Principal Component Analysis

Statistical Methods in Finance

Motivation

- Many financial data are characterized by a high degree of collinearity.
- We often need to generate a large set of scenarios based on movements in many risk factors for risk management or pricing models. The computation is expensive due to the large dimensions of factors.
- Variables are highly correlated when there are only a few important sources of information in the data that are common to many variables.
- Examples:
  - Term structures of interest rates
  - Implied volatility of different assets on the same underlying
  - Futures of different lags/maturities on the same underlying

Idea: Extract the most important uncorrelated sources of variation in a multivariate system — principal component analysis (PCA).

Motivation

- Dimension Reduction
  - Introduce a data transformation tool for correlated financial systems.
  - The reduction in dimensionality is achieved by taking only the first *m* principal components.
  - This transformation significantly reduces the computation time.
- The orthogonal property
  - The sample covariance matrix is not always positive definite (the determinant is about 0). This can be caused by high linear dependency of one variable to another, or large amounts of missing data. In this case, the sample covariance matrix can not be used for portfolio optimization as the singular matrix does not have an inverse.
  - Principal components are orthogonal, and their covariance matrix is diagonal. The principal components can be transformed into a positive definite covariance matrix.

Basics

- The data used by PCA analysis must be stationary.
  - Prices, rates or yields are generally non-stationary and so they will be transformed into returns, before PCA is applied.
- The returns need to be normalized before PCA analysis.
  - Otherwise, the first PC will be dominated by the input variable with the greatest volatility.
- PCA analysis is based on the eigenvalue and eigenvector analysis of the correlation/covariance matrix.

Eigenvalues and Eigenvectors

- Let $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ be a random vector with mean μ and covariance matrix V.
- The first principal component $\boldsymbol{a}_1^T \boldsymbol{x}$ is defined such that

$$\boldsymbol{a}_1 := \underset{\boldsymbol{a}:||\boldsymbol{a}||=1}{\arg\max} \operatorname{Var}(\boldsymbol{a}^T \boldsymbol{x}) = \underset{\boldsymbol{a}:||\boldsymbol{a}||=1}{\arg\max} (\boldsymbol{a}^T V \boldsymbol{a}).$$

- To find $\boldsymbol{a}_1$, introduce the Lagrange multiplier λ and set

$$\frac{\partial}{\partial a_i} \left(\boldsymbol{a}^T V \boldsymbol{a} + \lambda (1 - \boldsymbol{a}^T \boldsymbol{a}) \right) = 0, \quad \text{for } i = 1, \ldots, p.$$

  This gives

$$V \boldsymbol{a} = \lambda \boldsymbol{a}$$

  whichh implies that

  - λ is the largest eigenvalue of V; $\boldsymbol{a}_1$ is the corresponding eigenvector.

Eigenvalues and Eigenvectors

- The second principal component $\boldsymbol{a}_2^T \boldsymbol{x}$ is defined such that

$$\boldsymbol{a}_2 := \underset{\boldsymbol{a}:\boldsymbol{a}^T\boldsymbol{a}_1=0,||\boldsymbol{a}||=1}{\arg\max} \operatorname{Var}(\boldsymbol{a}^T\boldsymbol{x}) = \underset{\boldsymbol{a}:\boldsymbol{a}^T\boldsymbol{a}_1=0,||\boldsymbol{a}||=1}{\arg\max} (\boldsymbol{a}^T V \boldsymbol{a}).$$

- One can show that $\boldsymbol{a}_2$ is the eigenvector corresponding to the second largest eigenvalue of V.
- The third principal component can be defined similarly.
- Another way to derive the minimum variance properties of principal components is to use the spectral decomposition $V = \sum_{i=1}^{p} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T$ and get

$$V\boldsymbol{u} = \lambda_1(\boldsymbol{u}^T\boldsymbol{a}_1)\boldsymbol{a}_1 + \ldots \lambda_p(\boldsymbol{u}^T\boldsymbol{a}_p)\boldsymbol{a}_p,$$

where λ_i is the ith largest eigenvalue of V and $\boldsymbol{a}_i$ is the corresponding eigenvector.

Eigenvalues and Eigenvectors

- Since V is symmetric, its eigenvalues are real and can be ordered as $\lambda_1 \geq \ldots \lambda_p$. They are all nonnegative since V is nonnegative definite. Moreover,

$$\operatorname{tr}(V) = \lambda_1 + \ldots + \lambda_p \text{ and } \det(V) = \lambda_1 \cdots \lambda_p.$$

- Let $\boldsymbol{a}_j$ be the eigenvector corresponding to λ_j, then the eigenvectors are orthogonal to each other.
- Def: $\boldsymbol{a}_i^T \boldsymbol{x}$ is called the ith principal component of $\boldsymbol{x}$.

Properties on eigensystem

(a) $V = \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1^T + \ldots + \lambda_p \boldsymbol{a}_p \boldsymbol{a}_p^T$.

Proof.

From the definition of eigenvectors and eigenvalues,

$$V(\boldsymbol{a}_1 \cdots \boldsymbol{a}_p) = (\boldsymbol{a}_1 \cdots \boldsymbol{a}_p)\mathrm{diag}(\lambda_1, \ldots, \lambda_p).$$

Since $(\boldsymbol{a}_1 \cdots \boldsymbol{a}_p)$ is an orthogonal matrix,

$$V = (\boldsymbol{a}_1 \cdots \boldsymbol{a}_p) \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_p^T \end{pmatrix}.$$

This can be written as

$$V = (\boldsymbol{a}_1 \lambda_1 \cdots \boldsymbol{a}_p \lambda_p) \begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_p^T \end{pmatrix} = \sum_{i=1}^{p} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T.$$

Constructing the principal components

- Suppose we have a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of n independent observations from a multivariate population with mean μ and covariance matrix V.

$$\hat{\boldsymbol{\mu}} = \bar{x} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \hat{V} = \frac{\mathbf{X}^T \mathbf{X}}{n-1},$$

where

$$\mathbf{X} := \begin{pmatrix} x_{11} - \bar{x}_1 & \ldots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \ldots & x_{np} - \bar{x}_p. \end{pmatrix} =: (\mathbf{X}_1, \ldots, \mathbf{X}_p).$$

Constructing the principal components

- The jth principal component of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is the linear combination as follows

$$\mathbf{Y}_j = (\mathbf{X}_1, \ldots, \mathbf{X}_p)\hat{\boldsymbol{a}}_j = \hat{a}_{j1}\mathbf{X}_1 + \ldots + \hat{a}_{jp}\mathbf{X}_p,$$

  where $\hat{\boldsymbol{a}}_j = (\hat{a}_{j1}, \ldots, \hat{a}_{jp})^T$ is the eigenvector corresponding to the jth largest eigenvalue $\hat{\lambda}_j$ of the sample covariance matrix $\hat{V}$.

