Thresholded Generalized Principal Component Regression: Forecasting with Many Predictors

Mohsen Pourahmadi
Ranye Sun

Texas A&M University

Recent Advances and Trends in Time Series Analysis:
Nonlinear Time Series, High-Dimensional Inference and Beyond

Banff, CN: April 27-May 2, 2014

Modeling Two-Way Dependent Data

- Problem I: How to model 2-way dependency based on **only one realization** of a data matrix?

Modeling Two-Way Dependent Data

- Problem I: How to model 2-way dependency based on **only one realization** of a data matrix?

✓ Time Series: Assume Stationarity,
the ACF or Spectral Density Matrix Will Do the Job.

The Data Matrix

- In traditional multivariate analysis the rows are independent.
- In mult. time series **both rows and columns are corr.**

The Data Matrix

- In traditional multivariate analysis the rows are independent.
- In mult. time series **both rows and columns are corr.**
- Now, it is common to have data matrices where both rows and columns are correlated: Spatial Data, Spatio-temporal fMRI, Microarray (Efron, 2010), e-Commerce (Netflix), Finance, ...
- Names: Transposable Data (Allen and Tibshirani, 2010); Two-way Structured Data (Huang, Shen and Buja, 2009).

How to Model Transposable Data?

- Problem I: How to model 2-way dependency using **only one realization** of a data matrix?

✓ Time Series: Assume Stationarity, the ACF or Spectral Density Matrix Will Do the Job.

How to Model Transposable Data?

- Problem I: How to model 2-way dependency using **only one realization** of a data matrix?

✓ Time Series: Assume Stationarity, the ACF or Spectral Density Matrix Will Do the Job.

- Nowadays: Assume a matrix normal distribution:

$$\begin{aligned} Y &\sim MN_{n,q}(B, \Omega^{-1}, \Sigma^{-1}), \\ OR \;\; vec(Y) &\sim N_{nq}(vec(B), \Sigma^{-1} \otimes \Omega^{-1}), \end{aligned}$$

with **separable covariances**.

- Unrealistic/limited dependence structure.

Multivariate Linear Regression/Prediction

- Model

$$Y = XB + E,$$

where $Y \in R^{n\times q}$, $X \in R^{n\times p}$, $B \in R^{p\times q}$ and E has a matrix normal dist.

- OLS estimator: $\hat{B}_{OLS} = (X'X)^{-1}X'Y$.

- Problem II: How to improve $\hat{B}_{OLS}$ in HD for better prediction?

Reduced Rank Regression

- Finds the LS estimator of B subject to a rank constraint $\text{rank}(B) = r$ (Anderson, 1951).
- Reduces the pq parameters in B to $r(p+q)$ which is linear in p and q.
- Solution involves SVD/PCA of B.

A Simpler Model

- Reduce the regression model

$$Y = XB + E$$

to the "signal plus noise" model:

$$\begin{aligned} (X'X)^{-1}X'Y &= B + (X'X)^{-1}X'E \\ \hat{B}_{OLS} &= B + \tilde{E} \end{aligned}$$

- Low-rank/sparse estimation of B has been studied when the entries of the error matrix are i.i.d.:
Shen and Huang (2008); Yang, Buja and Ma (2013);
Allen, Grosenick and Taylor (2013).

The Singular Value Decomposition (SVD)

Let Y be an $n \times q$ matrix of rank m. Then,
(a) there exist matrices U, V and D such that

$$Y = UDV' = \sum_{i=1}^{m} d_i \mathbf{u}_i \mathbf{v}_i',$$

where the columns of $U = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$, $V = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ are orthonormal, and the diagonal entries of $D = \text{diag}(d_1, \ldots, d_m)$ are ordered: $d_1 \geq d_2 \geq \ldots \geq d_m > 0$.

The columns of U and V are called the *left- and right-singular vectors* of Y, and the diagonal entries of D are the corresponding *singular values*.

