High-Dimensional Time Series Analysis

Ruey S. Tsay
Booth School of Business
University of Chicago

December 2015

Outline

Analysis of high-dimensional time-series data (or dependent big data)

- Problem and some examples
- New challenges
- Parsimony vs sparsity
- Traditional methods may fail
- Some useful methods
- Concluding remark

Goal: Discuss directions for further research

High dimension: an example

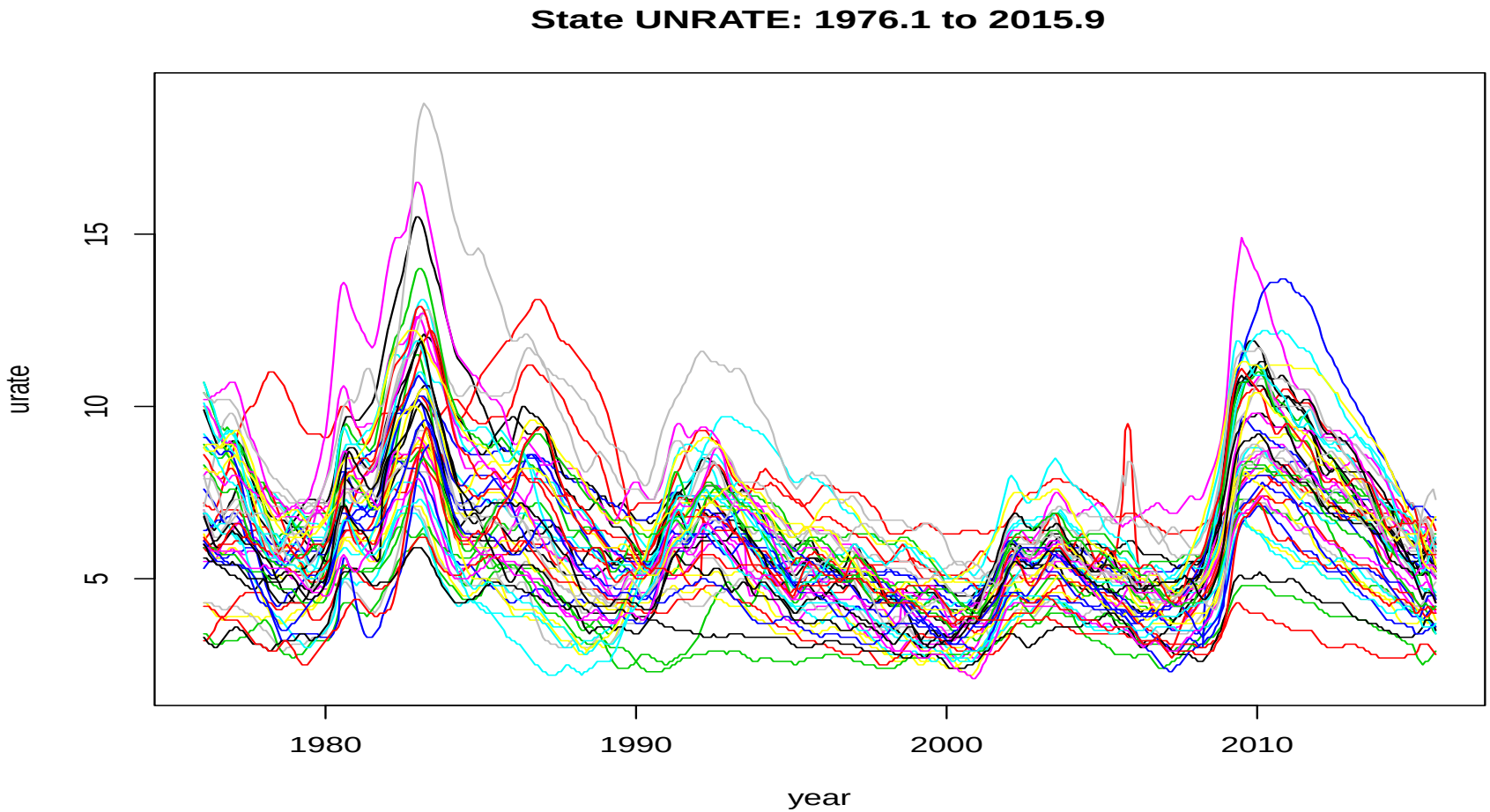

Figure: Time plots of monthly unemployment rates of the 50 States in the U.S. from January 1976 to September 2015. The data are seasonally adjusted.

Daily log returns of components of S&P 100 index

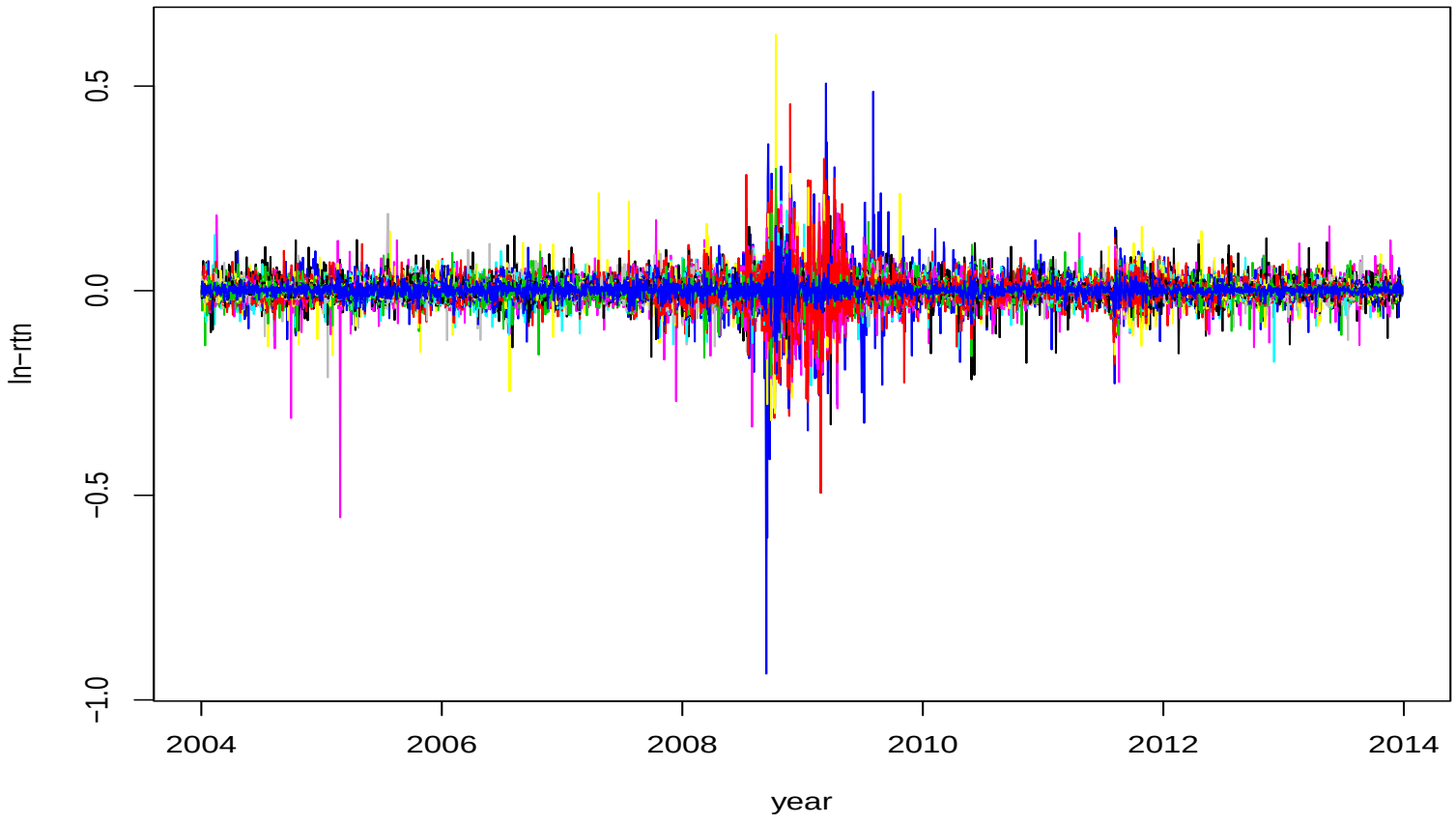

Figure: Time plots of 92 daily log return series: 2004-2013

3rd example: daily stock returns for two years

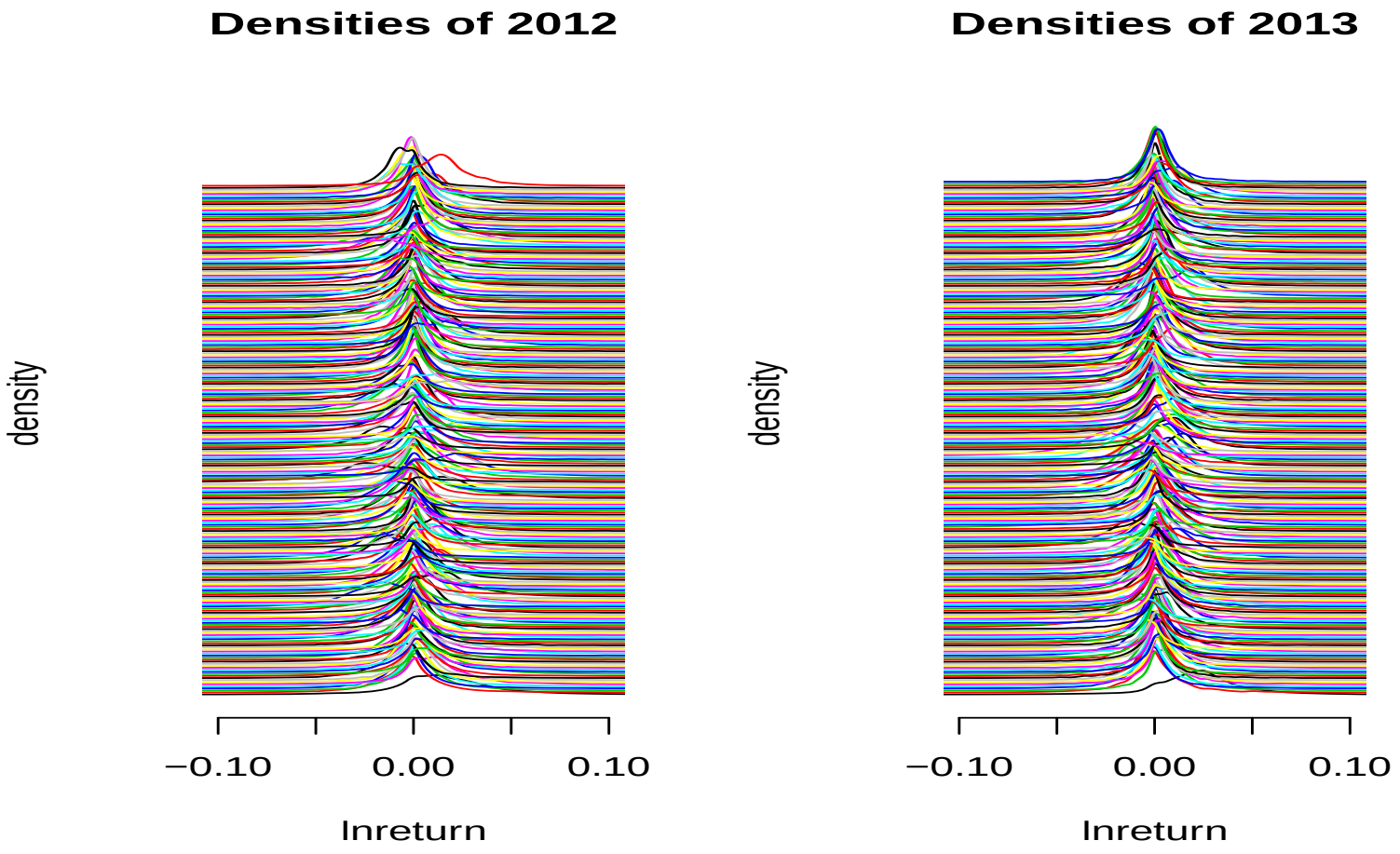

Figure: Densities of daily log returns of U.S. stocks in 2012 and 2013.

A fourth example: demands of electricity

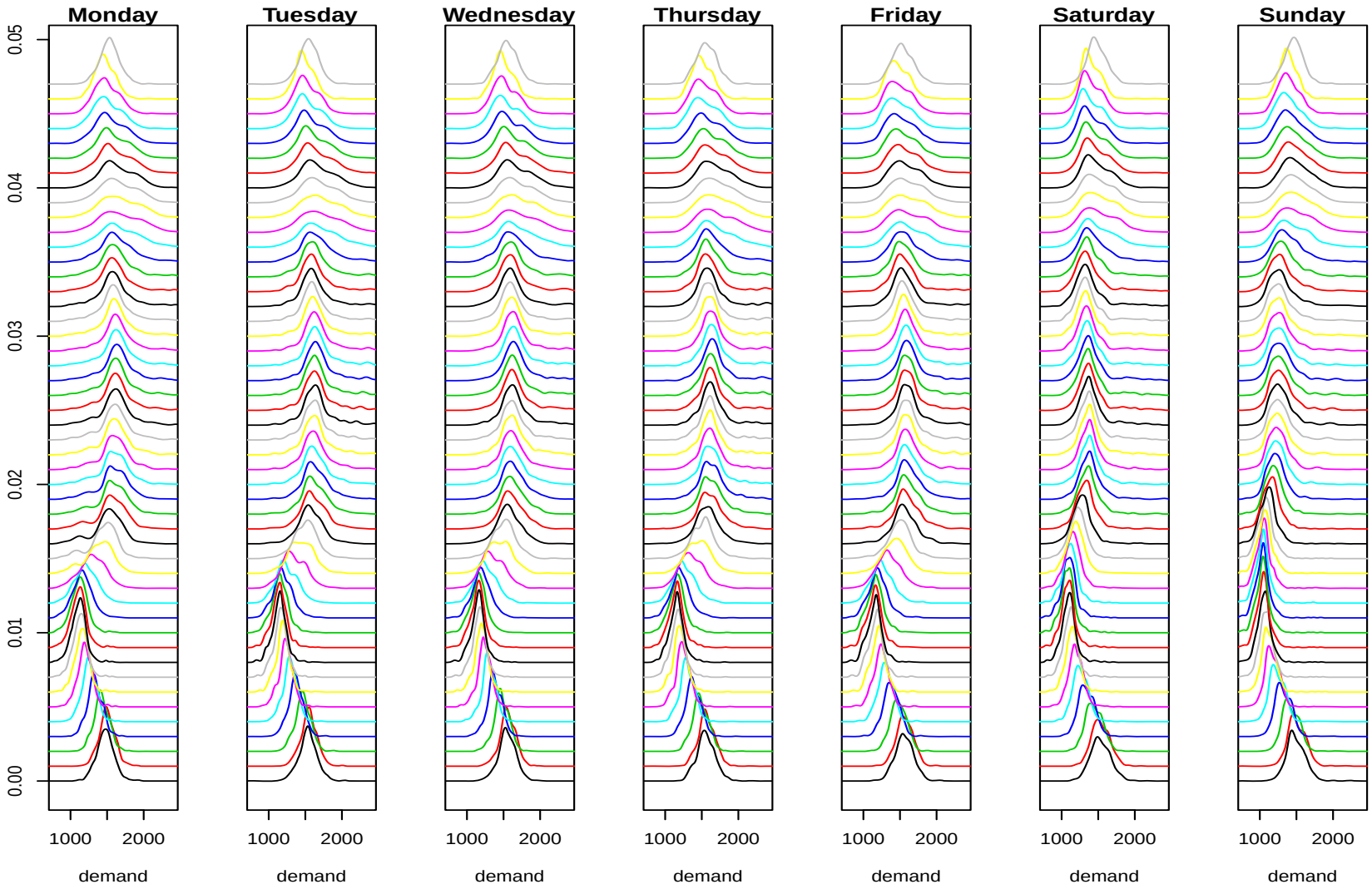

Figure: Empirical densities of electricity demand, 30 minute intervals,

Why are these series of interest?