Properties on eigensystem

(a) $V = \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1^T + \ldots + \lambda_p \boldsymbol{a}_p \boldsymbol{a}_p^T$.

(b) $\sum_{i=1}^{p} \mathrm{Var}(x_i) = \mathrm{tr}(V) = \lambda_1 + \ldots + \lambda_p$.

(c) $\mathrm{Var}(\boldsymbol{a}_i^T \boldsymbol{x}) = \lambda_i$.

(d) (usually) Only a few components accounts for the total variance:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\mathrm{tr}(V)} \text{ is near 1 for some all } k.$$

(e) Factor loadings are columns giving the elements of the column vectors $\boldsymbol{a}_i$ for the principal components $\boldsymbol{a}_i^T \boldsymbol{x}$.

Applications: As a dimension reduction technique

(1) When p is large, we need to estimate $\frac{p(p+1)}{2}$ parameters for the $p \times p$ covariance matrix.

(2) If many λ_i's are small, only a few, say k, principal components are involved in the standard error of the estimate $\hat{\boldsymbol{a}}_i$ for $1 \leq i \leq k$.

Real examples: Term structure of interest rates, implied volatilities,. . ..

Applications: the multi-factor model

- Choose the first k principal components of asset returns as k leading factors in multifactor models.
- The first principal component is the (normalized) linear combination of asset returns that gives the largest proportion of the variance.

An alternative to PAC: Factor Analysis

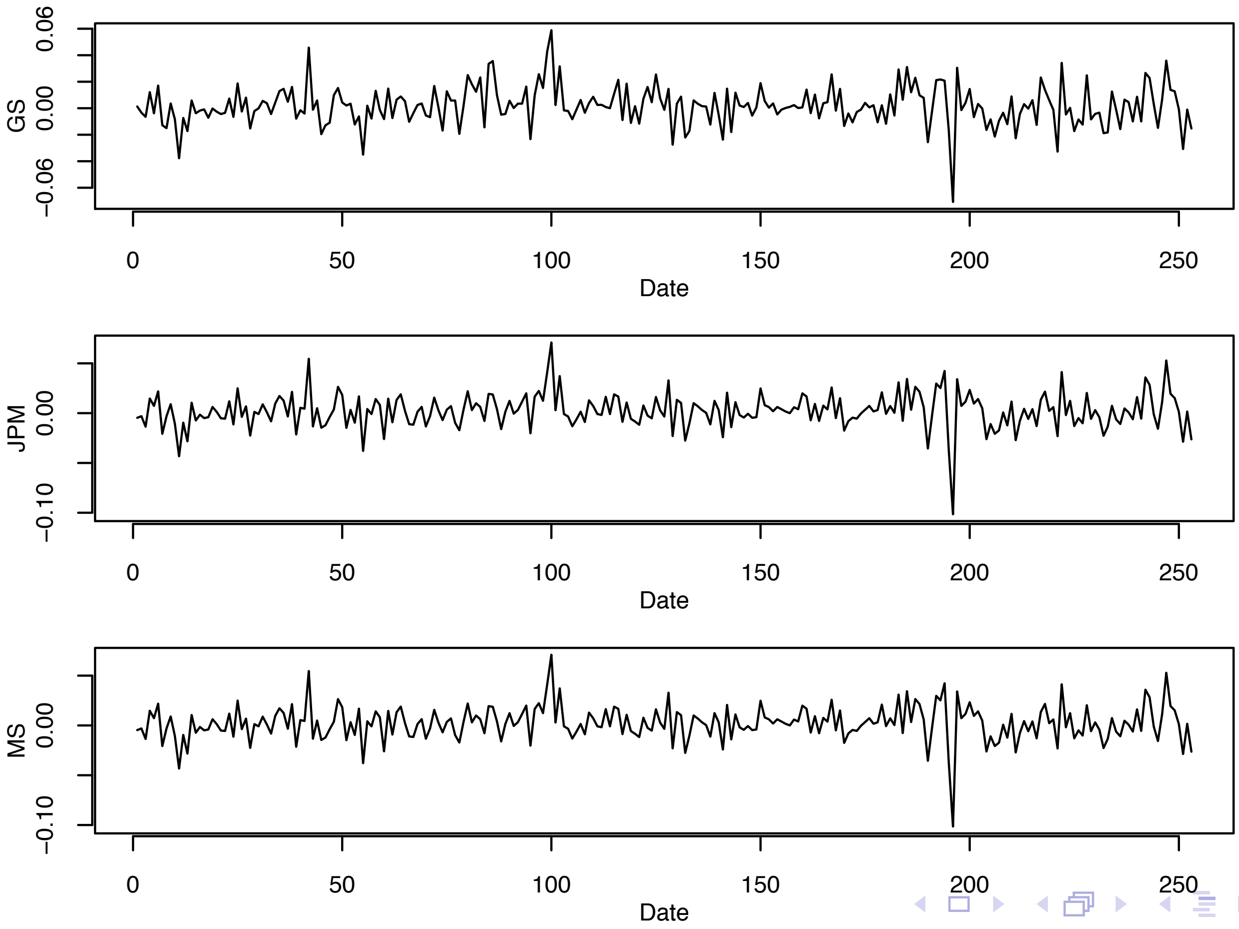
- Factor model assumes that there are indeed k factors so that