Rank-r approximation

(b) (Eckart-Young Theorem,1936): For any $r \leq m$, the best rank-r approximation to Y in the Frobenius norm is

$$Y^{(r)} = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i.$$

More precisely,

$$\begin{aligned} Y^{(r)} &= \arg \min_{rank(B)=r} ||Y - B||_F^2 \\ &= \arg \min_{rank(B)=r} tr\{(Y - B)'(Y - B)\}. \end{aligned}$$

Rank-r approximation: PCA

- The SVD represents Y as the sum of m orthogonal *layers* of decreasing importance.
- Use the first few SVD layers corresponding to larger d_i values, ignore the rest or treat them as noise.

Rank-r approximation: PCA

- The SVD represents Y as the sum of m orthogonal *layers* of decreasing importance.
- Use the first few SVD layers corresponding to larger d_i values, ignore the rest or treat them as noise.
- SVD and PCA deal with decompositions of Y and $Y'Y$, respectively.
- The right singular vectors in V are the eigenvectors of the sample cov. matrix or its PC loading matrix. The PCs are the columns of YV.
- **Remark**: Principal Component Regression (PCR) uses the first few PCs as the predictors.

Computing the SVD: Power method

- Starting with $\mathbf{v}^{(0)}$, iterate
  1. $\mathbf{u}^{(k)} = Y\mathbf{v}^{(k-1)} / ||Y\mathbf{v}^{(k-1)}||$,
  2. $\mathbf{v}^{(k)} = Y'\mathbf{u}^{(k)} / ||Y'\mathbf{u}^{(k)}||$,

  sequentially until convergence to $\mathbf{u}$ and $\mathbf{v}$. Compute $d = \mathbf{u}'Y\mathbf{v}$.

- Then, apply steps 1-2 to the residual matrix $Y - d\mathbf{u}\mathbf{v}'$.

- Next, ALL the singular vectors are computed **simultaneously**.

The Orthogonal Subspace Iteration

Starting with $V^{(0)}$

1. Multiplication: $Y_L^{(k)} = YV^{(k-1)}$,
2. QR Decomposition: $U^{(k)}R_u^{(k)} = Y_L^{(k)}$,
3. Multiplication: $Y_R^{(k)} = Y'U^{(k)}$,
4. QR Decomposition: $V^{(k)}R_v^{(k)} = Y_R^{(k)}$.

Golub and Van Loan (1996)

1. $\mathbf{u}^{(k)} = Y\mathbf{v}^{(k-1)} / ||Y\mathbf{v}^{(k-1)}||$,
2. $\mathbf{v}^{(k)} = Y'\mathbf{u}^{(k)} / ||Y'\mathbf{u}^{(k)}||$,

Inconsistency of U, V in high dim.

- Silverman (1996); Paul (2007); Johnstone and Lu (2009).
- Penalize the singular values to control the rank (Yuan et al., 2007; Bunea et al., 2011).
- Penalize the singular vectors to induce sparsity (Huang et al. 2009; Witten et al., 2009).

Regularization of the singular vectors

- Minimize the objective function:

$$||Y - d\mathbf{u}\mathbf{v}||_F^2 + P_\lambda(\mathbf{u}, \mathbf{v})$$

  - $P_\lambda(\mathbf{u}, \mathbf{v}) = \lambda_u ||\mathbf{u}||_1 + \lambda_v ||\mathbf{v}||_1$

- Sequentially solve for $(d_i, \mathbf{u}_i, \mathbf{v}_i)$, $i \in \{1 \cdots m\}$:
  e.g. $Y_2 = Y - d_1 \mathbf{u}_1 \mathbf{v}_1$

Regularization of the singular vectors

- Minimize the objective function:

$$||Y - d\mathbf{u}\mathbf{v}||_F^2 + P_\lambda(\mathbf{u}, \mathbf{v})$$

  - $P_\lambda(\mathbf{u}, \mathbf{v}) = \lambda_u ||\mathbf{u}||_1 + \lambda_v ||\mathbf{v}||_1$

- Sequentially solve for $(d_i, \mathbf{u}_i, \mathbf{v}_i)$, $i \in \{1 \cdots m\}$:
  e.g. $Y_2 = Y - d_1 \mathbf{u}_1 \mathbf{v}_1$

- Drawbacks:
  - Orthogonality of the singular vectors is not guaranteed.
  - Computational cost.

Thresholding: Optimization-Free

Yang et al. (2013): **A sparse SVD method for high dimensional data**.