Many reasons, including

1. Predicting state unemployment rates is important to local governments
2. Asset allocation and risk management require multivariate volatility
3. Finding relationships, both spatial and temporal, is of interest
4. Searching for common features among the data, and many more

Global economies and business are more integrated, more complicated, more competitive than before

What are available?

Statistical methods in the literature:

1. Focus on sparsity
2. Various penalized auto-regressions, e.g. Lasso and its extensions
3. Various dimension reduction methods, e.g. factor models, index models, clustering.

Some useful concepts in analyzing big data:

1. Parsimony vs sparsity: Sparsity $\Rightarrow$ Parsimony
2. Simplicity vs reality: trade-off btw feasibility & sophistication

Parsimony, but not sparsity

A simple example

$$\boldsymbol{Z}_t = \beta \left(\frac{1}{km} \sum_{j=1}^{m} \sum_{i=1}^{k} z_{i,t-j} \right) + \boldsymbol{a}_t$$

A highly restricted VAR(m) model.
This model has only one coefficient so it is parsimonious, but it is not sparse because $\boldsymbol{Z}_t$ depends on all elements of past m lagged values.
In some applications, $\sum_{i=1}^{k} z_{i,t-j}$ is a close approximation to the first principal component, and long lag-average denotes momentum effect (or local trend).

Lasso may fail for dependent data

1. Data generating model: scalar Gaussian autoregressive, AR(3), model

$$x_t = 1.9x_{t-1} - 0.8x_{t-2} - 0.1x_{t-3} + a_t, \quad a_t \sim N(0,1).$$

   Generate 2000 observations. See Figure 5.

2. Big data setup
   - Dependent x_t: $t = 11, \ldots, 2000$
   - Regressors: $X_t = [x_{t-1}, x_{t-2}, \ldots, x_{t-10}, \epsilon_{1t}, \ldots, \epsilon_{10,t}]$, where ϵ_{it} are iid $N(0,1)$.
   - Dimension $= 20$, sample size 1990.

3. Run the Lasso regression via the **lars** package of R. See Figure 2 for results. Lag 3, x_{t-3} was not selected.

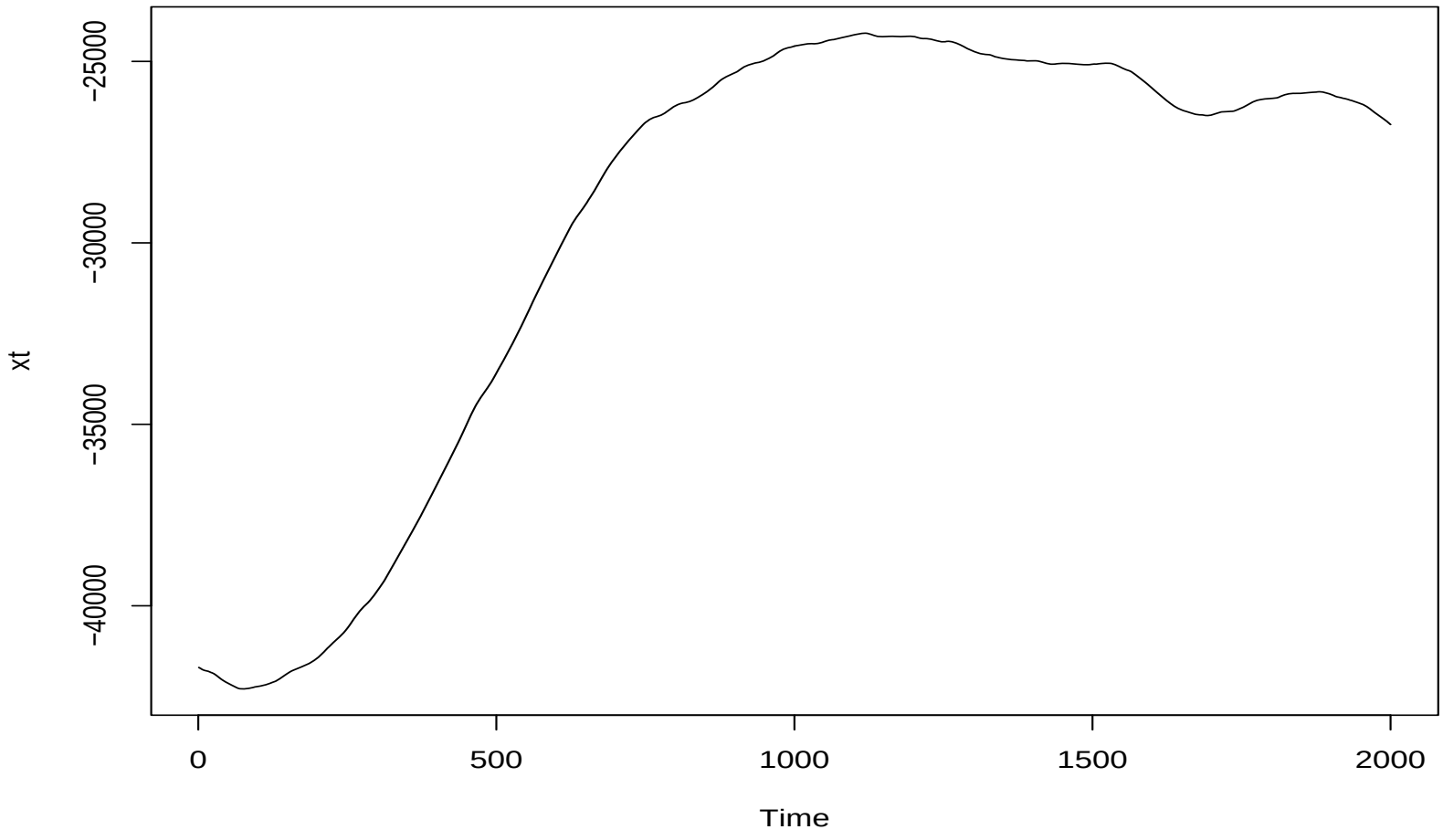

Figure: Time plot of simulated AR(3) time series with 2000 observations

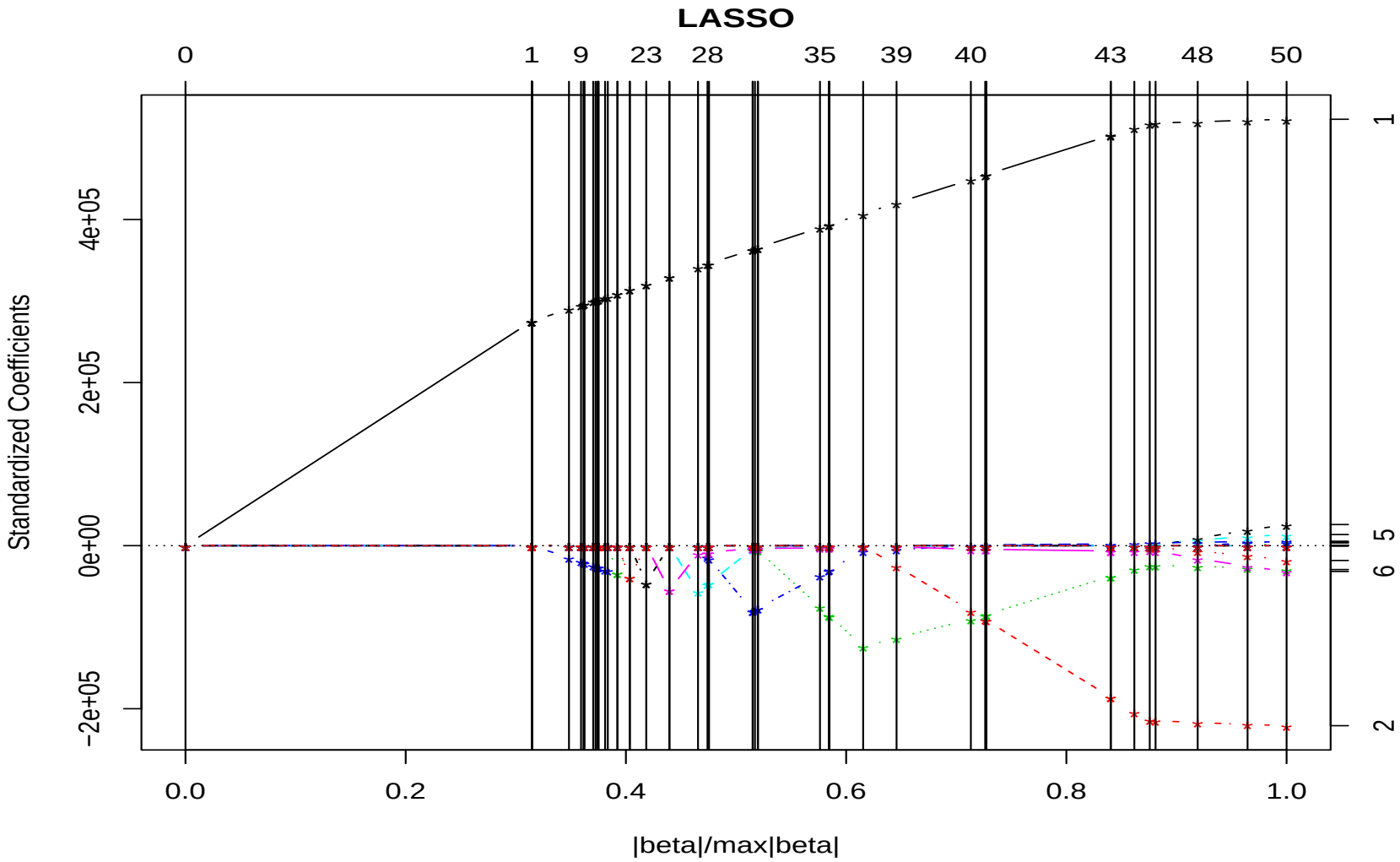

Figure: Results of Lasso regression for the AR(3) series

OLS works if we entertain AR models

Run the linear regression using the first three variables of X_t.

- Fitted model

$$x_t = 1.902x_{t-1} - 0.807x_{t-2} - 0.095x_{t-3} + \epsilon_t, \quad \sigma_\epsilon = 1.01.$$

- All estimates are statistically significant with p-value less than 2.22×10^{-5}.
- The residuals are well behaved, e.g. $Q(10) = 12.23$ with p-value 0.20 (after adjusting the df).

Why?

Two possibilities:

1. Scaling effect: Lasso standardizes each variable in X_t. For unit-root non-stationary time series, standardization might wash out the dependence of the stationary part
2. Multicollinearity: Unit-root time series have strong serial correlations. [ACF approach 1 for all lags.]

This artificial example highlights the difference between independent and dependent data.

Need to develop methods for high-dimensional time series!

Possible solutions

1. Re-parameterization using time series properties
2. Use different penalties for different parameters

The first approach is easier.
For the particular time series, we can define $\Delta x_t = (1 - B)x_t$ and $\Delta^2 x_t = (1 - B)^2 x_t$. Then,

$$\begin{aligned} x_t &= 1.9x_{t-1} - 0.8x_{t-2} - 0.1x_{t-3} + a_t \\ &= x_{t-1} + \Delta x_{t-1} - 0.1\Delta^2 x_{t-1} + a_t \\ &= \text{double} + \text{single} + \text{stationary} + a_t. \end{aligned}$$

The coefficients of $x_{t-1}, \Delta x_{t-1}, \Delta^2 x_{t-1}$ are 1, 1, an -0.1, respectively.

Different frameworks for LASSO

The X-matrix of conventional LASSO consists of

$$(x_{t-1}, x_{t-2}, \ldots, x_{t-10}, z_{1t}, \ldots, z_{10,t}),$$

where z_{it} are iid $N(0,1)$.
Under the re-parameterization, the X-matrix becomes

$$(x_{t-1}, \Delta x_{t-1}, \Delta^2 x_{t-1}, \ldots, \Delta^2 x_{t-8}, z_{1t}, \ldots, z_{10,t}).$$

These two X-matrices provide theoretically the same information. However, the first one has high multicollinearity, but the 2nd one does not, especially after standardization.

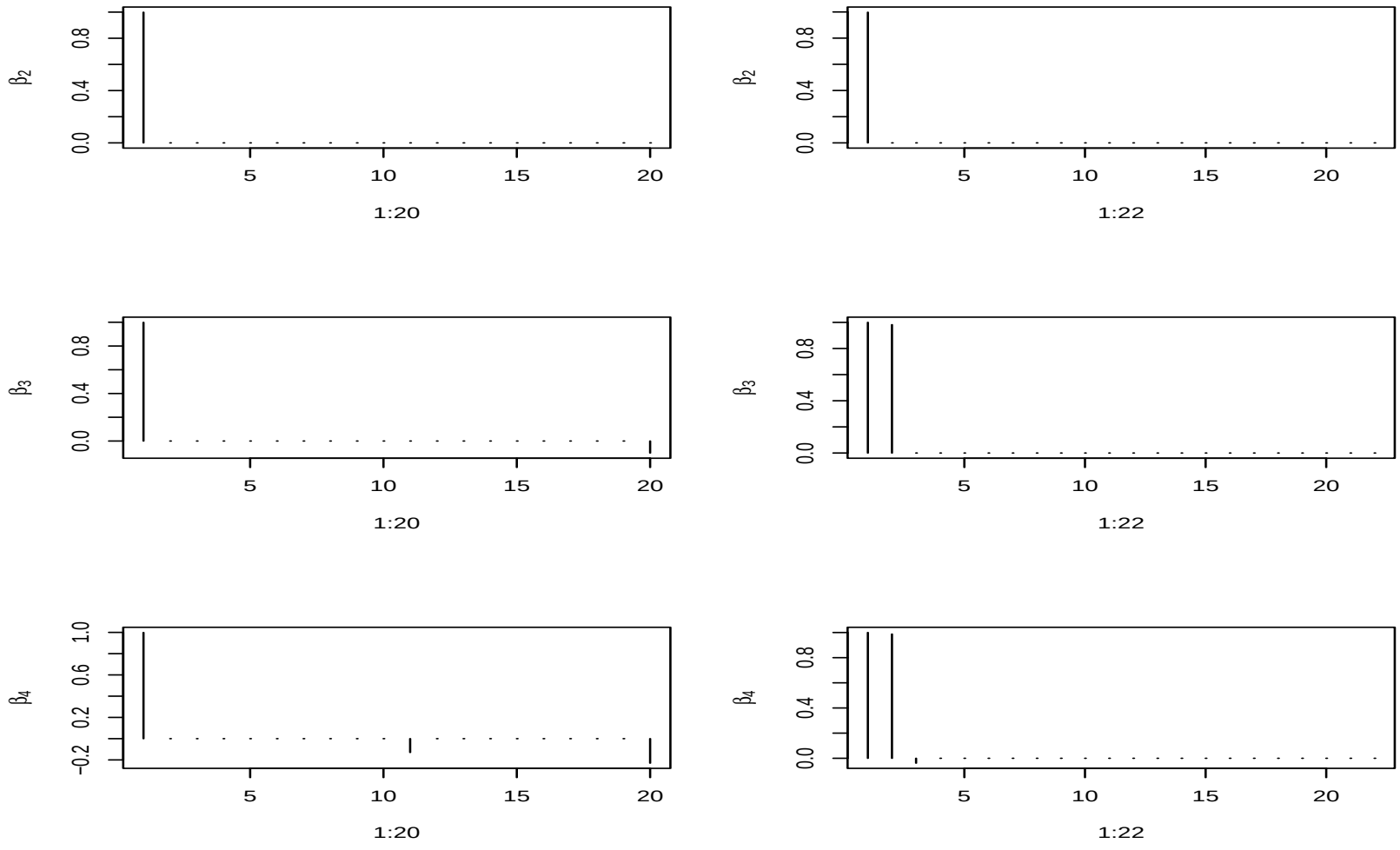

Figure: Comparison of β-estimates of **lars** results

Theoretical justification

Focus on the particular series x_t used. Some properties of the series are

1. $T^{-4}\sum_{t=1}^{T} x_t^2 \Rightarrow \int_0^1 \bar{W}^2$, where $\bar{W} = \int_0^1 W(s)ds$ with $W(s)$ the standard Brownian motion.
2. $T^{-5/2}\sum_{t=1}^{T} x_t \Rightarrow \int_0^1 \bar{W}$
3. $T^{-3}\sum_{t=1}^{T} x_t \Delta x_t \Rightarrow \int_0^1 \bar{W}W$
4. $T^{-2}\sum_{t=1}^{T} (\Delta x_t)^2 \Rightarrow \int_0^1 W^2$

Standardization may wash out the Δx_{t-1} and $\Delta^2 x_{t-1}$ parts.