$$\boldsymbol{r}_t = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t,$$

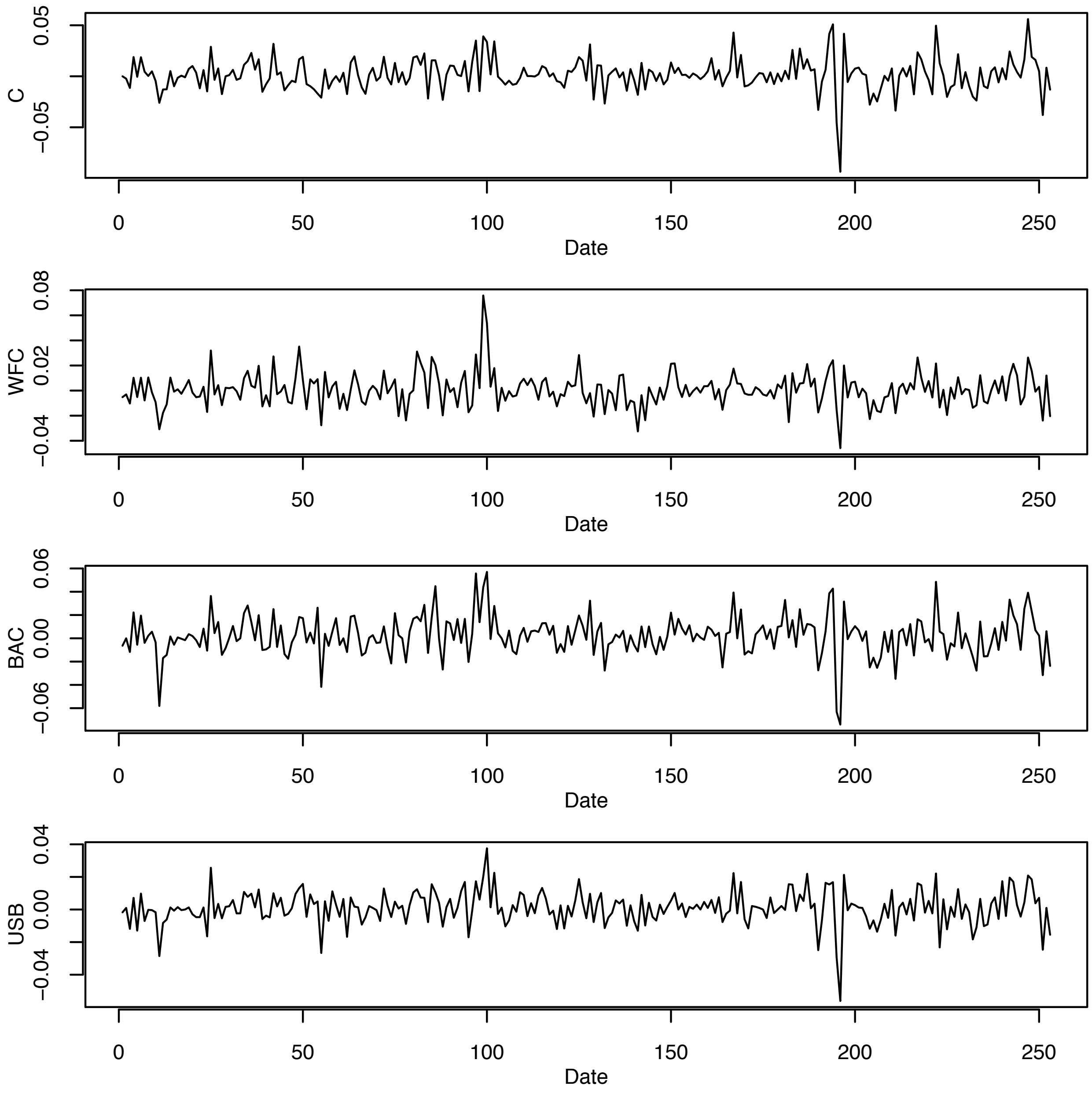
  where $\boldsymbol{F}_t = (F_{t1}, \ldots, F_{tk})^T$ is a vector of unobservable factors and $\text{Corr}(\boldsymbol{F}_t, \boldsymbol{\varepsilon}_t) = 0$. $\mathbf{B}$ is a matrix of factor loadings.
- Unlike principal component analysis, factor analysis requires prespecification of the number of factors and distribution assumptions on $\boldsymbol{\varepsilon}_t$.

Stocks from the Same Sector: Time Series Plots

- Time series plots of the daily returns of 7 financial companies: Goldman Sachs (GS), JPMorgan (JPM), Morgan Stanles (MS), CITI (C), Wellls Fargo (WFC), Bank of America (BAC) and U.S. Bancorp (USB).
- Apr 4, 2016 - Apr 4, 2017