- Simultaneously computes the subspaces spanned by the leading singular vectors in U, V using the orthogonal subspace iterations.

Thresholding: Optimization-Free

Yang et al. (2013): **A sparse SVD method for high dimensional data**.

- Simultaneously computes the subspaces spanned by the leading singular vectors in U, V using the orthogonal subspace iterations.
- Thresholding is used to replace by zero the smaller entries of U and V.
- The Fast Iterative Thresholding Sparse SVD (FIT-SSVD)

The FIT-SSVD Algorithm

1. Multiplication and Thresholding: $U^{(k),thr} = \eta(YV^{(k-1)}, \gamma_u)$,
2. QR Decomposition: $U^{(k)}R_u^{(k)} = U^{(k),thr}$,
3. Multiplication and Thresholding: $V^{(k),thr} = \eta(Y'U^{(k)}, \gamma_v)$,
4. QR Decomposition: $V^{(k)}R_v^{(k)} = V^{(k),thr}$.

For **a given threshold level** γ,

- Hard-thresholding: $\eta(y, \gamma) = y \cdot \mathbf{1}_{(|y|>\gamma)}$.
- Soft-thresholding: $\eta(y, \gamma) = sign(y) \cdot (|y| - \gamma)_+$

SVD for Transposable Data

- Recall low-rank approx. in the Frobenius norm:

$$||Y - UDV'||_F^2 = tr\{(Y - UDV')'(Y - UDV')\}.$$

- Weighted F-norm or (Ω, Σ)-norm:

$$||Y - UDV'||_{\Omega,\Sigma}^2 = tr\{(Y - UDV')'\Omega(Y - UDV')\Sigma\},$$

- Motivation: Log-likelihood function of $Y = B + E$ is

$$l(Y|\Omega^{-1}, \Sigma^{-1}) \propto tr\{(Y - B)'\Omega(Y - B)\Sigma\}.$$

Escoufier (1977+) and Allen et al. (2013).

The Generalized Ortho. Subspace Iter.

- Compute

$$(\hat{U}, \hat{D}, \hat{V}) = \arg \min_{U,D,V} ||Y - UDV'||^2_{\Omega,\Sigma}$$

subject to $U'\Omega U = I$, $V'\Sigma V = I$.

The Generalized Ortho. Subspace Iter.

- Compute

$$(\hat{U}, \hat{D}, \hat{V}) = \arg \min_{U,D,V} ||Y - UDV'||^2_{\Omega,\Sigma}$$

subject to $U'\Omega U = I$, $V'\Sigma V = I$.

1. Multiplication: $Y_L^{(k)} = Y\Sigma V^{(k-1)}$,
2. Ω-QR Decomposition: $U^{(k)} R_u^{(k)} = Y_L^{(k)}$,
3. Multiplication: $Y_R^{(k)} = Y'\Omega U^{(k)}$,
4. Σ-QR Decomposition: $V^{(k)} R_v^{(k)} = Y_R^{(k)}$.

- Computes Thresholded Gen. PCA (TGPCA).
  For $\Omega = \Sigma = I$, it reduces to the standard SVD/PCA.

Advantages of the TGPCA

- Uses the correlations effectively.
- Finds generalized singular vectors and guarantees their orthogonality.
- Inherits the good computational and statistical properties of the FIT-SSVD in Yang et al. (2013)