Remark: This example suggests an approach for handling co-integration in high-dimensional time series analysis.

New Challenges: canonical correlation analysis

Illustration of canonical correlation analysis. See Bao, Hu, Pan and Zhou (2014).
Framework: $\dim(\boldsymbol{X}) = p$, $\dim(\boldsymbol{Y}) = q$. $E(\boldsymbol{X}) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p$. $E(\boldsymbol{Y}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{I}_q$. Assume $p < q$.
Let $r_i = \rho_i^2$, where ρ_i is the ith population canonical correlation coefficient (in decreasing order). Let λ_i be the ith eigenvalue of the usual sample matrix

$$\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1}\boldsymbol{S}_{yx}$$

Let $F_T(x) = \frac{1}{p}\sum_{i=1}^{p} 1_{\{\lambda_i \le x\}}$ be the empirical spectral density.

Assumptions: high-dimensional case

1. $p/T \to c_1$, $q/T \to c_2$, $p/q \nrightarrow 1$ as $T \to \infty$ and $c_1 + c_2 \in (0, 1)$.
2. $\text{Rank}(\mathbf{\Sigma}_{xy}) = k$. Further, let k_0 be the nonnegative integer satisfying

$$r_0 = 1 \geq \cdots \geq r_{k_0} > r_c \geq r_{k_0+1} \geq \cdots \geq r_k > r_{k+1} = 0,$$

where

$$r_c = \frac{c_1 c_2 + \sqrt{c_1 c_2 (1 - c_1)(1 - c_2)}}{(1 - c_1)(1 - c_2) + \sqrt{c_1 c_2 (1 - c_1)(1 - c_2)}}.$$

The Null Case with $k = 0$

$\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent and the Assumption 1 holds.

Result 1: Almost surely $F_T(x)$ converges weakly to $F(x)$ with density

$$\rho(x) = \frac{1}{2\pi c_1} \frac{\sqrt{(d_r - x)(x - d_\ell)}}{x(1-x)} 1_{\{d_\ell \le x \le d_r\}}, \quad \text{with}$$

$$\begin{aligned} d_r &= c_1 + c_2 - 2c_1c_2 + 2\sqrt{c_1c_2(1-c_1)(1-c_2)} \\ d_\ell &= c_1 + c_2 - 2c_1c_2 - 2\sqrt{c_1c_2(1-c_1)(1-c_2)}. \end{aligned}$$

Result 2: The sample eigenvalues satisfy

$$\lambda_i \to d_r, \quad a.s.$$

for any fixed integer i. In fact, we have

$$\text{Prob}(\lambda_1 \le d_r + \eta) \ge 1 - T^{-h}$$

for any positive number h and any small positive constant η.

Important implication

To have zero canonical correlations under the null, we need c_1 and c_2 approach zero as $T \to \infty$. Dimensions cannot grow at the same rate as the sample size.

An example: $T = 3000$, $p = 30$, $q = 300$ so that $c_1 = 0.01$ and $c_2 = 0.1$. In this case, $r_c = 0.0335$, $d_r = 0.1677$ and $d_\ell = 0.0483$.

Expect all 30 sample eigenvalues to be between 0.0483 and 0.1677.

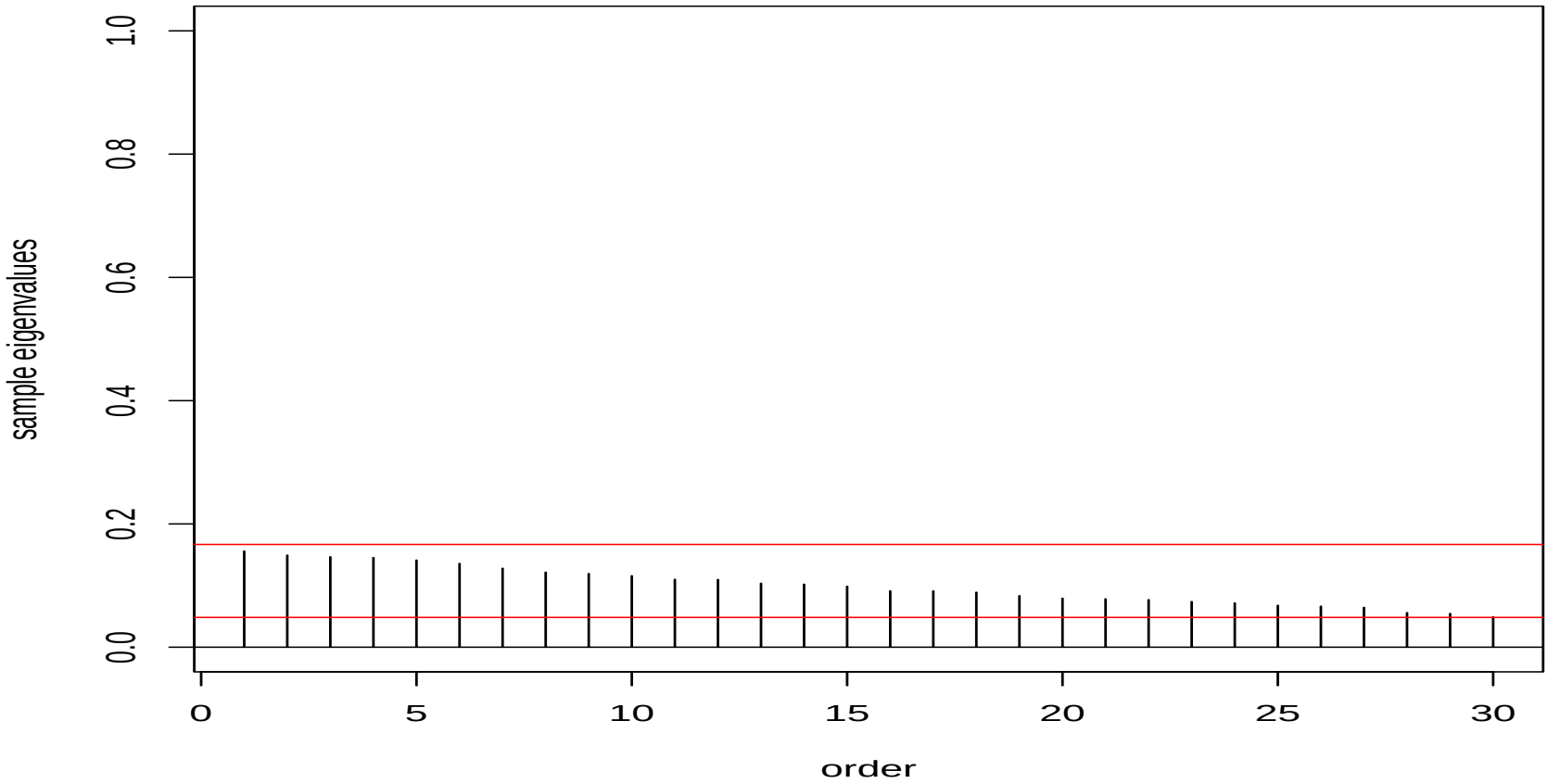

Figure: Sample squared canonical correlations under the Null of independent variables with dimensions 30 and 300. Sample size is 3000.

Dependent case, i.e. $k > 0$

Assumptions 1 and 2 hold.

Result 3: For $1 \leq i \leq k_0$, we have

$$\lambda_i \to_{a.s.} \gamma_i = r_i(1 - c_1 + \frac{c_1}{r_i})(1 - c_2 + \frac{c_2}{r_i}).$$

For each fixed $i \geq k_0 + 1$, we have

$$\lambda_i \to_{a.s.} d_r.$$

An example

$c_1 = 0.1,\ c_2 = 0.2,\ T = 5000,$
$(r_1, \ldots, r_5) = (0.8, 0.7, 0.6, 0.16, 0.15)$ so that $k = 5$.
In this case, $r_c = 0.167$ and $(d_\ell, d_r) = (0.02, 0.5)$.
Since $r_c = 0.167$, we can only expect the first three sample eigenvalues to approach $(\gamma_1, \gamma_2, \gamma_3) \approx (0.861, 0.793, 0.725)$ and others to approach 0.5.

Indeed, we have ρ_i as 0.858, 0.798, 0.724, 0.494, 0.492, 0.485, 0.482, ..., 0.0199.
The other eigenvalues are approximately between (0.02, 0.5).

Important implication: It is not possible to recover the non-zero eigenvalues 0.16 and 0.15.

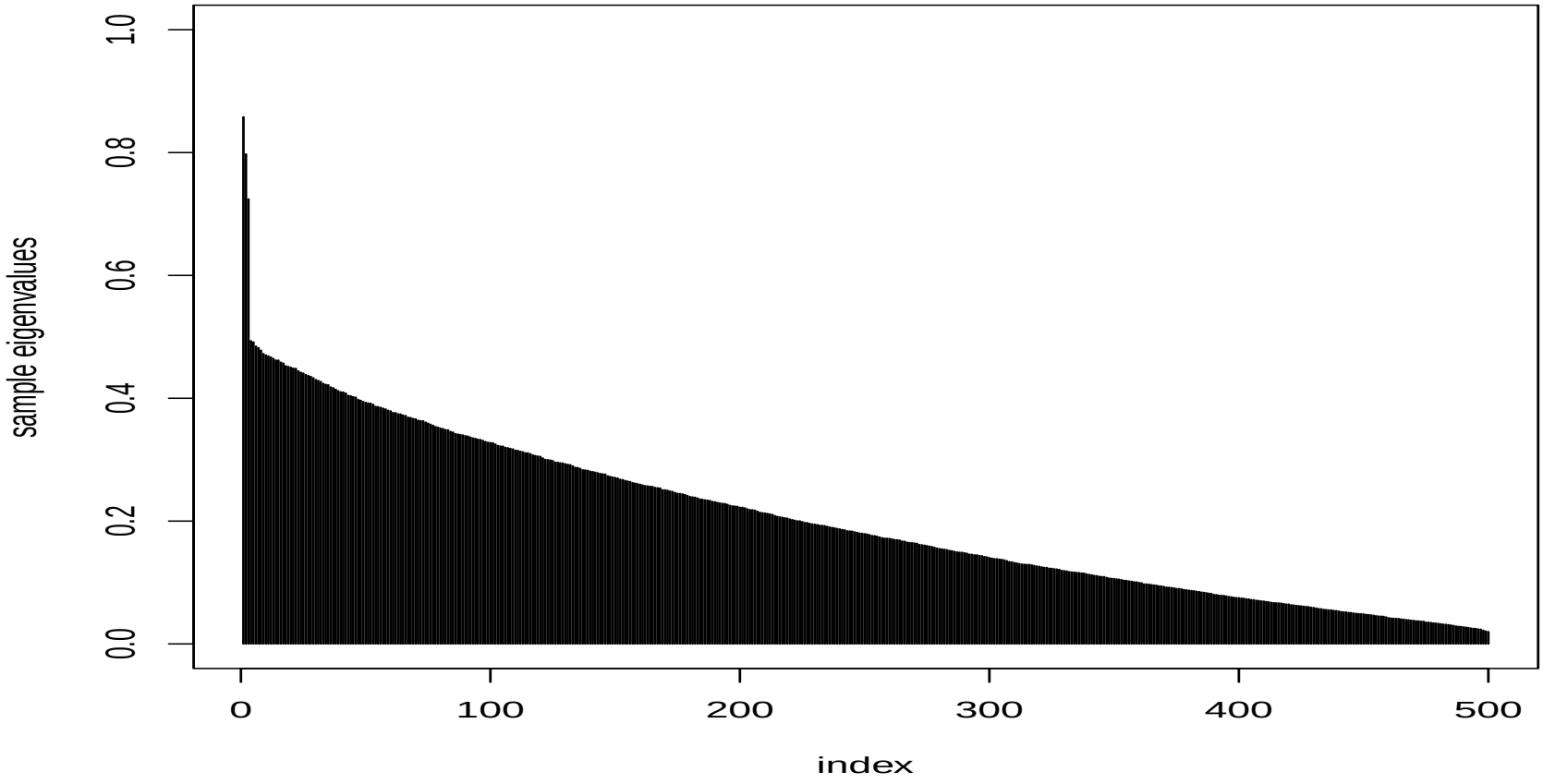

Figure: Sample squared canonical correlations for two dependent variables with dimension 500 and 1000. Sample size is 5000. The true eigenvalues are 0.8, 0.7, 0.6, 0.16, 0.15, and zero.

Some statistical methods

1. Classification and cluster analysis
   - K means
   - Tree-based classification
   - Model-based classification
2. Factor models, including volatility modeling
   - Prediction with many predictors
   - Prediction using mixed-frequency data
   - Approximate factor models
3. Generalizations of Lasso methods

Classification

A possible approach: Use a two-step procedure

1. Transform dependent big data into functions, e.g. probability densities
2. Apply classification methods to functional data

The density functions of daily log returns of U.S. stocks serve as an example.
We can then classify the density functions to make statistical inference

Illustration of classification

Cluster Analysis of density functions

Consider the time series of density functions $\{f_t(x)\}$.
For simplicity, assume the densities are evaluated at equally-spaced grid point $\{x_1 < x_2 < \ldots < x_N\} \in D$ with increment Δx. The data we have become $\{f_t(x_i)|t = 1, \ldots, T; i = 1, \ldots, N\}$.

Using Hellinger distance (HD), we consider two methods:

- K means
- Tree-based classification

Hellinger distance of two density functions

Let $f(x)$ and $g(x)$ be two density functions on the common domain $D \subset R$. Assume both density functions are absolutely continuous w.r.t. the Lebesgue measure. The Hellinger distance (HD) between $f(x)$ and $g(x)$ is defined as

$$H(f,g)^2 = \frac{1}{2}\int_D \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx = 1 - \int_D \sqrt{f(x)g(x)}dx$$

Basic properties:

1. $H(f,g) \geq 0$
2. $H(f,g) = 0$ if and only if $f(x) = g(x)$ almost surely.

K-means method

For a given K, the K-means method seeks partitions of the densities, say, $C_1, \ldots, C_K$, such that

1. $\bigcup_{k=1}^{K} C_k = \{f_t(x)\}$
2. $C_i \bigcap C_j = \emptyset$ for $i \neq j$
3. Sum of within-cluster variation $V = \sum_{k=1}^{K} V(C_k)$ is minimized, where the within-cluster variation is

$$V(C_k) = \sum_{t_1, t_2 \in C_k} H(f_{t_1}, f_{t_2})^2$$

It turns out this can easily be done by applying the K-means method with squared Euclidean distance to the squared-root densities $\{\sqrt{f_t(x)}\}$.

Example of K-means

Consider the 48 density functions of half-hour demand of electricity on Monday in Adelaide, Australia.
With $K = 4$ clusters, we have

k	Elements (time index)	Calendar Hours
1	17 to 44	8:00 AM to 10:00 PM
2	15, 16, 45 to 48, 1, 2, 3	7:00 – 8:00 AM; 10:00 PM – 1:30 AM
3	4, 5, 13, 14	1:30 – 2:30 AM; 6:00 – 7:00 AM
4	6 to 12	2:30 – 6:00 AM

Result: capture daily activities, namely, (1) active period, (2) transition period, (3) light sleeping period, and (4) sound sleeping period.