Stocks from the Same Sector: Time Series Plots



Stocks from the Same Sector: Heat map

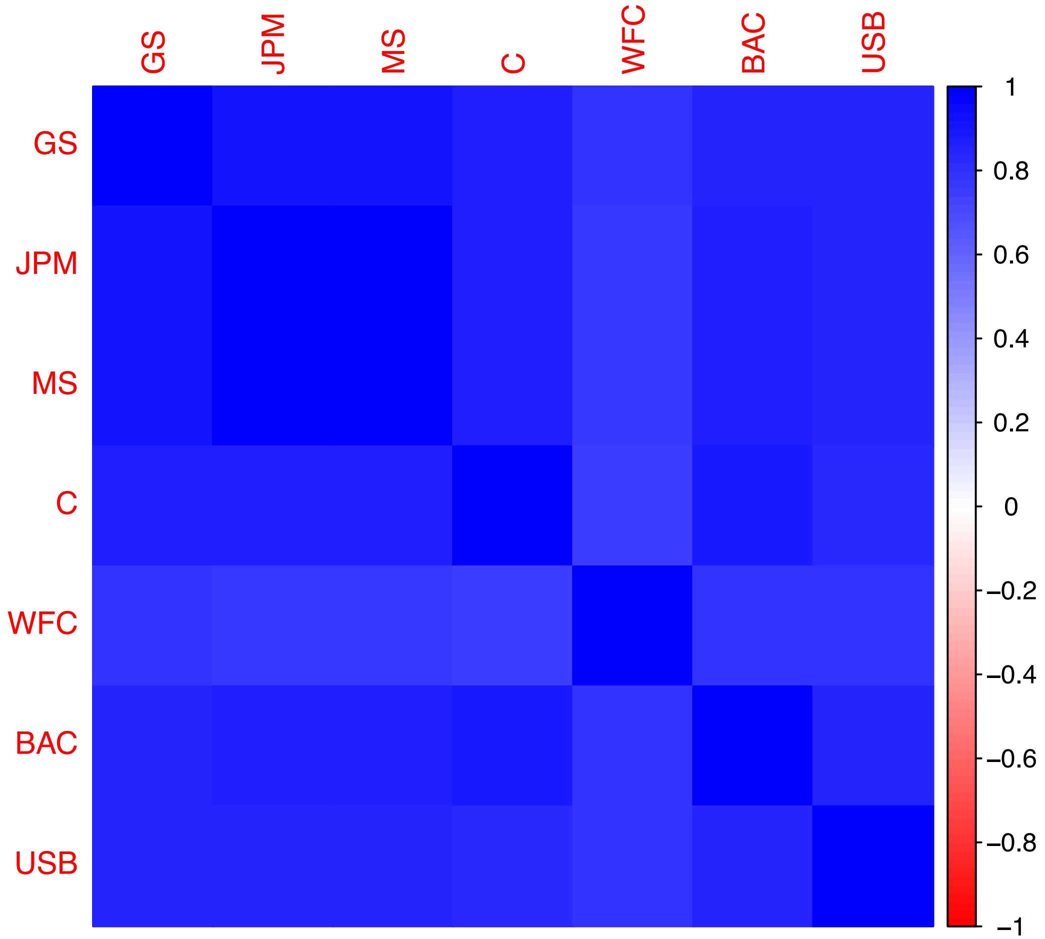
- A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors.
- One useful `R` package for generating heat map is `corrplot`. To use it:

  ```
  install.packages("corrplot")
  library(corrplot)
  ```

- Read more about the package at https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

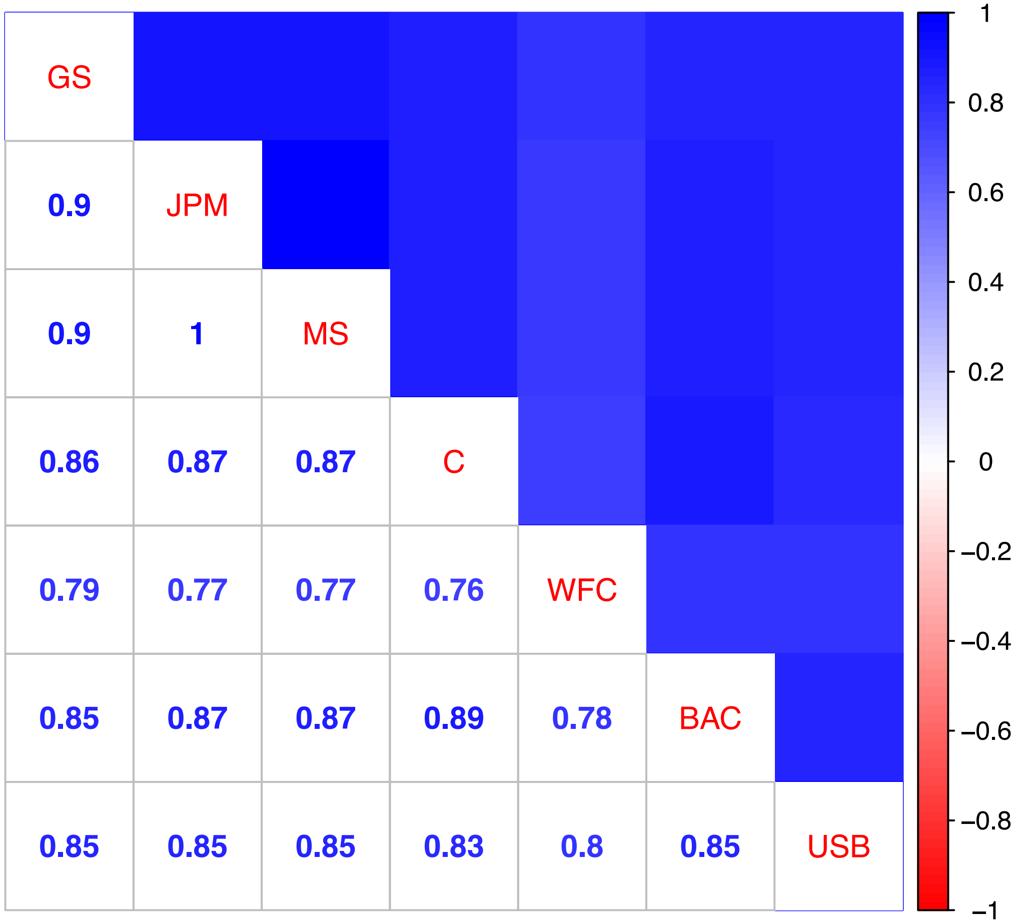
Stocks from the Same Sector: Heat map

Heat map of the correlation matrix of returns



Stocks from the Same Sector: Heat map

Heat map of the correlation matrix of returns (an alternative way to plot)



GS						
0.9	JPM					
0.9	1	MS				
0.86	0.87	0.87	C			
0.79	0.77	0.77	0.76	WFC		
0.85	0.87	0.87	0.89	0.78	BAC	
0.85	0.85	0.85	0.83	0.8	0.85	USB

Stocks from the Same Sector: PCA

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.47	0.56	0.45	0.40	0.35	0.32	0.00
Proportion of Variance	0.87	0.04	0.03	0.02	0.02	0.01	0.00
Cumulative Proportion	0.87	0.91	0.94	0.97	0.99	1.00	1.00