A Macroeconomics Dataset

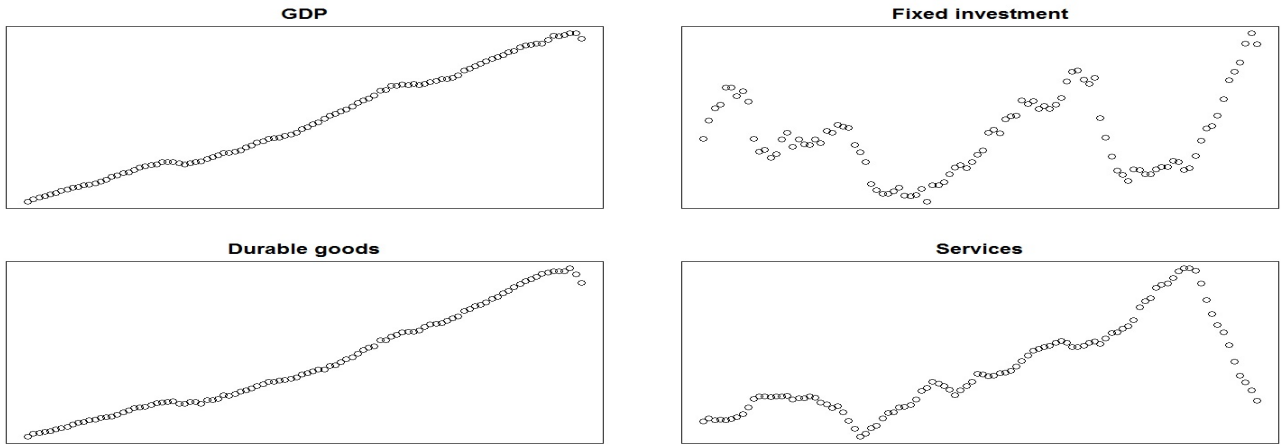
- Stock and Watson (2002/2012): 144 U.S. macroeconomic TS with $n = 195$ quarterly obs. from 1960:II through 2008:IV. Some series are aggregated.

- For example, the (aggregate) gross domestic product (GDP) is the sum of disaggregate series in goods, services, ...

A Macroeconomics Dataset

- Stock and Watson (2002/2012): 144 U.S. macroeconomic TS with $n = 195$ quarterly obs. from 1960:II through 2008:IV. Some series are aggregated.
- For example, the (aggregate) gross domestic product (GDP) is the sum of disaggregate series in goods, services, ...
- The $q = 35$ high-level aggregates series are used as the responses Y, and $p = 109$ lower-level disaggregated series as the predictors X.
- Each of the 144 series were transformed to (near) univariate stationarity.

A Macroeconomics Dataset



Problem III: Should one transform HD time series data to stationarity?

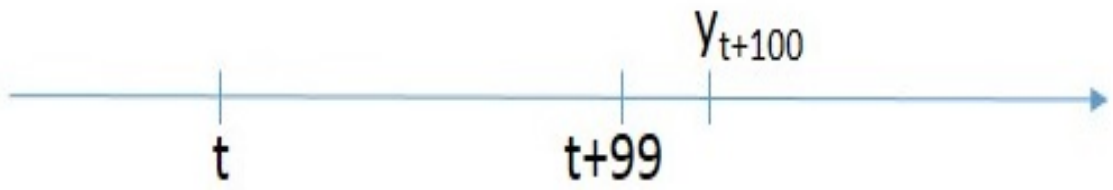
A Macroeconomics Dataset

- Do a classical PCA of X (not accounting for the correlation),
- Use the first 5 PCs of X as predictors (PCR-5),
- Find the PCR-5 forecasts for each of the 144 series and their RMSEs.
- PCR-5 beats most existing shrinkage/regularized methods such as the Bayesian model averaging, empirical Bayes, Bagging (Bootstrap aggregation),...
- Dobrev and Schaumburg (2003): Using regularized RRR, report slightly better performance than PCR-5.

Forecasting The Macro Data

Following Stock and Watson (2012):

- Out-of-sample one-step-ahead forecast with rolling window size 100 (quarterly observations).



$t = 1, ..., 95$.

- Forecast the 35 aggregated series in Y.

Forecasting The Macro Data

- Forecast equation:

$$\hat{\mathbf{y}}_{t+1} = \mathbf{x}_t \hat{U}\hat{D}\hat{V}',$$

where

- $\hat{U}\hat{D}\hat{V}' = \sum_{i=1}^{r} \hat{d}_i \hat{u}_i \hat{v}_i'$
is obtained by applying TGPCA to $\hat{B}_{OLS} = (X'X)^{-}X'Y$
with a predetermined number of layers r.

Forecasting The Macro Data

- Forecast equation:

$$\hat{\mathbf{y}}_{t+1} = \mathbf{x}_t \hat{U} \hat{D} \hat{V}',$$

  where

- $\hat{U}\hat{D}\hat{V}' = \sum_{i=1}^{r} \hat{d}_i \hat{u}_i \hat{v}_i'$
  is obtained by applying TGPCA to $\hat{B}_{OLS} = (X'X)^{-}X'Y$
  with a predetermined number of layers r.