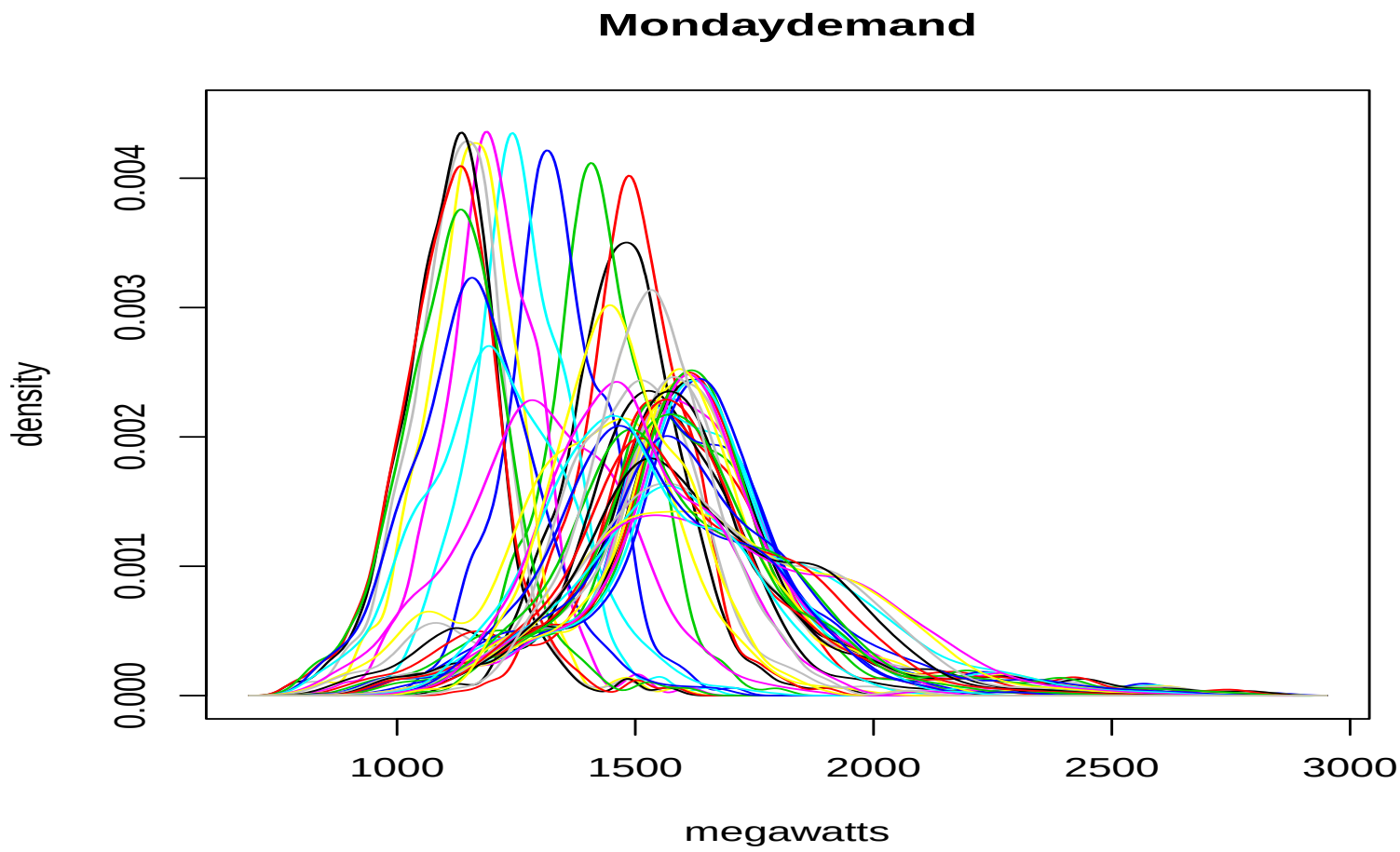

Figure: Density functions of half-hour electricity demand on Monday at Adelaide, Australia. The sample period is from July 6, 1997 to March 31, 2007.

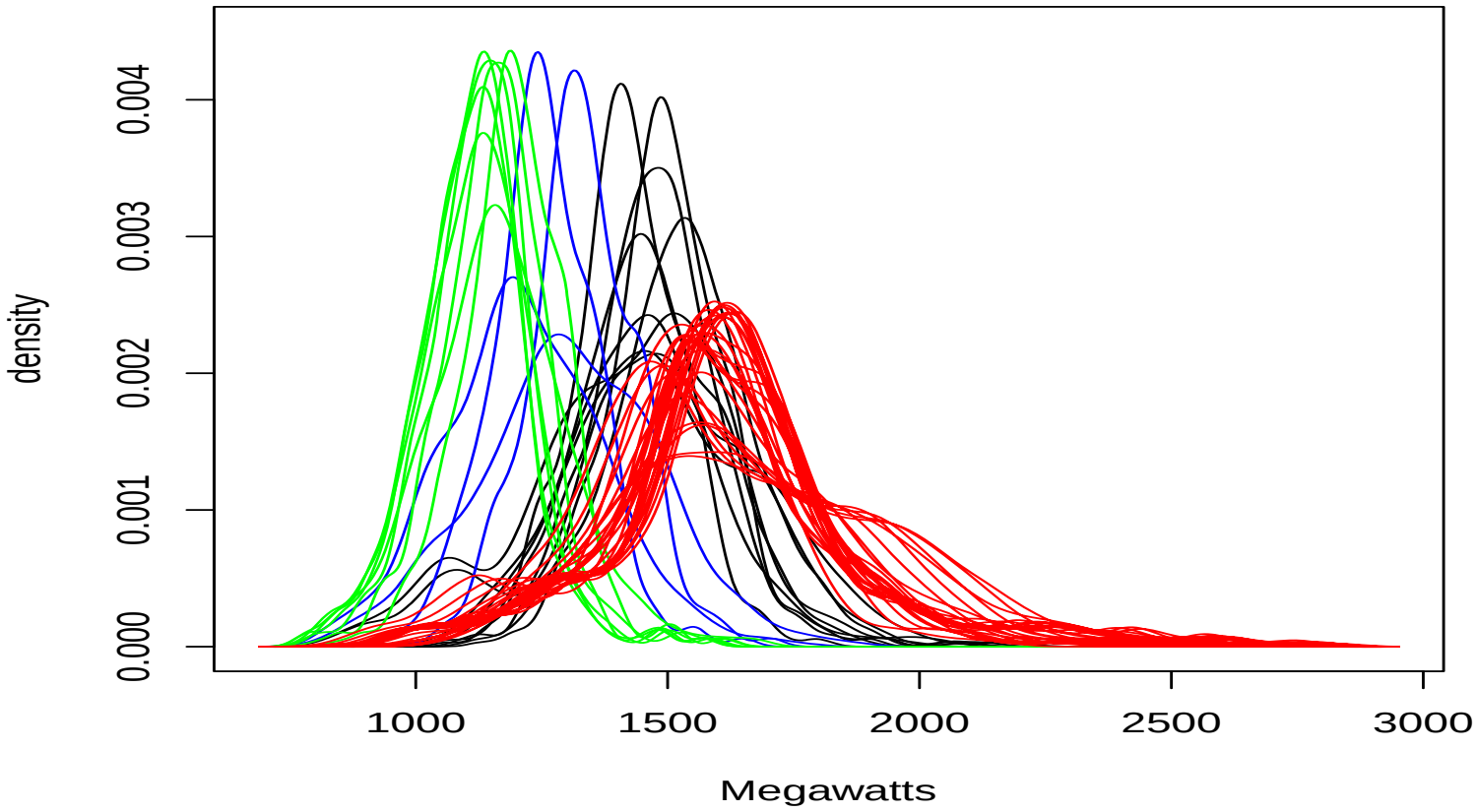

Figure: Results of K-means Cluster Analysis Based on Squared Hellinger Distance for Electricity Demands on Monday. Different colors denote different clusters.

Tree-based classification

Let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{pt})'$ denote p covariates. We use an iterative procedure to build a binary tree, starting with the root $C_0 = \{f_t(x)\}$.

1. For each covariate z_{it}, let $z_i(j)$ be the jth order statistic
   1.1 Divide C_0 into two sub-clusters

$$C_{i,j,1} = \{f_t(x) | z_{it} \leq z_{i(j)}\}; \quad C_{i,j,2} = \{f_t(x) | z_{it} > z_{i(j)}\}$$

   1.2 Compute the sum of within-cluster variations

$$H(i,j) = V(C_{i,j,1}) + V(C_{i,j,2})$$

   1.3 Find the smallest j, say v_i, such that $H(i, v_i) = \min_j\{H(i,j)\}$.
2. Select $i \in \{1, \ldots, p\}$, say I, such that $H(I, v_I) = \min_i\{H(i, v_i)\}$.
3. Use covariate z_{It} with threshold v_I to grow two new leaves, i.e.

$$C_{1,1} = C_{I,v_I,1}, \quad C_{1,2} = C_{I,v_I,2}$$

Tree-based procedure continued

Next, consider $C_{1,1}$ and $C_{1,2}$ as the root of a branch and apply the same procedure with their associated covariates to find candidate for growth.
The only modification is as follows: When considering $C_{1,1}$, we treat $C_{1,2}$ as a leaf in computing the sum of within-cluster variations. Similarly, when considering $C_{1,2}$ for further division, we treat $C_{1,1}$ as a leaf in computing the sum of within-cluster variations.
This growth-procedure is iterated until the number of clusters K is reached.

Example of tree-based classification

Consider the density functions of U.S. daily log stock returns in 2012 and 2013.
Using the first-differenced VIX index as the explanatory variable and $K = 4$, we obtain 4 clusters as follows:

$$(-\infty, -0.73], \quad (-0.73, 0.39], \quad (0.39, 1, 19], \quad (1.19, \infty).$$

The cluster sizes are 104, 259, 86, and 53, respectively.
Note that positive z_t signifies an increase in market volatility (uncertainty).

What drove the U.S. financial market?

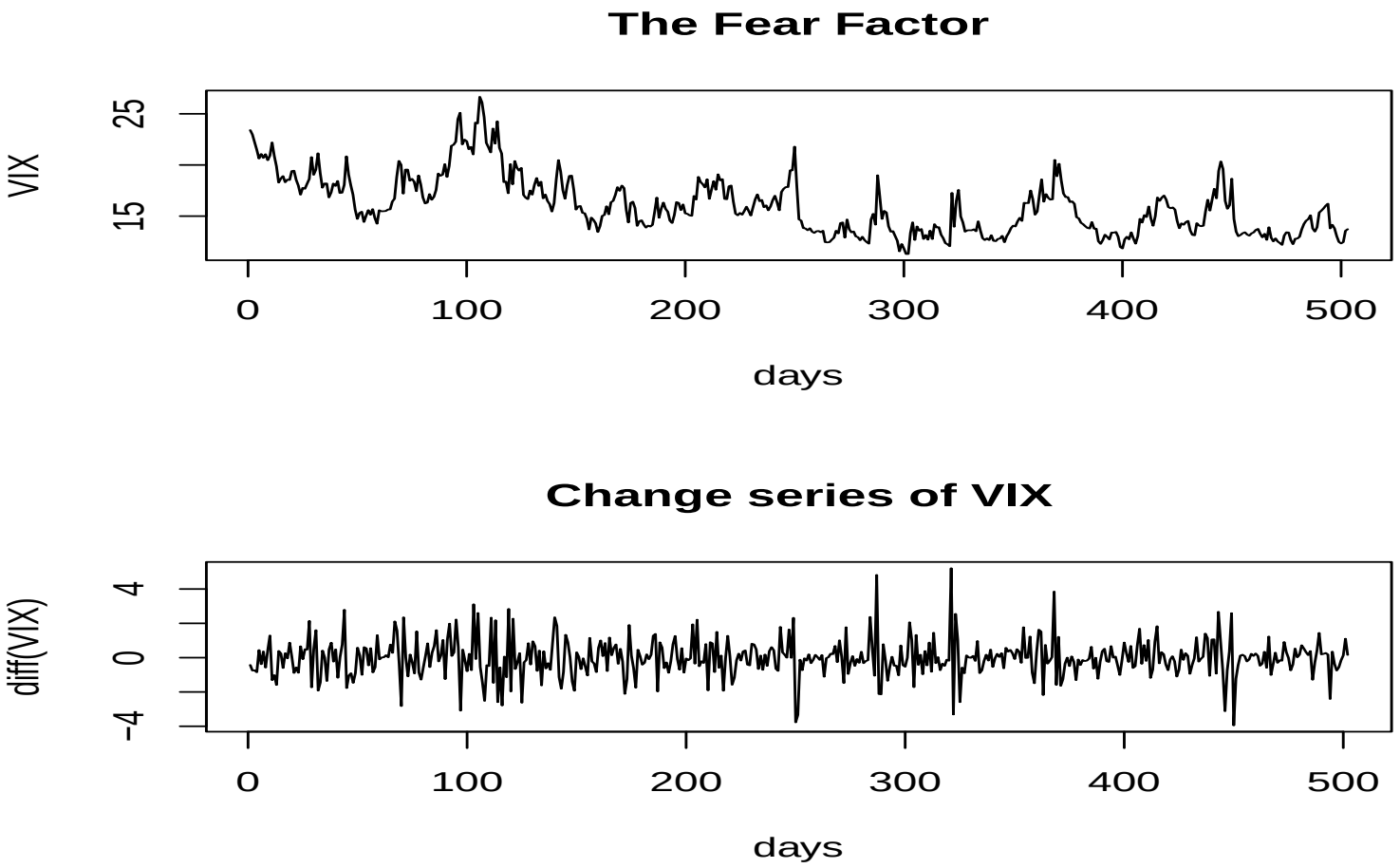

Figure: Time plots of the market fear factor (VIX index) and its change series: 2012-2013

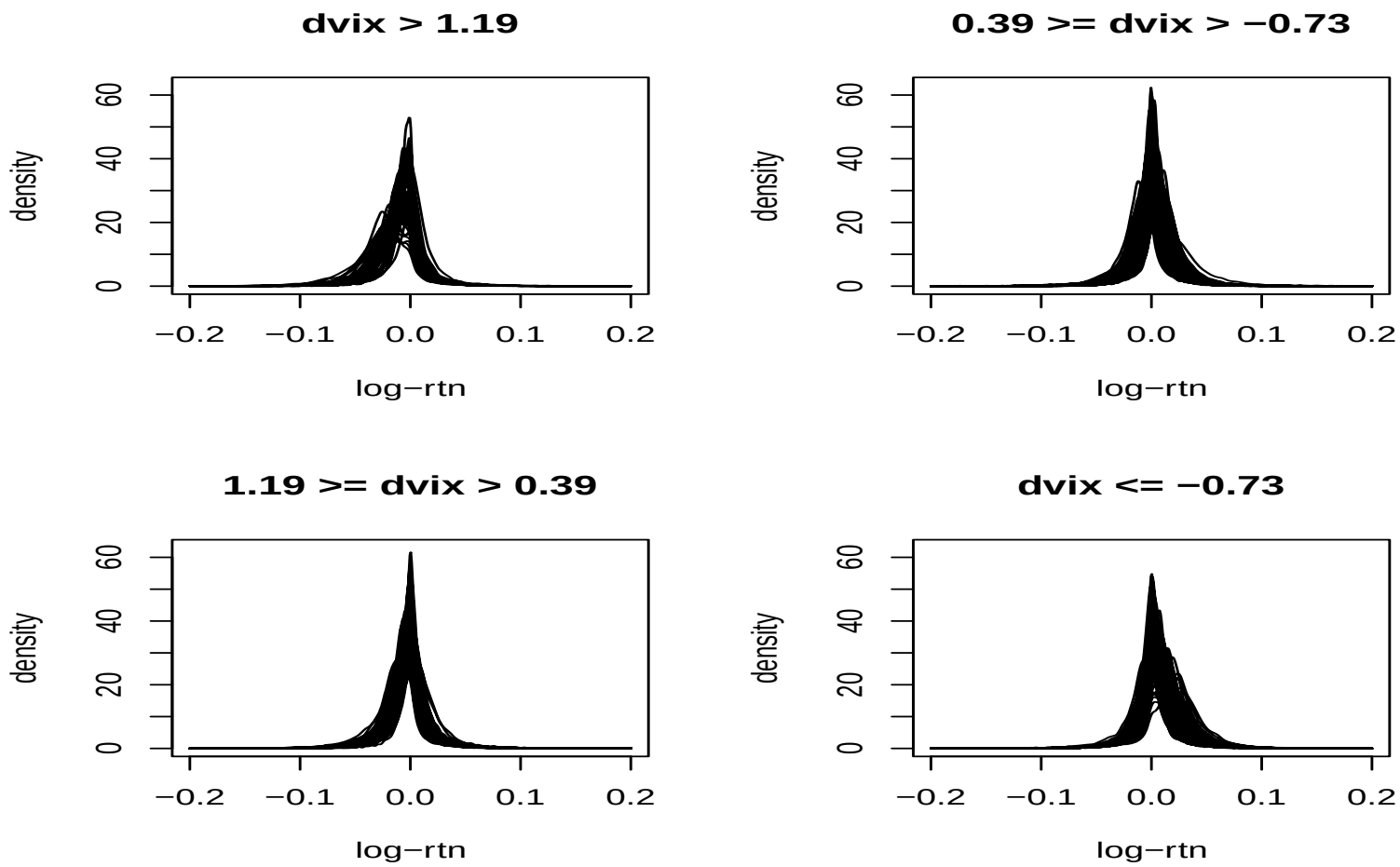

Figure: Results of Tree-based Cluster Analysis for the Daily Densities of Log Returns of the U.S. Stocks in 2012 and 2013. The first-differenced series of the VIX index is used as the explanatory variable. The numbers of element for the clusters are 53, 86, 259, and 104, respectively. The cluster classification is given in the heading of each plot.