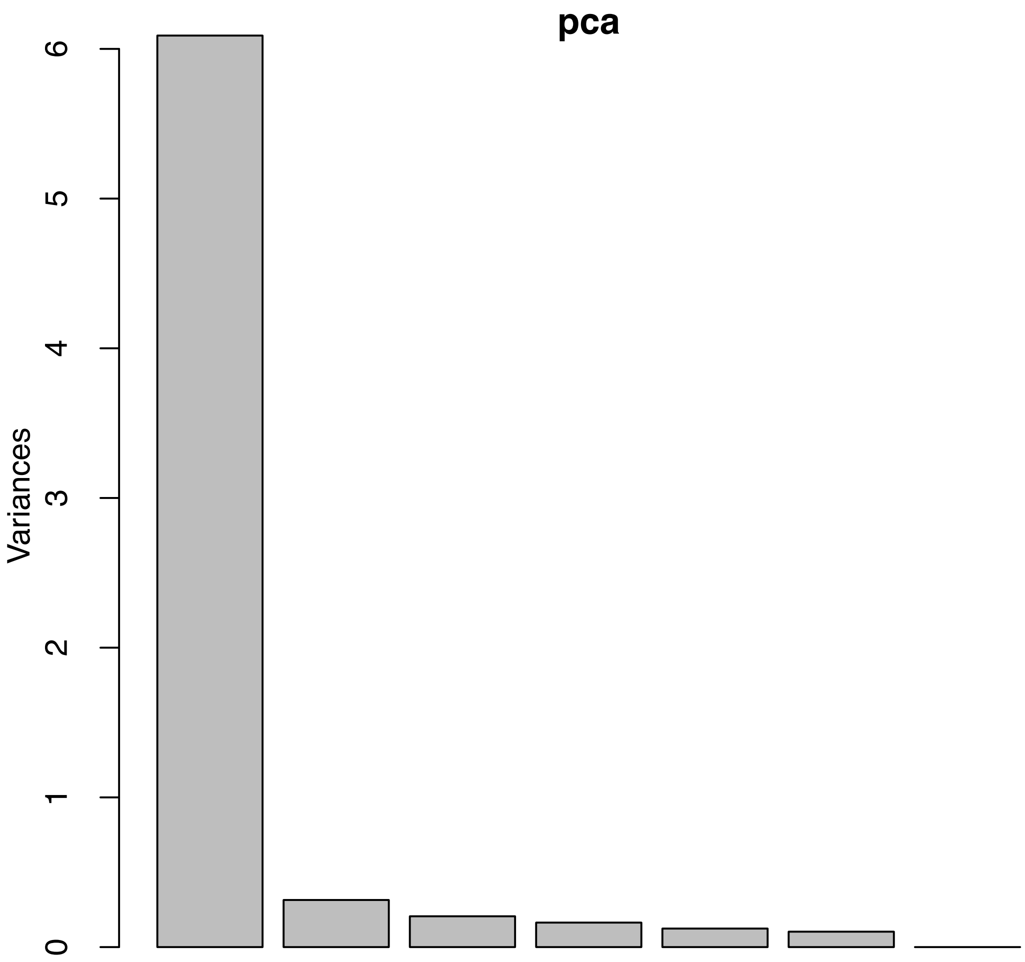
Loadings:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
1	0.38	-0.10	0.26	-0.12	0.73	0.49	0.00
2	0.39	-0.34	0.36	-0.00	-0.30	-0.12	0.71
3	0.39	-0.34	0.36	-0.00	-0.30	-0.12	-0.71
4	0.38	-0.16	-0.51	-0.35	0.31	-0.60	0.00
5	0.35	0.84	0.26	-0.26	-0.12	-0.14	0.00
6	0.38	0.01	-0.56	-0.16	-0.42	0.59	0.00
7	0.37	0.17	-0.19	0.88	0.11	-0.11	-0.00

Interpretation: 1st PC represents the general movement of the sector

Stocks from the Same Sector: PCA

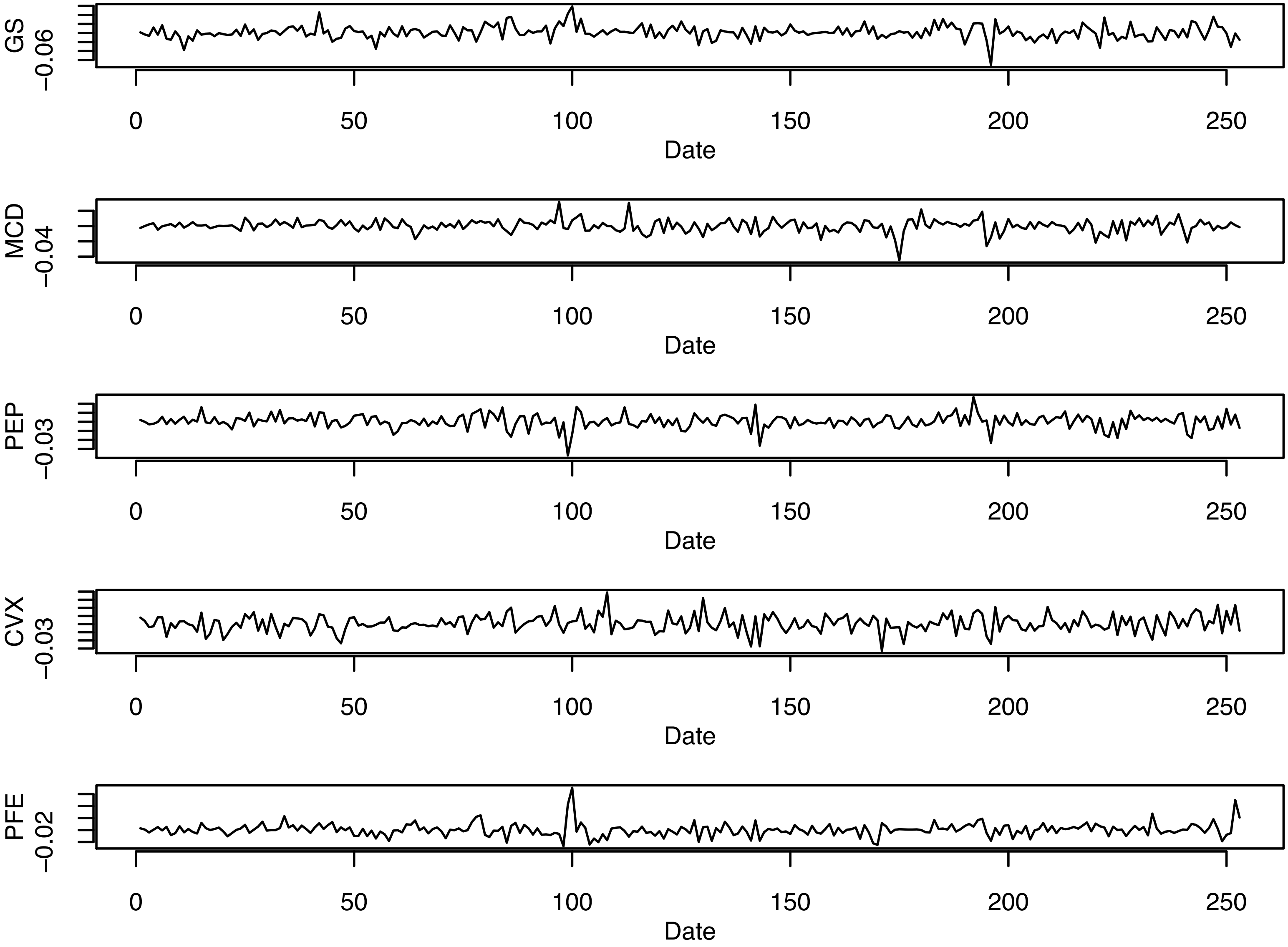
Variance of each principal component:



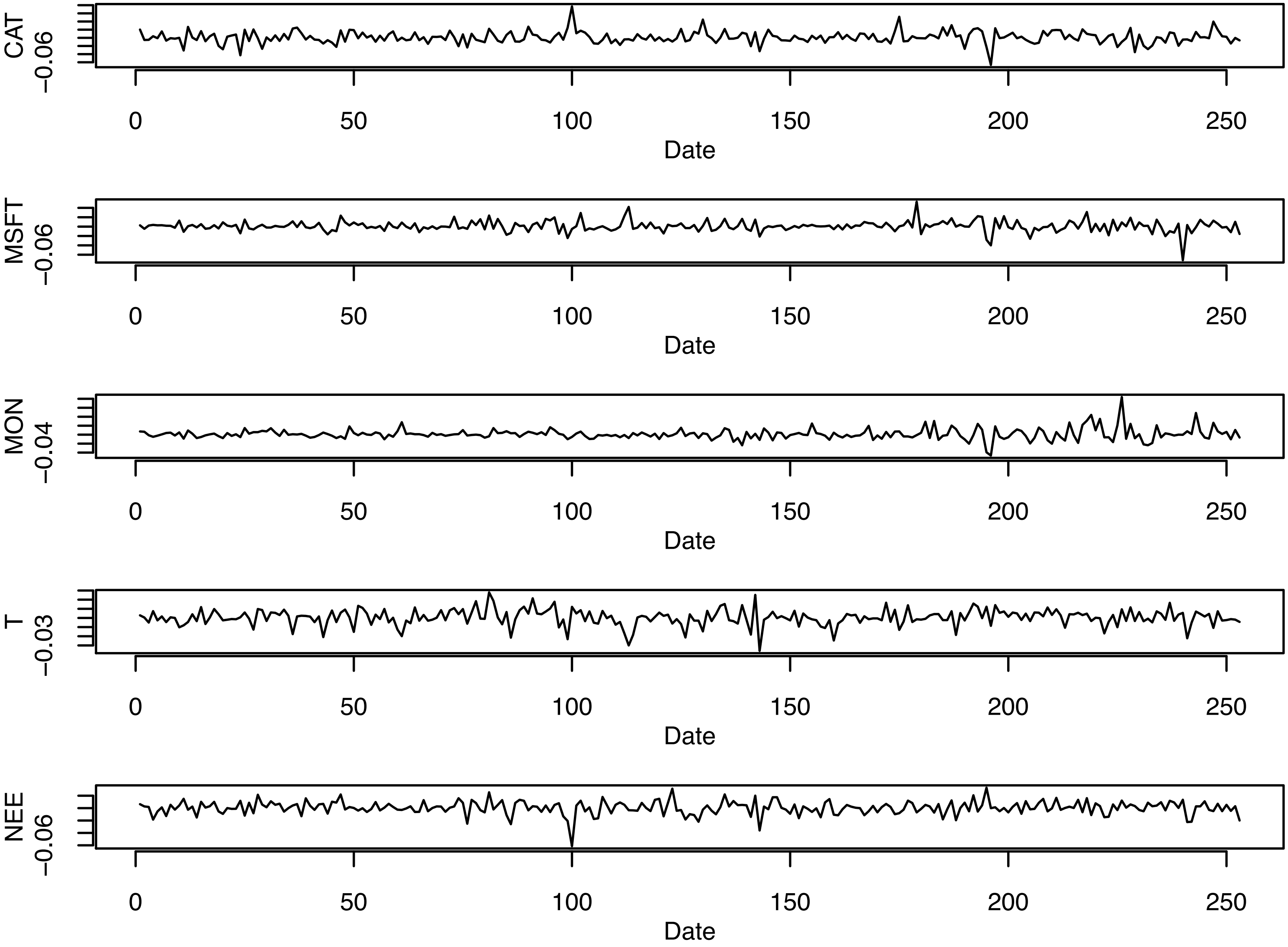
Stocks from Different Sectors

- Now, we consider stocks from different sectors: GS (XLF), MCD (XLY), PEP (XLP), CVX (XLE), PFE (XLV), CAT (XLI), MSFT (XLK), MON (XLB), T (XLK) and NEE (XLU).
- Apr 4, 2016 - Apr 4, 2017
- XLF–Financials, XLY–Consumer Discretionary, XLP–Consumer Staples, XLE–Energy, XLV–Health Care, XLI–Industrials, XLK–Technology, XLB–Materials, XLU–Utilities

