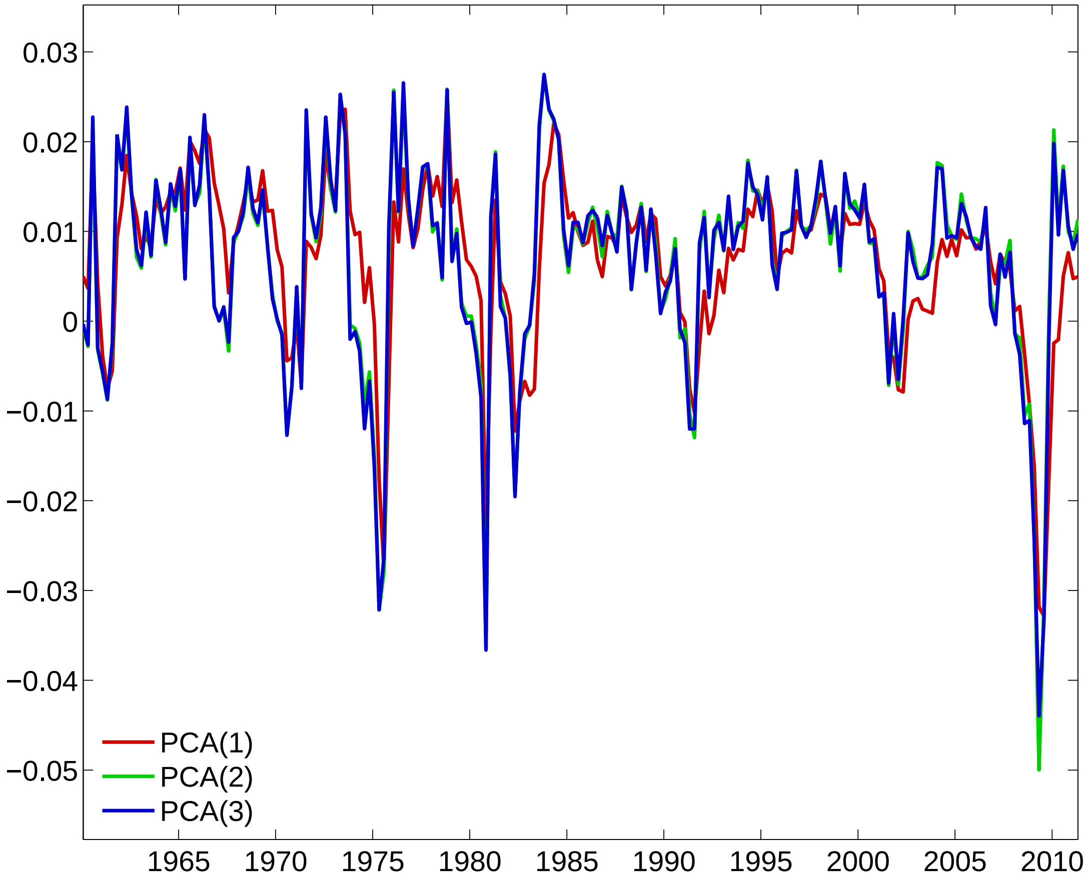
S&P 500 Return

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

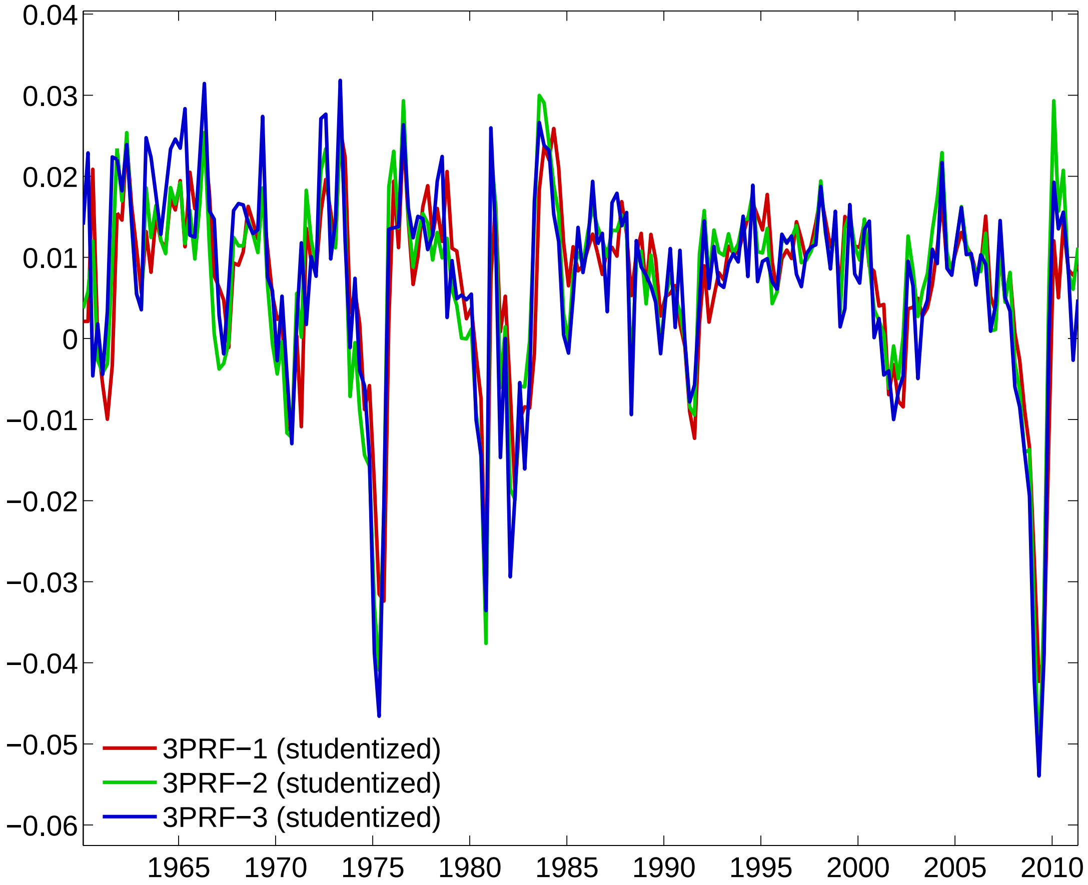
Alternative Fits of GDP



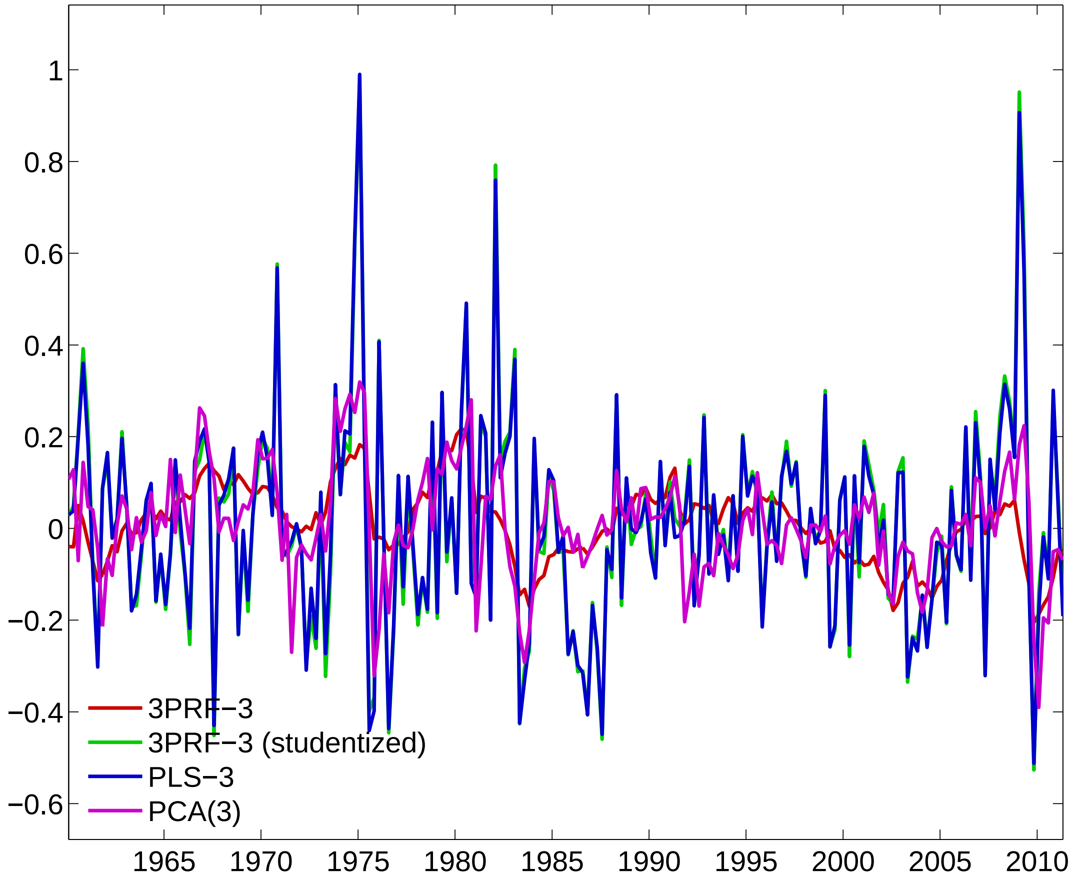
Number of PC and Fit of GDP



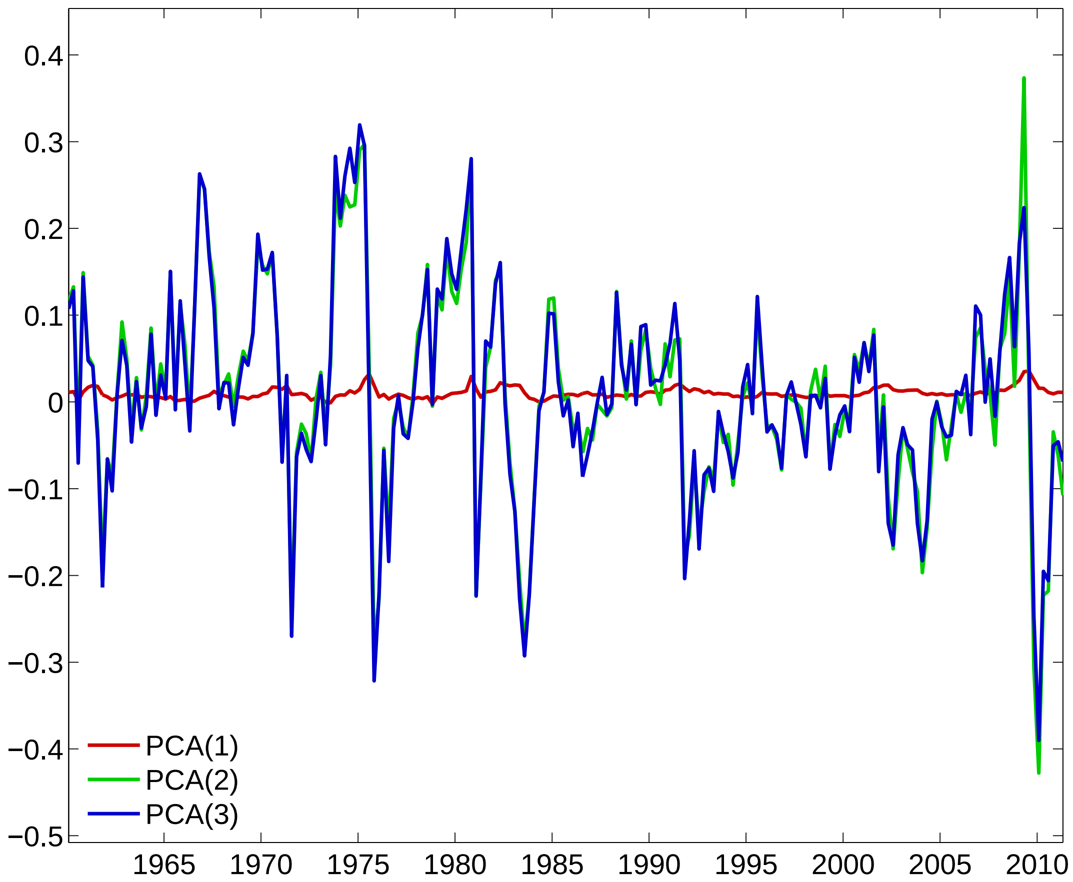
Number of 3PRF Factors and Fit of GDP



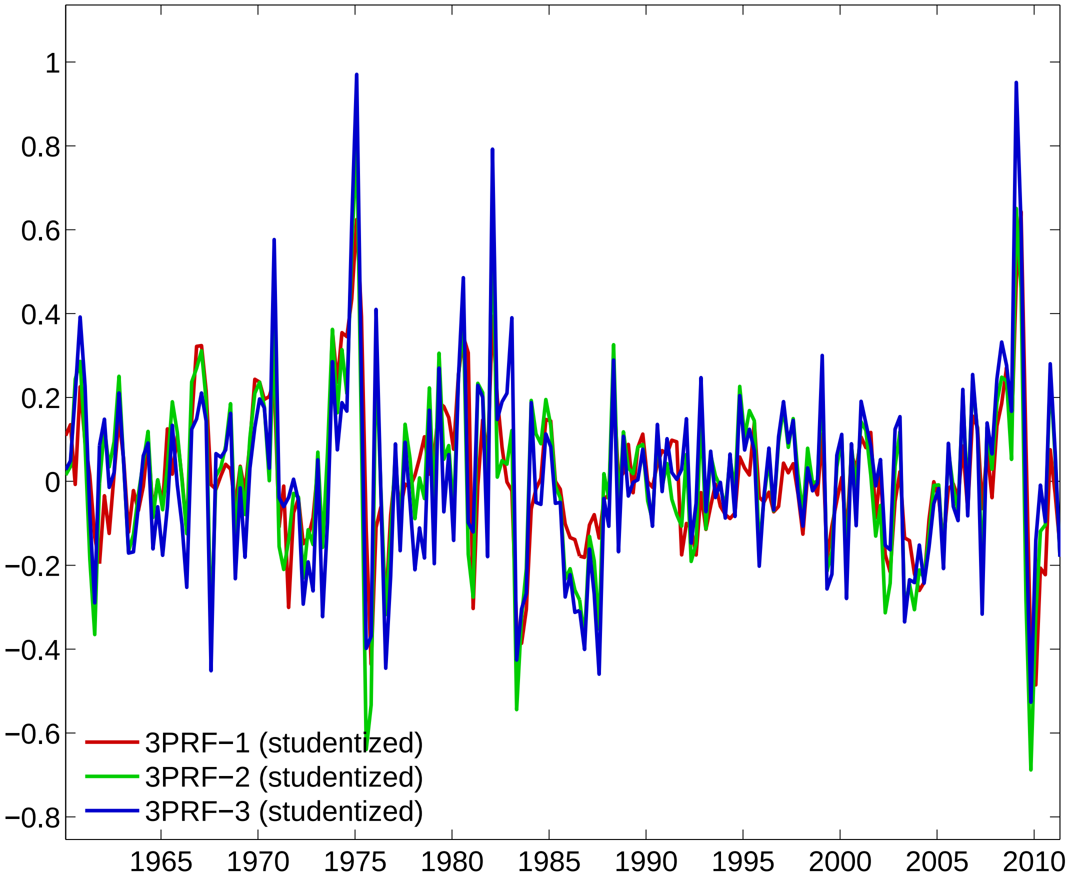
Alternative Fits of Baa-10 year spread



Number of PC and Fit of Spread



Number of 3PRF Factors and Fit of Spread



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

$$\left|\mathbf{\Sigma_{xy}} \mathbf{W} \mathbf{\Sigma}'_{\mathbf{xy}} - \lambda \left(\mathbf{\Sigma_x} + \rho \mathbf{Q}\mathbf{Q}'\right)\right| = 0$$

  - $\mathbf{X}$ 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
    - r will always be less than m, the number of y series: When there is only 1 series, the first eigenvector selects the optimal linear combination which will predict that series the best. Only tension if more than 1 series.

RRRR-Spectral Cutoff

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

$$\left|\boldsymbol{\Sigma}_{\mathbf{fy}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{fy}}' - \lambda\boldsymbol{\Sigma}_{\mathbf{f}}\right| = 0$$

- $\boldsymbol{\Sigma}_{\mathbf{f}}$ is the covariance of the first r_f principal components
  - r_f to distinguish from r (the number of columns in $\boldsymbol{\beta}$)
  - $\boldsymbol{\Sigma}_{\mathbf{fy}}$ 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)

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

- $-$ 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 $\mathbf{C}$