Model-based classification

Work directly on observed multiple time series

1. Postulate a general univariate model for all time series, e.g. an AR(p) model
2. Time series in a cluster follow the same model: Pooling data to estimate common parameters
3. Time series in different clusters follow different models
4. May be estimated by Markov chain Monte Carlo methods
5. May employ scaled-mixture of normal innovations to handle outliers

Have been widely studied, e.g. Wang et al (2013) and Fruehwirth-Schnatter (2011), among others.

Application

1. Apply to monthly unemployment rates of 50 states of the U.S.
2. Use out-of-sample predictions to compare with other methods, including lasso.
3. For 1-step to 5-step ahead predictions, the model-based method works well in comparison. Wang et al (2013, JoF).

	RMSE$\times 10^4$				MAE$\times 10^4$			
Method	$m=1$	$m=2$	$m=3$	$m=4$	$m=1$	$m=2$	$m=3$	$m=$
UAR	1616	1492	1791	2073	879	994	1268	13
VAR	2676	2095	2129	2759	1349	1353	1506	16
Lasso25	1798	1833	2063	2504	1245	1250	1332	14
Lasso15	1714	1798	1855	2028	1186	1228	1296	13
G-Lasso	1877	1865	1882	1905	1291	1290	1306	13
LVAR	1550	1716	1806	1904	1065	1298	1210	13
Pls10	1239	1531	1679	1873	909	1028	1263	12
Pls30	1395	1651	1835	1890	933	1092	1281	13
Pls50	1685	1871	2006	1967	940	1158	1304	13
Pls70	1914	2040	2182	1953	996	1222	1362	14
Pls100	2187	2279	2313	2123	1099	1342	1480	15
Pcr10	1276	1829	2077	2108	890	1073	1247	14
Pcr30	1577	1837	2049	1769	888	1093	1261	13
Pcr50	1546	1805	2017	1759	880	1035	1209	12
Pcr70	1594	1837	2049	1769	886	1042	1221	12
Pcr100	1649	2117	2202	2163	1068	1243	1324	14
MBC	1607	1703	1809	1961	885	1035	1225	13
rMBC	1225	1481	1691	1839	873	1027	1193	12

Factor models and beyond

1. Generalizations of PC regression, PLS, etc.
   Diffusion index of Stock and Watson (2002, JASA)
2. Functional PCA, Ramsay and Silverman (2005) and Yao, M$'$uller and Wang (2005).
3. Factor models for multivariate volatility

Functional PCA: Singular value decomposition

1. A tool to study the time evolution of the return distributions
2. Data set: In this particular instance, each density function is evaluated at 512 points and we have

$$\boldsymbol{Y} = [Y_{it} = f_t(x_i) | i = 1, \ldots, N; t = 1, \ldots, T]_{512 \times 502}$$

3. Perform singular value decomposition

$$\tilde{\boldsymbol{Y}} = (N-1)\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$$

   where $\tilde{\boldsymbol{Y}}$ denotes column-mean adjusted data matrix, $\boldsymbol{U}$ is an $N \times N$ unitary matrix, $\boldsymbol{D}$ is an $N \times T$ rectangular diagonal matrix, and $\boldsymbol{V}$ is a $T \times T$ unitary matrix.
4. This is a simple form of functional PCA. [Large samples, smoothing of PC is not needed.]

Scree plot

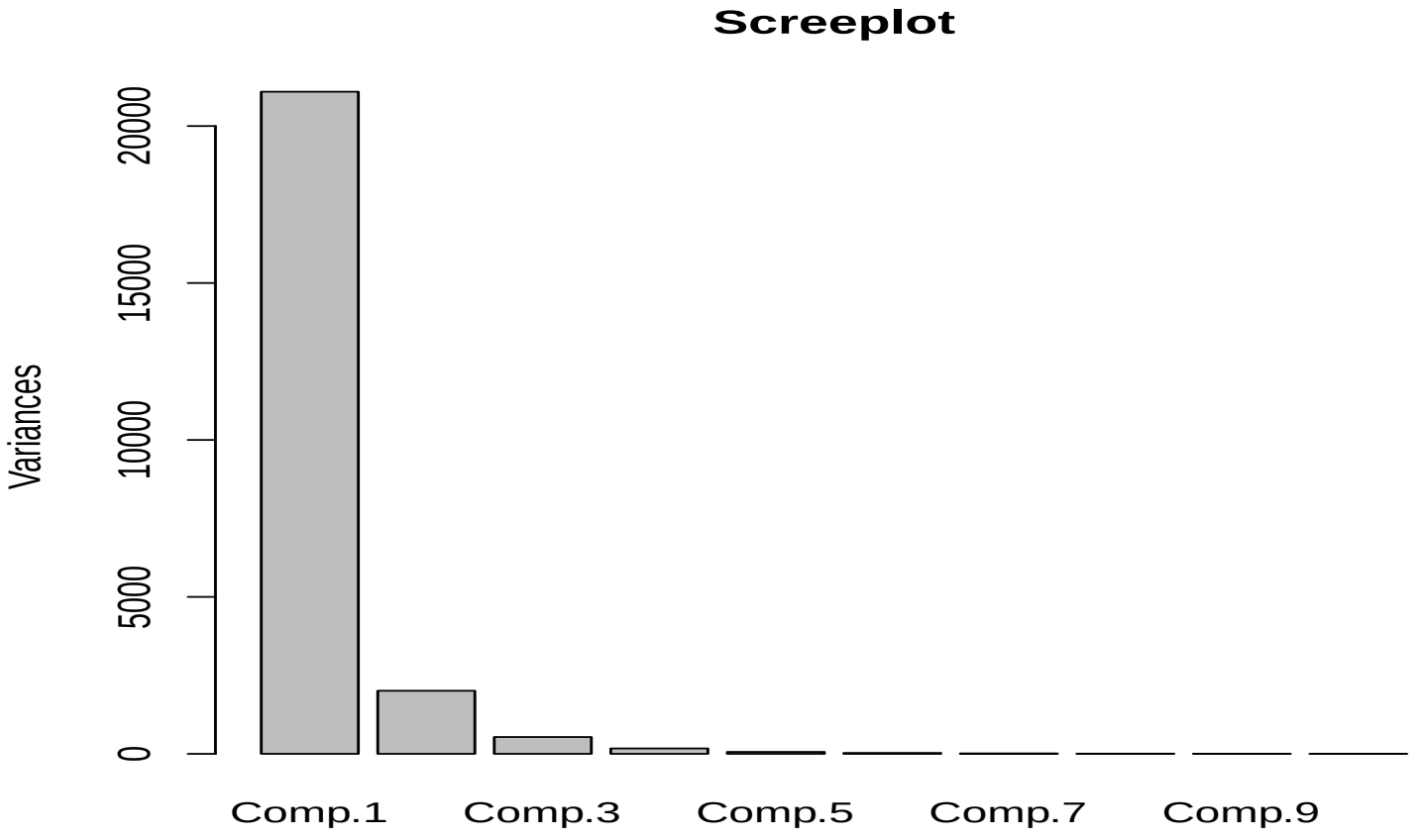

Figure: Scree plot of PCA for daily return densities in 2012 and 2013.

The first 6 PC functions

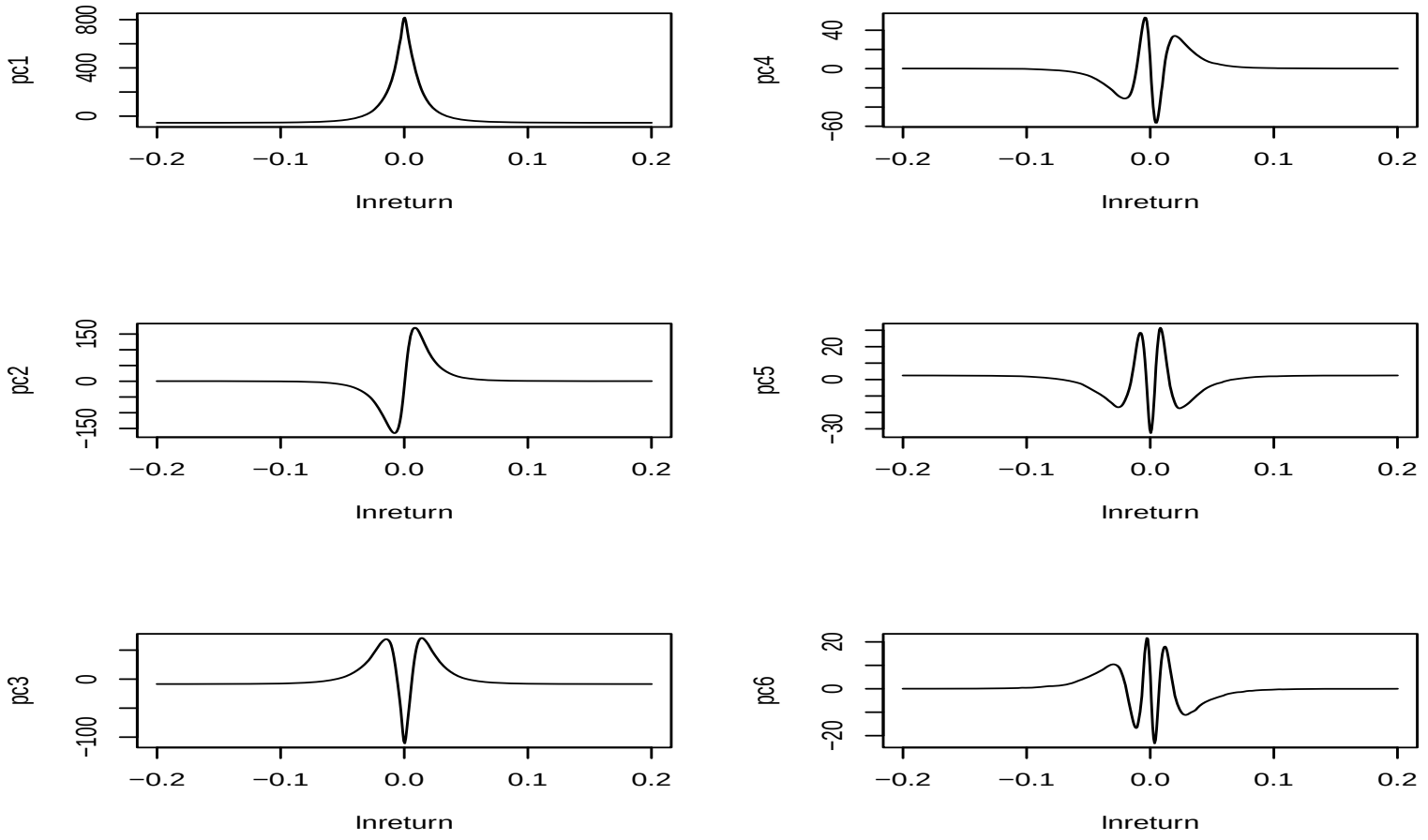

Figure: The first 6 PC functions for daily log return densities in 2012 and 2013.

Meaning of PC functions? 1st

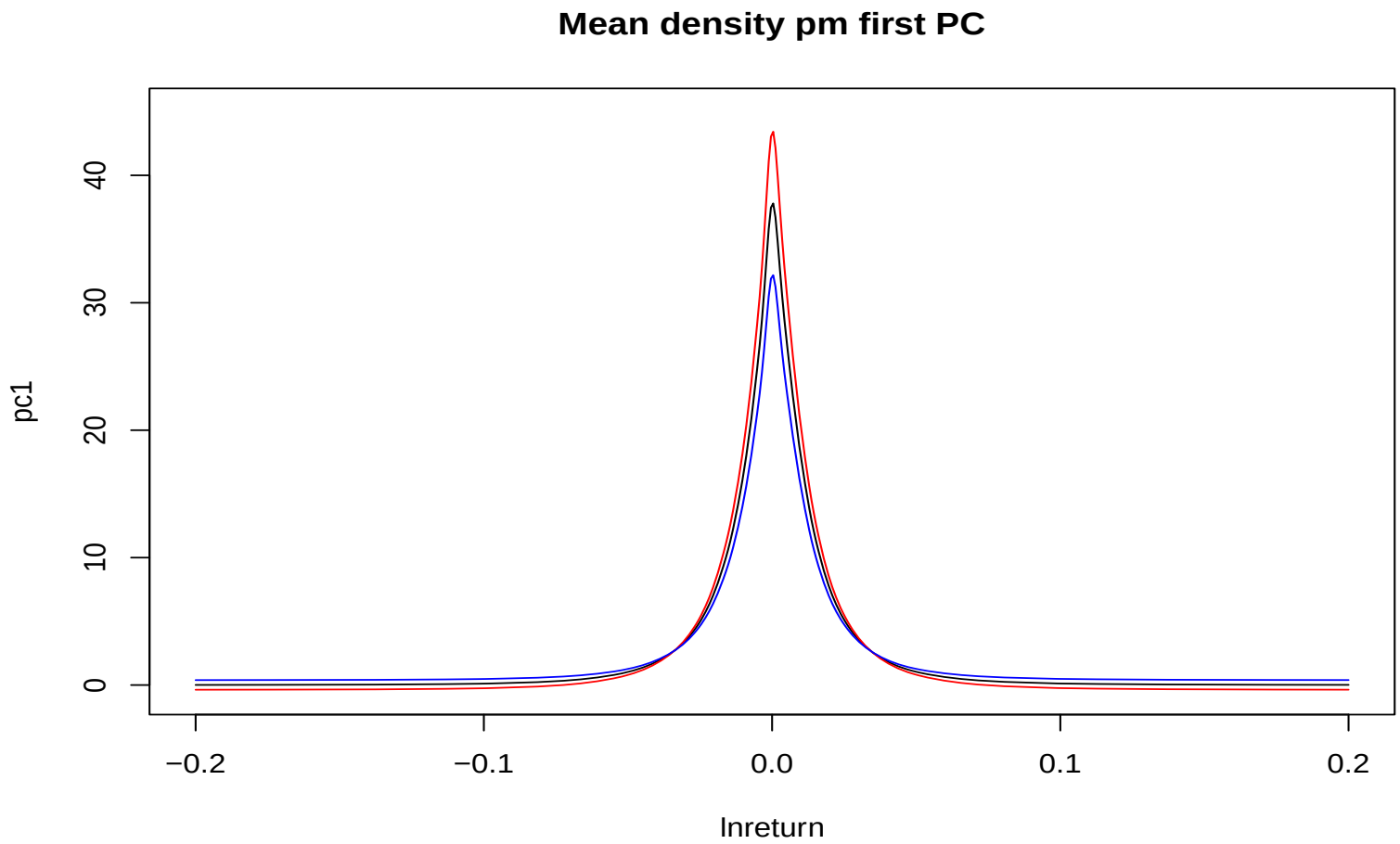

Figure: Mean density $\pm$ 1st PC: Peak and tails: mean+ standardized 1st PC (red).