- Threshold level: $\gamma = \sqrt{2 \log p}$,
  where p is the length of the relevant singular vector.

- Ω and Σ **are estimated by the sample covariances of** X **and** Y.

Forecast Performance

- The measure of forecast accuracy:
  - Root of mean square error $RMSE_j = \sqrt{\sum_t (y_{jt} - \hat{y}_{jt})^2/95}$.
  - Ratios of RMSE:

$$\text{Ratio}_j = \frac{RMSE_{j,TGPCA}}{RMSE_{j,PCR-5}}, \text{ for } j = 1, ..., 35.$$

  - If Ratio$<$ 1, the TGPCA-r is better than the PCR-5.

Forecast Performance

TGPCA-r	5%	25%	50%	75%	95%
2	0.53	0.94	**1.00**	1.09	1.28
3	0.51	0.95	**1.03**	1.13	1.29
4	0.44	0.97	**1.05**	1.23	1.38
5	0.44	1.00	**1.08**	1.19	1.39
6	0.44	0.97	**1.04**	1.10	1.15
7	0.45	0.98	**1.05**	1.08	1.18
8	0.47	0.91	**1.03**	1.16	1.49
9	0.32	0.41	**1.01**	1.09	1.25

Percentiles of ratios of RMSE of TGPCA relative to the PCR-5 for **transformed data**.

Forecast Performance

TGPCA-r	5%	25%	50%	75%	95%
2	0.84	1.10	1.27	1.76	2.71
3	0.44	0.70	**0.96**	1.28	1.76
4	0.49	0.62	**0.85**	1.07	1.50
5	0.51	0.64	**0.78**	**0.95**	1.38
6	0.56	0.68	**0.77**	**0.98**	1.16
7	0.50	0.56	**0.62**	**0.78**	**0.92**
8	0.38	0.45	**0.53**	**0.73**	**0.82**
9	0.42	0.51	**0.64**	**0.84**	**0.94**

Percentiles of ratios of RMSE of TGPCA relative to the PCR-5 for **original data**.

Problem III: Transform the Data?

- Compared to Stock and Watson (2012), the TGPCA approach obviates the need to transform the data to stationarity which can be a major advantage over the PCR in high-dimensional data situations.
- Deciding what transformations to use is a difficult task even for univariate time series data.

Simulating nonstationary data

- **Case I**: Random walk.

$$X_{j,t} = X_{j,t-1} + \epsilon_{jt}.$$

- **Case II**: AR(2) with unit root plus drift.

$$X_{j,t} = 1.03X_{j,t-1} - 0.03X_{j,t-2} + c_j + \epsilon_{jt}.$$

- **Case III**: AR(3) with unit root plus seasonality.

$$\begin{aligned} X_{j,t} = \quad & 1.2X_{j,t-1} - 0.21X_{j,t-2} + 0.01X_{j,t-3} + c_j \\ & + sin(\pi * t/16) * 5 + \epsilon_{jt}. \end{aligned}$$

Simulation model

- $Y = XB + E$
- $B = \sum_{i=1}^{q} d_i \mathbf{u}_i \mathbf{v}_i'$ with the first five largest singular values $(177, 32, 30, 26, 22)$, while others are less than 5.
- This indicates that the model with $r = 5$ is appropriate.

Simulation (cont.)

TGPCA-r	5%	25%	50%	75%	95%
			Original Data		
5	0.10	0.98	1.14	1.33	1.66
6	0.07	0.88	1.01	1.15	1.39
7	0.05	0.80	**0.92**	1.03	1.22
8	0.04	0.72	**0.83**	**0.94**	1.10
			Transformed Data		
5	0.02	0.90	**0.96**	1.00	1.05
6	0.02	0.86	**0.93**	**0.98**	1.04
7	0.02	0.83	**0.90**	**0.95**	1.02
8	0.02	0.79	**0.87**	**0.93**	1.00

Case I: Percentiles of ratios of RMSE relative to the PCR-5 for the original and transformed simulated data.

Simulation (cont.)