Approximate factor models

$$f_t(x) = \sum_{i=1}^{p} \lambda_{t,i} g_i(x) + \epsilon_t(x),$$

where $g_i(x)$ denotes the ith common factor and $\epsilon_t(x)$ is the noise function.

1. A generalization of the orthogonal factor model, but allows the error functions to be correlated.
2. Only asymptotically identified under some regularity conditions.
3. FPCA provides a way to estimate approximate factor models.

Loadings of the first PC function

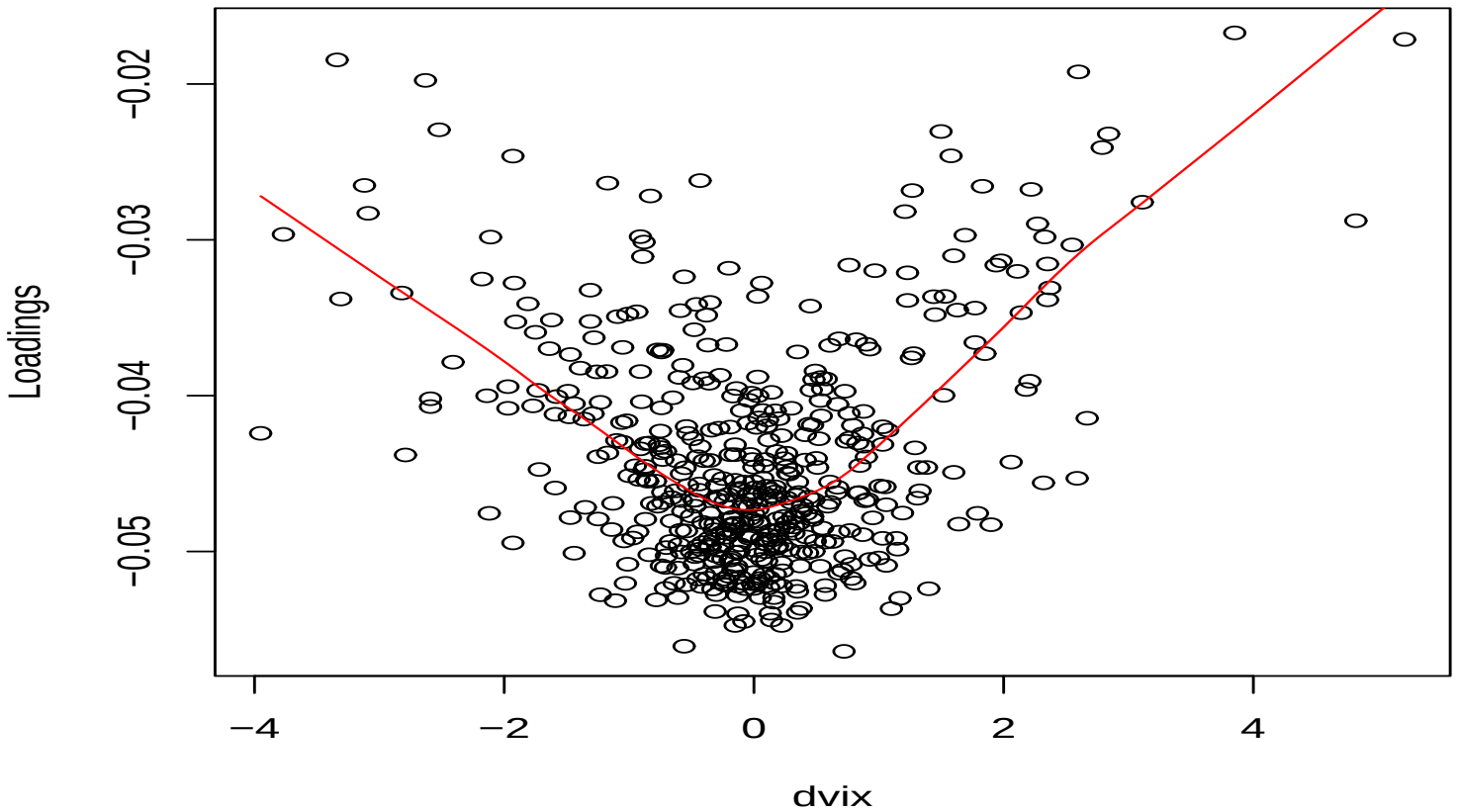

Figure: Scatter plot of loadings vs changes in VIX index. Red line denotes lowess fit

Functional PC via Thresholding

1. Zero appears to be a reasonable and natural threshold
2. Regime 1: dvix ≥ 0 with 244 days. [Volatile (bad) state]
3. Regime 2: dvix < 0 with 258 days. [Calm (good) state]
4. Perform PCA of density functions for each regime.
5. The differences are clearly seen.
6. Leads to different approximate factor models for the density functions

Scree plots

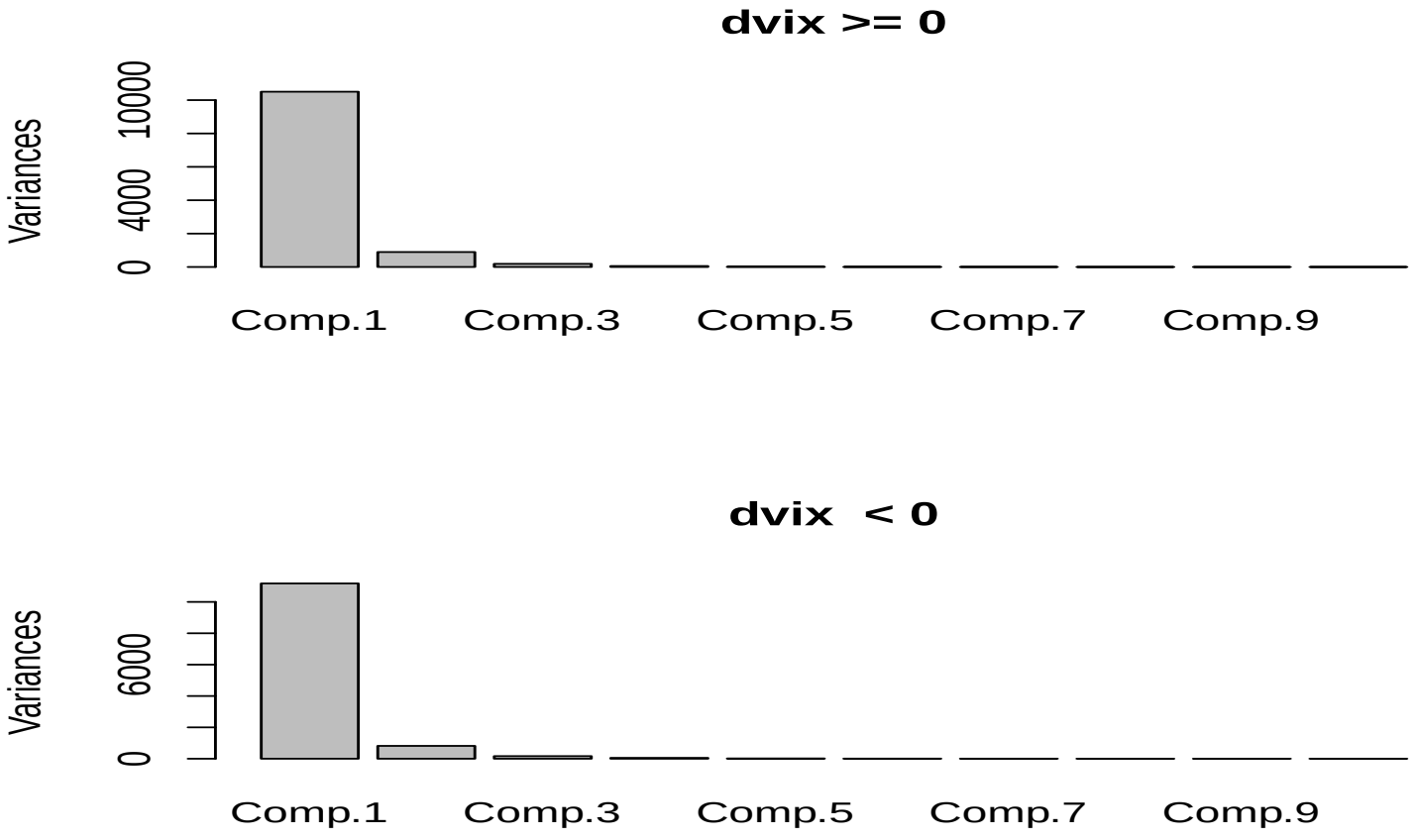

Figure: Scree plots of PCA for each regime

The first 6 PC functions

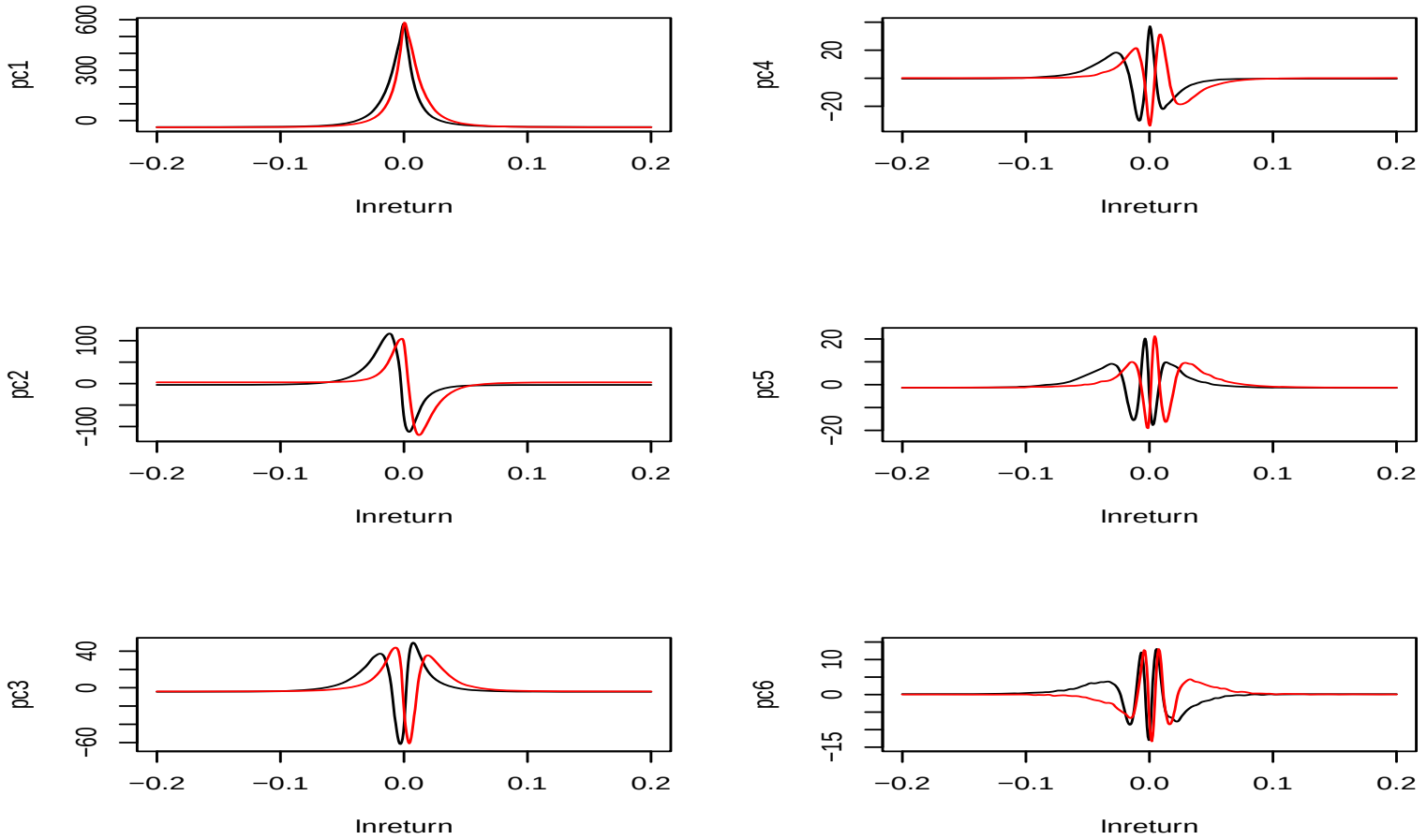

Figure: The first 6 PC functions for daily log return densities for each regime: red line is for the Calm state, Regime 2

Approximate factor models

1. Use approximate factor models with the first 12 principal component functions
2. Compare overall fits with/without thresholding
3. For Regime 1 (positive dvix): randomly select day 17
4. For Regime 2 (negative dvix): randomly select day 420.
5. Check: (a) observed vs fits and (b) residuals of with/without thresholding
6. With 12 components, both approaches fair well, but thresholding provides improvements.

Comparison: day 17 (in Regime 1)

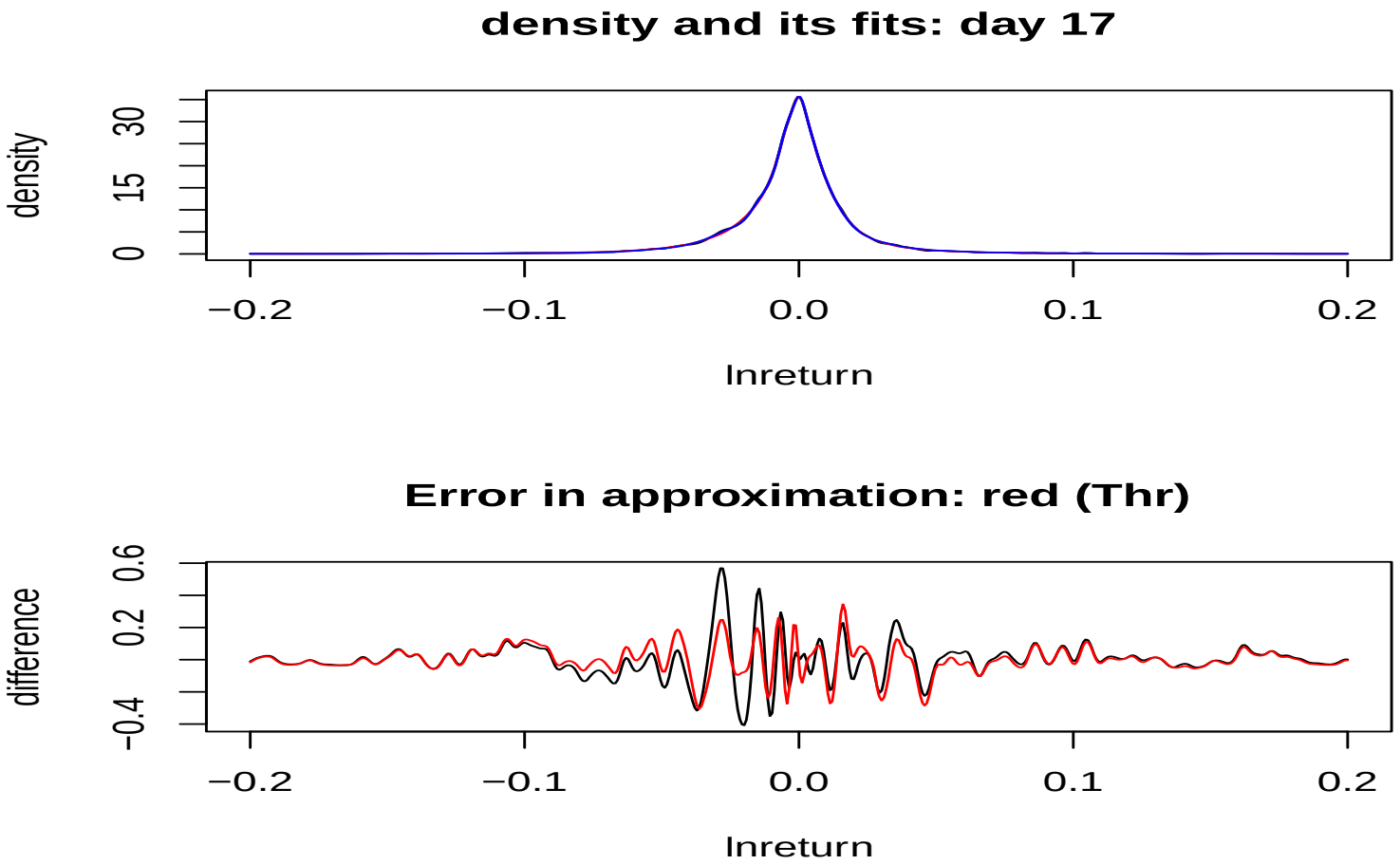

Figure: Top plot: observed (black), all (red), Thr (blue). Bottom plot: all (black), Thr (red)

Comparison: day 420 (in Regime 2)

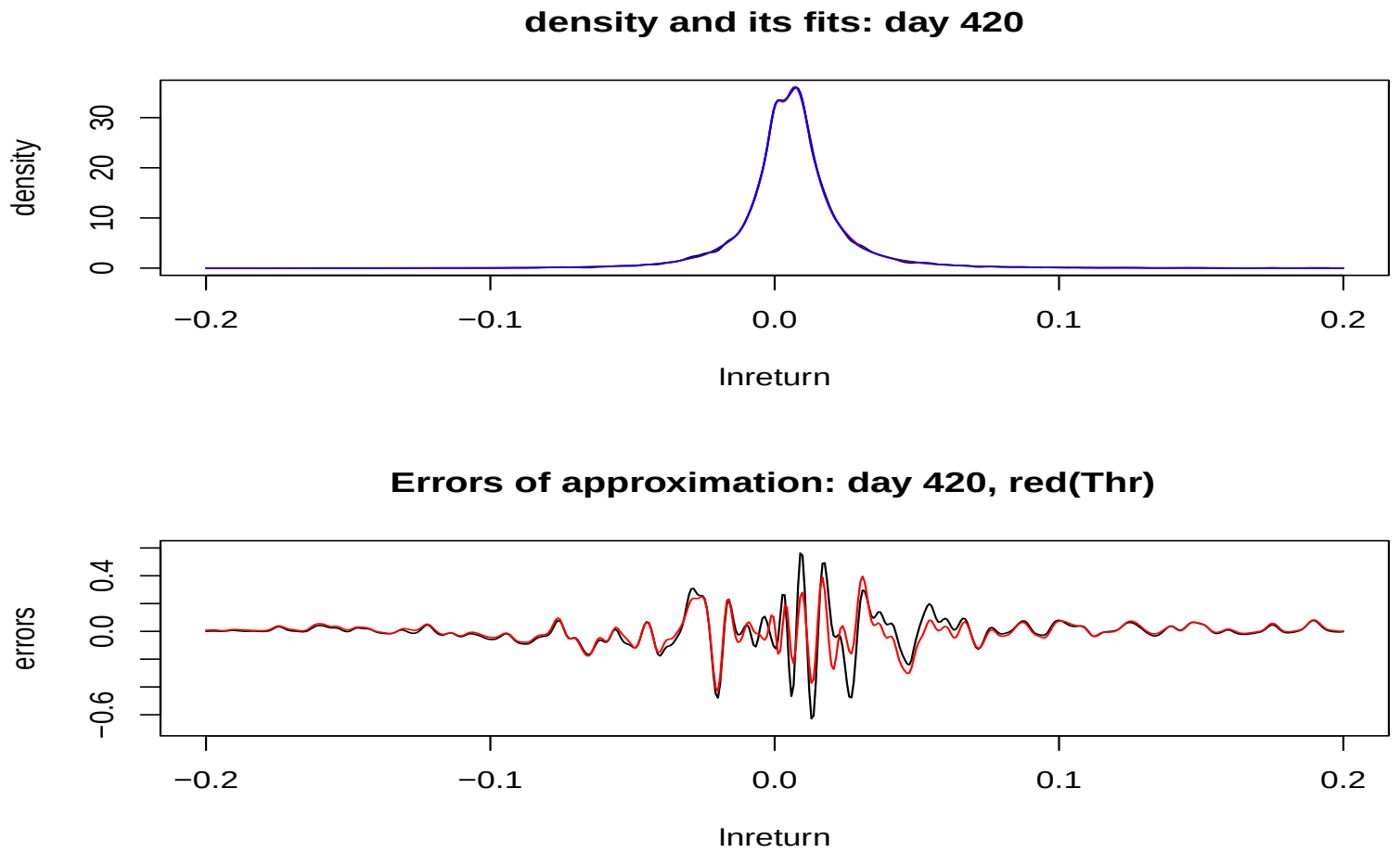

Figure: Top plot: observed (black), all (red), Thr (blue). Bottom plot: all (black), Thr (red)

Factor volatility models and implications

Some references

1. Conditionally uncorrelated components. Fan, Wang and Yao (2008)
2. Dynamic orthogonal components. Matteson and Tsay (2011). Use orthogonal transformation (parametrized by products of Given matrices) to obtain transformations that minimize the cross correlations of squared series.
3. Go-GARCH models. Van der Weide (2002)
   Perform transformation by PCA or the concept of independent component analysis.

Volatility co-movements

Volatility is driven by news, especially bad news. For assets with similar risk factors, price movements should be similar (APT).
If volatilities of assets have co-movements, then we can find portfolios that mitigate the fluctuations in volatility.
One way to search for such portfolios is to consider **Principal Volatility Component** (PVC) analysis. See Hu and Tsay (2014, JBES)

PVC analysis

Basic idea

Assume $E(\boldsymbol{r}_t|F_{t-1}) = \mathbf{0}$. Volatility is a function of the squared and cross-products of past asset returns, i.e.

$$vec(\boldsymbol{\Sigma}_t) = c_0 + \sum_{i=1}^{\infty} C_i vec(\boldsymbol{r}_{t-i}\boldsymbol{r}'_{t-i}).$$

Define the lag-ℓ generalized kurtosis matrix of $\boldsymbol{r}_t$ as

$$\gamma_\ell = \sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{Cov}^2(\boldsymbol{r}_t\boldsymbol{r}'_t, x_{ij,t-\ell}) = \sum_{i=1}^{k}\sum_{j=1}^{k} \gamma_{\ell,ij}\gamma'_{\ell,ij}$$

where $x_{ij,t-\ell}$ is a function of $r_{i,t-\ell}r_{j,t-\ell}$ for $1 \leq i,j \leq k$ and

$$\gamma_{\ell,ij} = \mathrm{Cov}(\boldsymbol{r}_t\boldsymbol{r}'_t, x_{ij,t-\ell}) = E[(\boldsymbol{r}_t\boldsymbol{r}'_t - \boldsymbol{\Sigma})(x_{ij,t-\ell} - E(x_{ij,t-\ell}))]$$

Discussion

1. The idea of generalized covariance matrix between a matrix variable and scalar variable is not new. It has been used in statistical literature before, e.g. Li (1992).
2. $\gamma_{\ell,ij}$ is symmetric
3. We use square to ensure γ_ℓ is non-negative definite.
4. $\gamma_\ell = 0$ if $\boldsymbol{r}_t \boldsymbol{r}_t'$ is not correlated with any element of $\boldsymbol{r}_{t-\ell} \boldsymbol{r}_{t-\ell}'$.
5. Let $\boldsymbol{z}_t = \boldsymbol{M}' \boldsymbol{r}_t$ where $\boldsymbol{M}$ is a $k \times k$ constant matrix, then

$$\text{Cov}(\boldsymbol{z}_t \boldsymbol{z}_t', x) = \text{Cov}(\boldsymbol{M}' \boldsymbol{r}_t \boldsymbol{r}_t' \boldsymbol{M}, x) = \boldsymbol{M}' \text{Cov}(\boldsymbol{r}_t \boldsymbol{r}_t', x) \boldsymbol{M}.$$

Cumulative generalized kurtosis matrix

For a given positive integer m, define

$$\boldsymbol{\Gamma}_m = \sum_{\ell=1}^{m} \boldsymbol{\gamma}_\ell.$$

Useful properties

1. Γ_m is symmetric and non-negative definite
2. If $y_t = \boldsymbol{u}'\boldsymbol{r}_t$ (a linear combination of $\boldsymbol{r}_t$) that has no ARCH effect, then $E(y_t^2|F_{t-1})$ is a constant. Consequently, $\boldsymbol{u}'\boldsymbol{\gamma}_{\ell,ij} = 0$, implying $\boldsymbol{\gamma}_{\ell,ij}$ is singular, so is $\boldsymbol{\Gamma}_m$.
3. On the other hand, if $\boldsymbol{\Gamma}_m\boldsymbol{u} = 0$ with $\boldsymbol{u} \leq \boldsymbol{0}$, then $\boldsymbol{\gamma}_\ell\boldsymbol{u} = 0$ for $1 \leq \ell \leq m$. From $\boldsymbol{u}'\boldsymbol{\gamma}_\ell u = 0$, it follows that $\boldsymbol{u}'\boldsymbol{\gamma}_{\ell,ij}\boldsymbol{\gamma}'_{\ell,ij}\boldsymbol{u} = 0$ for all i, j. Consequently, $\boldsymbol{\gamma}_{\ell,ij}\boldsymbol{u} = 0$, implying that $y_t = \boldsymbol{u}'\boldsymbol{r}_t$ is not correlated with $r_{i,t-\ell}r_{j,t-\ell}$.

Properties continued

In particular, if $m = \infty$, then $y_t = \boldsymbol{u}'\boldsymbol{r}_t$ satisfies that $E(y_t^2|F_{t-1})$ is not correlated with $r_{i,t-\ell}y_{j,t-\ell}$ for all i, j and ℓ. This implies y_t has no ARCH effects.

PVC analysis: Perform eigenvalue-eigenvector analysis of a sample estimate of Γ_m. In particular, exam the number of zero eigenvalues and their associate eigenvectors.

An application

Consider monthly returns of Fama-French 25 portfolios from July 1963 to May 2015. Denoted by $\boldsymbol{z}_t$. Data from French's web.
Also, use Fama-French 3 factors as explanatory variables. Denoted by F_t.
Employ the adjusted portfolio returns $\boldsymbol{r}_t$ (residuals)

$$\boldsymbol{z}_t = \beta \boldsymbol{F}_t + \boldsymbol{r}_t$$

The $\boldsymbol{r}_t$ series has no serial correlations, but show strong ARCH effects (as expected).
Applying the PVC analysis to $\boldsymbol{r}_t$, we found three eigenvalues close to zero. The associated eigenvectors give rise to three portfolios that have no ARCH effects.

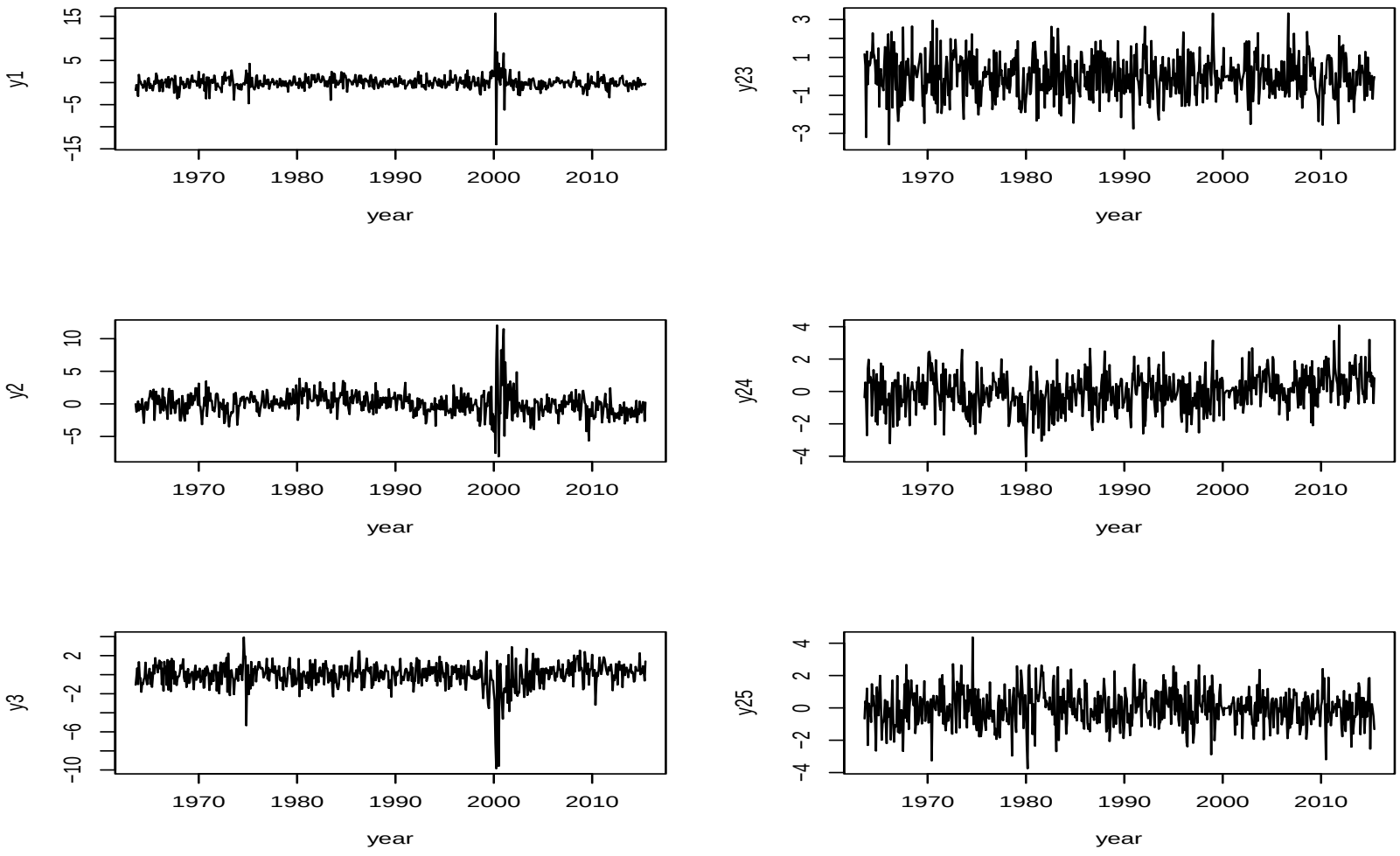

Figure: Time plots of the first 3 and last 3 transformed series of PVC analysis applied to Fama-French 25 portfolios. Monthly return from 1963.7 to 2015.5.