TGPCA-r	5%	25%	50%	75%	95%
			Original Data		
5	0.03	0.54	**0.72**	**0.91**	1.28
6	0.01	0.47	**0.61**	**0.76**	**0.99**
7	0.01	0.42	**0.54**	**0.64**	**0.82**
8	0.01	0.38	**0.48**	**0.57**	**0.72**
			Transformed Data		
5	1.52	1.72	1.87	2.03	2.29
6	1.53	1.72	1.88	2.04	2.30
7	1.55	1.73	1.89	2.06	2.32
8	1.55	1.74	1.90	2.07	2.33

Case II: Percentiles of ratios of RMSE relative to the PCR-5 for the original and transformed simulated data.

Simulation (cont.)

TGPCA-r	5%	25%	50%	75%	95%
			Original Data		
5	0.05	0.79	1.05	1.37	1.96
6	0.03	0.69	**0.89**	1.10	1.51
7	0.02	0.60	**0.76**	**0.91**	1.19
8	0.01	0.54	**0.67**	**0.79**	1.02
			Transformed Data		
5	1.51	1.75	1.92	2.09	2.38
6	1.52	1.76	1.93	2.09	2.39
7	1.52	1.76	1.93	2.10	2.41
8	1.53	1.77	1.94	2.11	2.42

Case III: Percentiles of ratios of RMSE relative to the PCR-5 for the original and transformed simulated data.

Summary

The TGPCA

- outperforms PCR-5 in forecasting the original aggregate macroeconomic series when using the disaggregate series as the predictors.
- obviates the need to transform the data to stationarity. This needs further research.
- has good computational and statistical properties and guarantees the (generalized) orthogonality of the PCs.

The thresholded GPCA

- Finds sparse B in

$$\hat{B}_{OLS} = B + \tilde{E} = UDV' + \tilde{E}$$

by thresholding U and V or generalizing the FIT-SSVD algorithm to transposable data.

- The algorithm in Allen et al (2013) is sequential and does not guarantee the orthogonality of the singular vectors.

Thresholded GPCA

1. Multiplication and Thresholding: $Y_L^{(k),thr} = \eta(Y\Sigma V^{(k-1)}, \gamma_u)$,
2. Ω-QR Decomposition: $U^{(k)} R_u^{(k)} = Y_L^{(k),thr}$,
3. Multiplication and Thresholding: $Y_R^{(k),thr} = \eta(Y'\Omega U^{(k)}, \gamma_v)$,
4. Σ-QR Decomposition: $V^{(k)} R_v^{(k)} = Y_R^{(k),thr}$.

- Remark: As in Allen et al. (2013), it avoids computing square root and the inverse of Ω and Σ when de-correlating.

Selecting the Threshold Levels

- Updating $U^{(k)}$, one column at a time.
- $Y_L^{(k),thr} = \eta(Y\Sigma V^{(k-1)}, \gamma_u)$.

For a given column l, right multiply both side of $Y = B + E$ by $\Sigma \mathbf{v}_l^{(k-1)}$:

$$\begin{array}{ccccc} Y\Sigma v_l^{(k-1)} & = & u_l^{(k)} d_l^{(k)} & + & E\Sigma v_l^{(k-1)} \\ \uparrow & & \uparrow & & \uparrow \\ y & = & \mu & + & e \end{array}$$

Selecting the Threshold Level

Thresholding for the mean μ:

$$y = \mu + e.$$

- The ideal threshold level $\gamma = E[||e||_\infty]$ is unknown and hard to compute. The alternatives are:

Selecting the Threshold Level

Thresholding for the mean μ:

$$y = \mu + e.$$

- The ideal threshold level $\gamma = E[||e||_\infty]$ is unknown and hard to compute. The alternatives are:
  - Asymptotic result for the Gaussian sequence model $\gamma = \sigma\sqrt{2\log n}$, Johnstone (2011).
  - "m out of n" bootstrap, Bickel et al. (1997).
- Thresholded: $(Y_{L,l})^{(k),thr}$.

Selecting the Threshold Level

- Thresholded version $(Y_{L,l})^{(k),thr}$ serves as an estimator of the mean vector.
- For $l = 1, \cdots, r$, repeating the previous procedure leads to $(Y_L)^{(k),thr}$.
- Updating $V^{(k)}$ is similar.