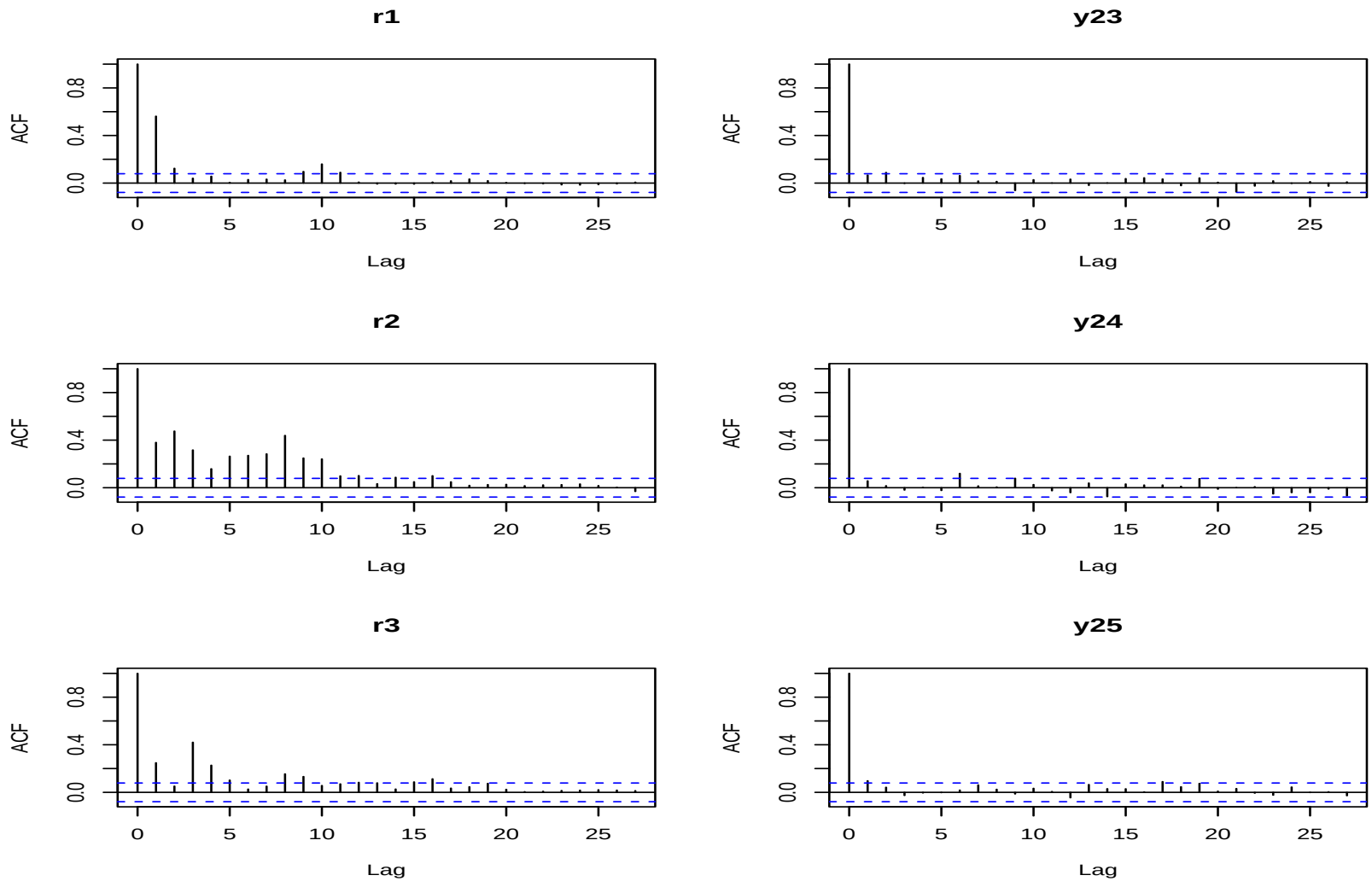

Figure: ACF of squared series of the first 3 and last 3 transformed series of PVC analysis applied to Fama-French 25 portfolios: monthly returns from 1963.7 to 2015.5.

Lesson learned: Are we happy?

1. The example demonstrates that PVC works well.
   - It detects 3 linear combinations that have constant conditional variances
   - There exist common factors in the volatility of $\boldsymbol{r}_t$.
2. Recall the factor model

$$\boldsymbol{z}_t = \beta \boldsymbol{F}_t + \boldsymbol{r}_t,$$

   where the dimensions of $\boldsymbol{z}_t$ and $\boldsymbol{F}_t$ are 25 and 3, respectively. What is the dimension of $\boldsymbol{r}_t$?

Many researchers in factor models assume $\boldsymbol{r}_t$ has the same dimension as $\boldsymbol{z}_t$. Is it really?
Artificial dimension: In the example, dimension of $\boldsymbol{r}_t$ should be 22, not 25. What PVC analysis found is simply to confirm that $\dim(\boldsymbol{F}_t) = 3$. There is nothing to be happy about!

Lasso and beyond

1. Need to exploit parsimony, beyond sparsity
2. Need to take into account prior knowledge. We have accumulated lot of knowledge in diverse scientific areas. How to take advantages of this knowledge?
3. Variable selection is not sufficient. More importantly, what are the proper measurements to take? What questions can a given big data answer?

An illustration

Every country has many interest series

1. have different maturities
2. serve different financial purposes
3. What is the information embedded in those interest rate series?

Consider U.S. weekly constant maturity interest rates

1. From January 8, 1982 to October 30, 2015
2. Maturities: 3m, 6m, 1y, 2y, 3y, 5y, 7y, 10y, and 30y*

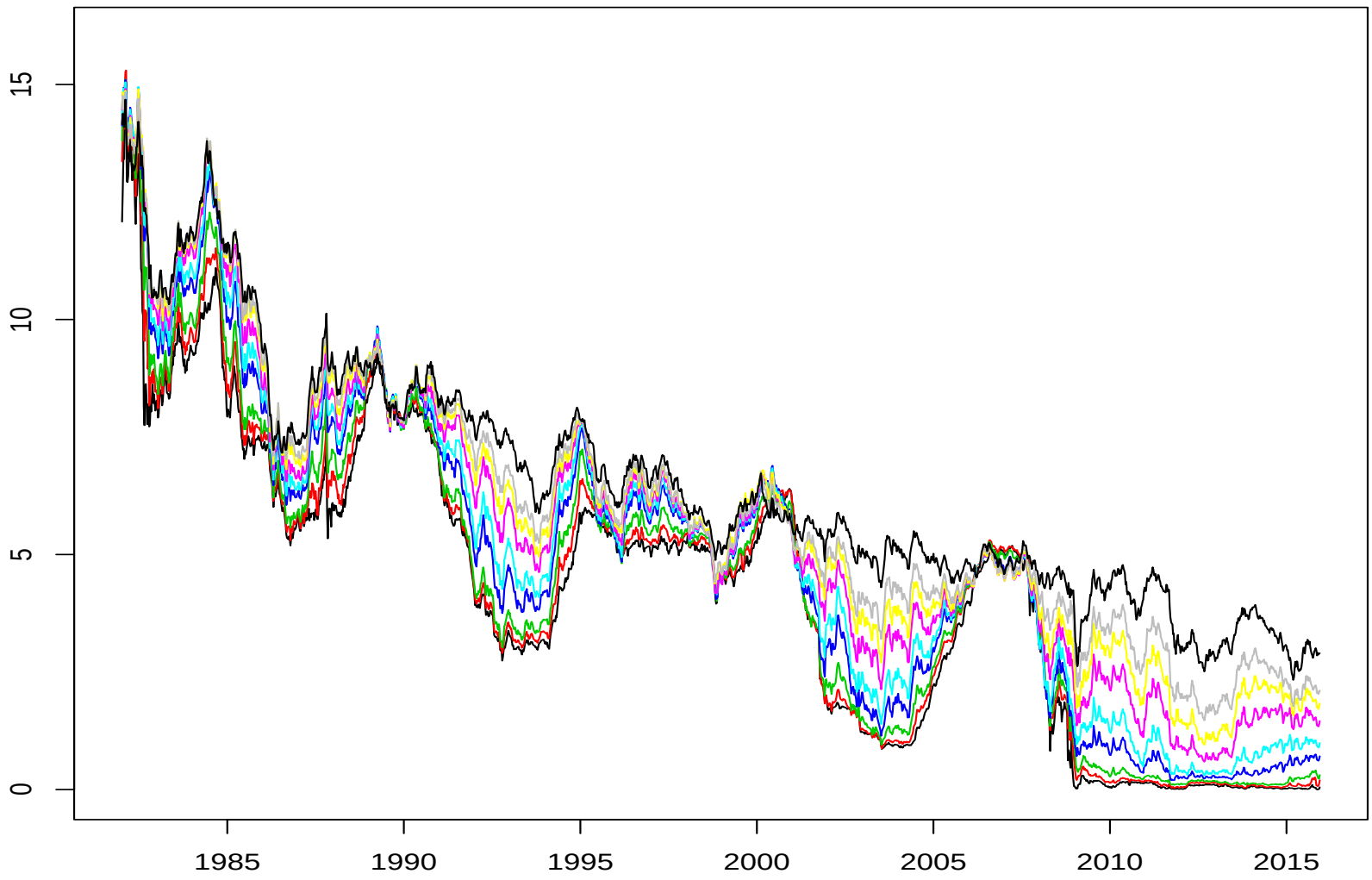

Figure: Time plots of U.S. weekly interest rates with different maturities: 1/8/1982 to 10/30/2015.

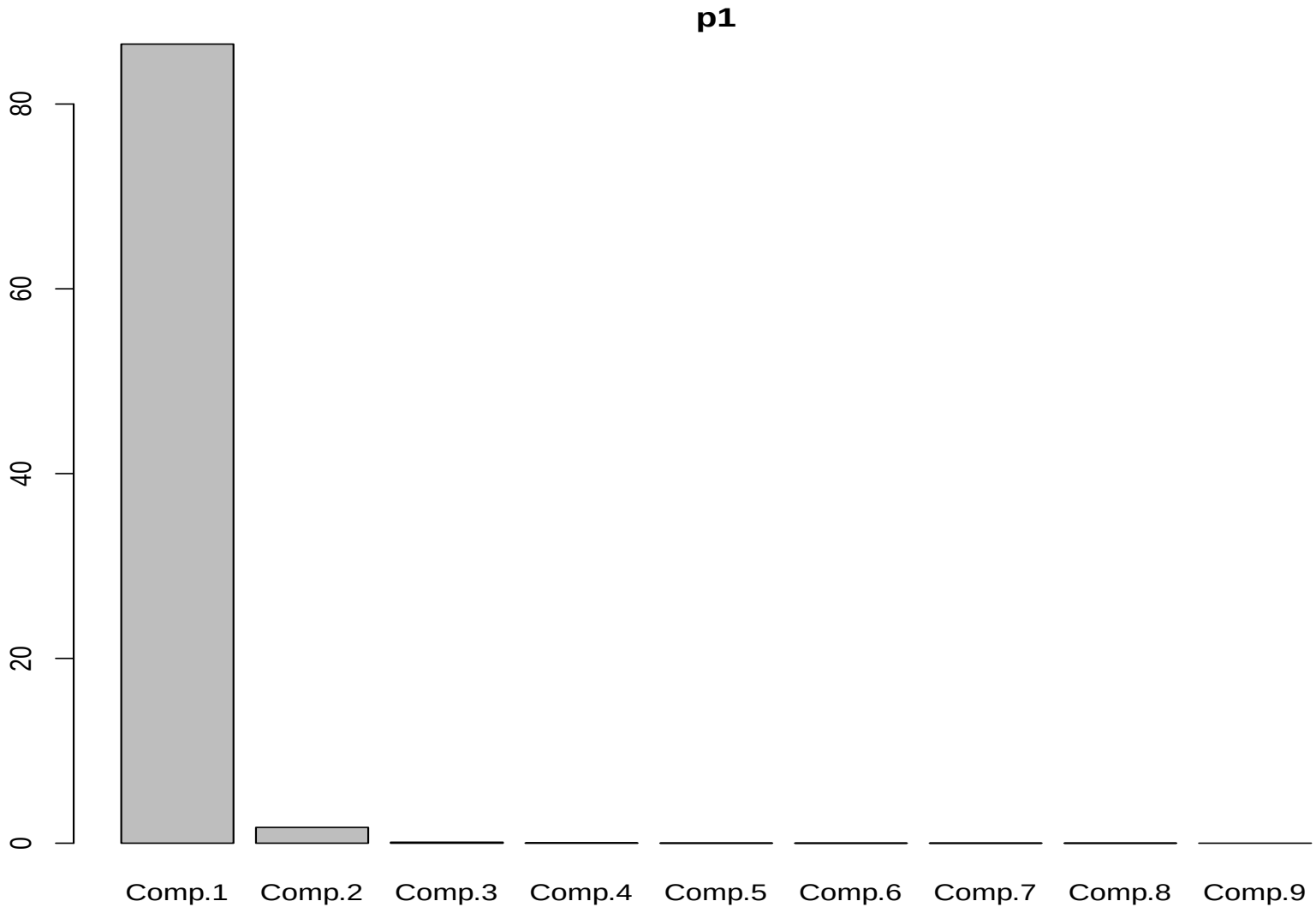

Figure: Screeplot of U.S. weekly interest rates.

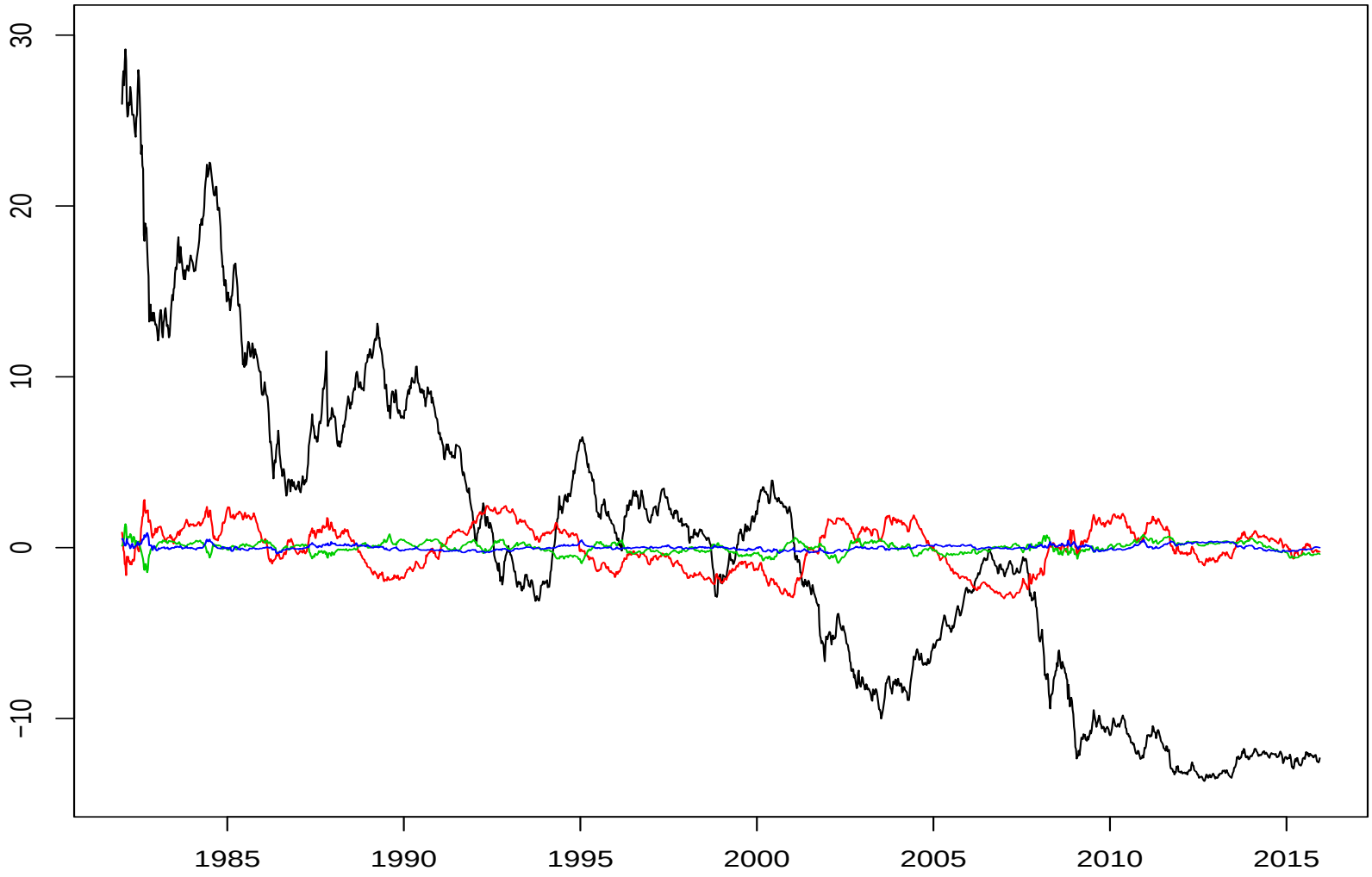

Figure: Time plots of the first four principal components of U.S. weekly interest rates

Implication?

In lasso-type of analysis,

1. should we use the interest rate series directly? Even with group lasso.
   This leads to sparsity.
2. should we apply PCA first, then use the PCs?
   This leads to parsimony.
3. should we develop other possibilities?

Concluding Remarks

1. There are many open and interesting questions for high-dimensional time series analysis
2. Computation? Extracting information? Presentation?
3. Relation to reduced rank regression? Random projection.
4. Simultaneous analysis of continuous and discrete-value time series
5. Spatio-temporal series beyond separable models