Stocks from Different Sectors: Time Series Plots

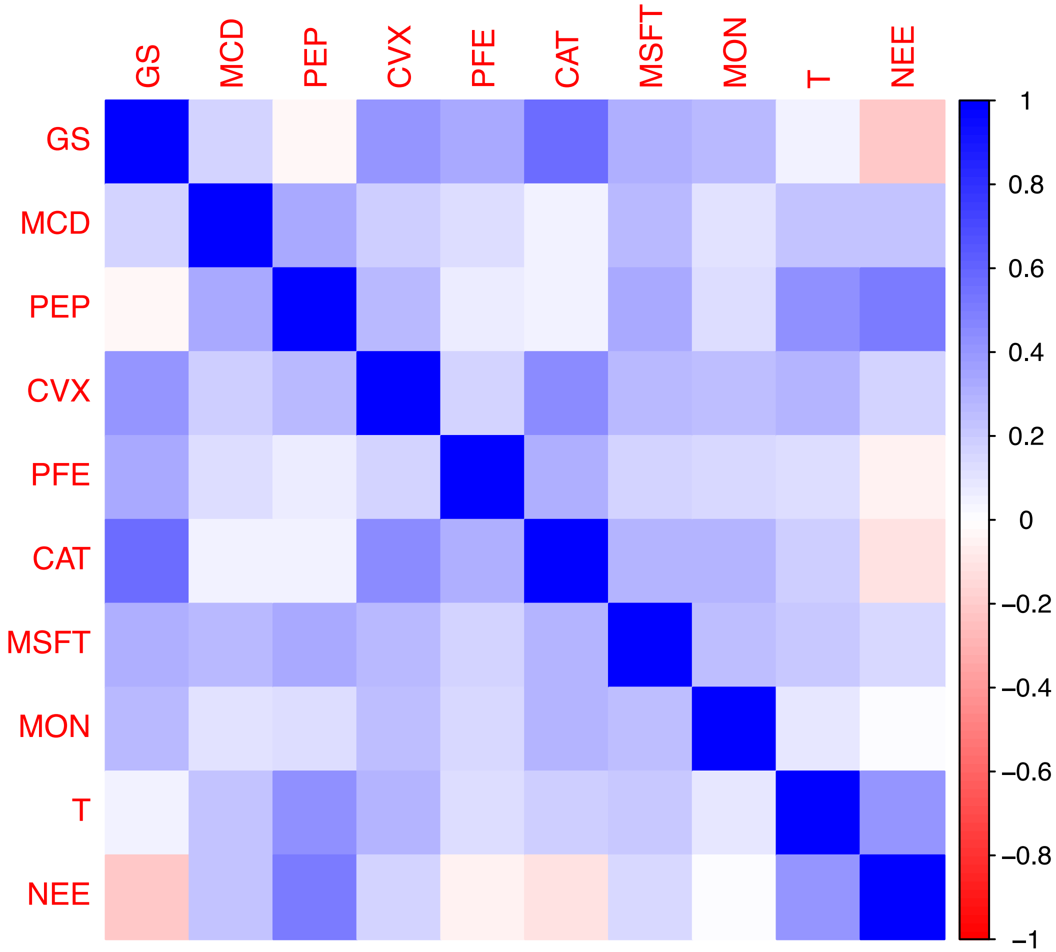


Stocks from Different Sectors: Time Series Plots



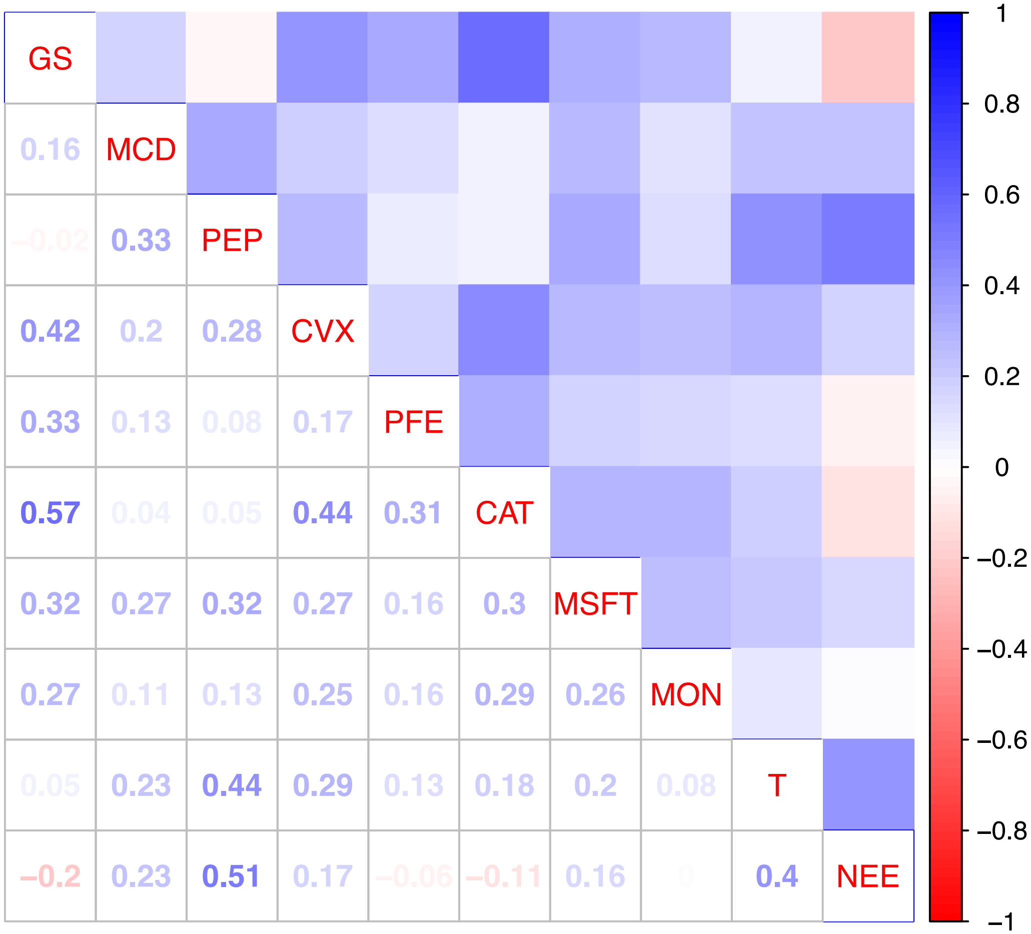
Stocks from Different Sectors: Heat map

Heat map of the correlation matrix of returns



Stocks from Different Sectors: Heat map

Heat map of the correlation matrix of returns (an alternative way to plot)



Stocks from Different Sectors: PCA

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.71	1.40	0.95	0.94	0.88	0.83	0.76	0.67	0.66	0.62
Proportion of Variance	0.29	0.20	0.09	0.09	0.08	0.07	0.06	0.05	0.04	0.04
Cumulative Proportion	0.29	0.49	0.58	0.67	0.75	0.82	0.87	0.92	0.96	1.00

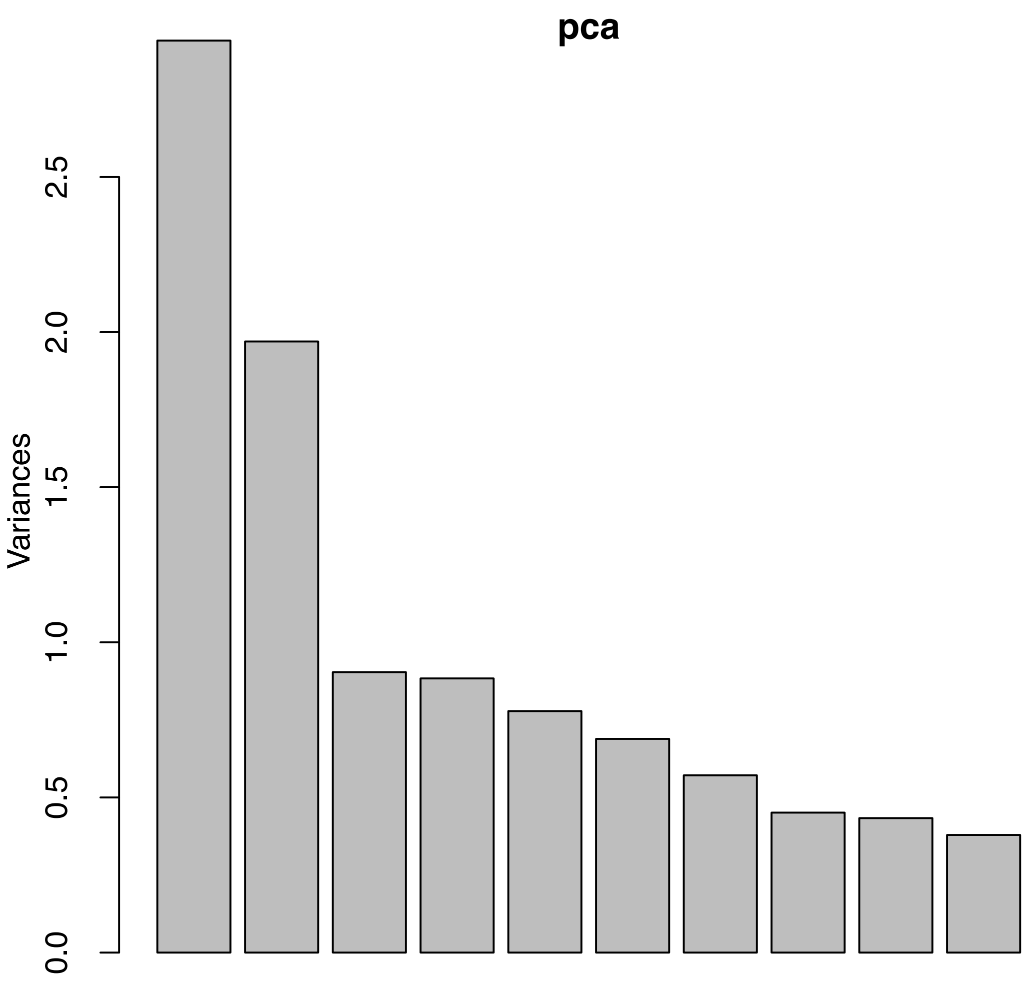
Loadings:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
1	0.35	-0.42	0.02	-0.06	0.27	-0.03	0.01	-0.13	0.20	0.75
2	0.28	0.21	0.62	-0.25	0.35	-0.47	0.17	-0.07	0.08	-0.22
3	0.32	0.43	0.02	0.03	-0.10	0.11	-0.23	0.68	0.38	0.16
4	0.40	-0.06	-0.36	0.15	0.30	-0.32	-0.43	0.13	-0.52	-0.15
5	0.25	-0.22	0.08	-0.70	-0.53	0.06	-0.27	-0.04	-0.15	-0.06
6	0.37	-0.36	-0.31	0.03	0.08	0.12	0.15	-0.08	0.55	-0.54
7	0.37	0.01	0.39	0.20	0.13	0.73	0.04	-0.08	-0.33	-0.10
8	0.27	-0.16	0.24	0.58	-0.61	-0.33	0.11	-0.05	-0.03	0.03
9	0.32	0.32	-0.39	-0.18	-0.11	-0.02	0.72	0.01	-0.25	0.13
10	0.18	0.54	-0.16	0.06	-0.09	0.03	-0.33	-0.69	0.21	0.08

Interpretation: 1st PC represents the general movement of the market and 2nd PC represents difference of industrial sectors.

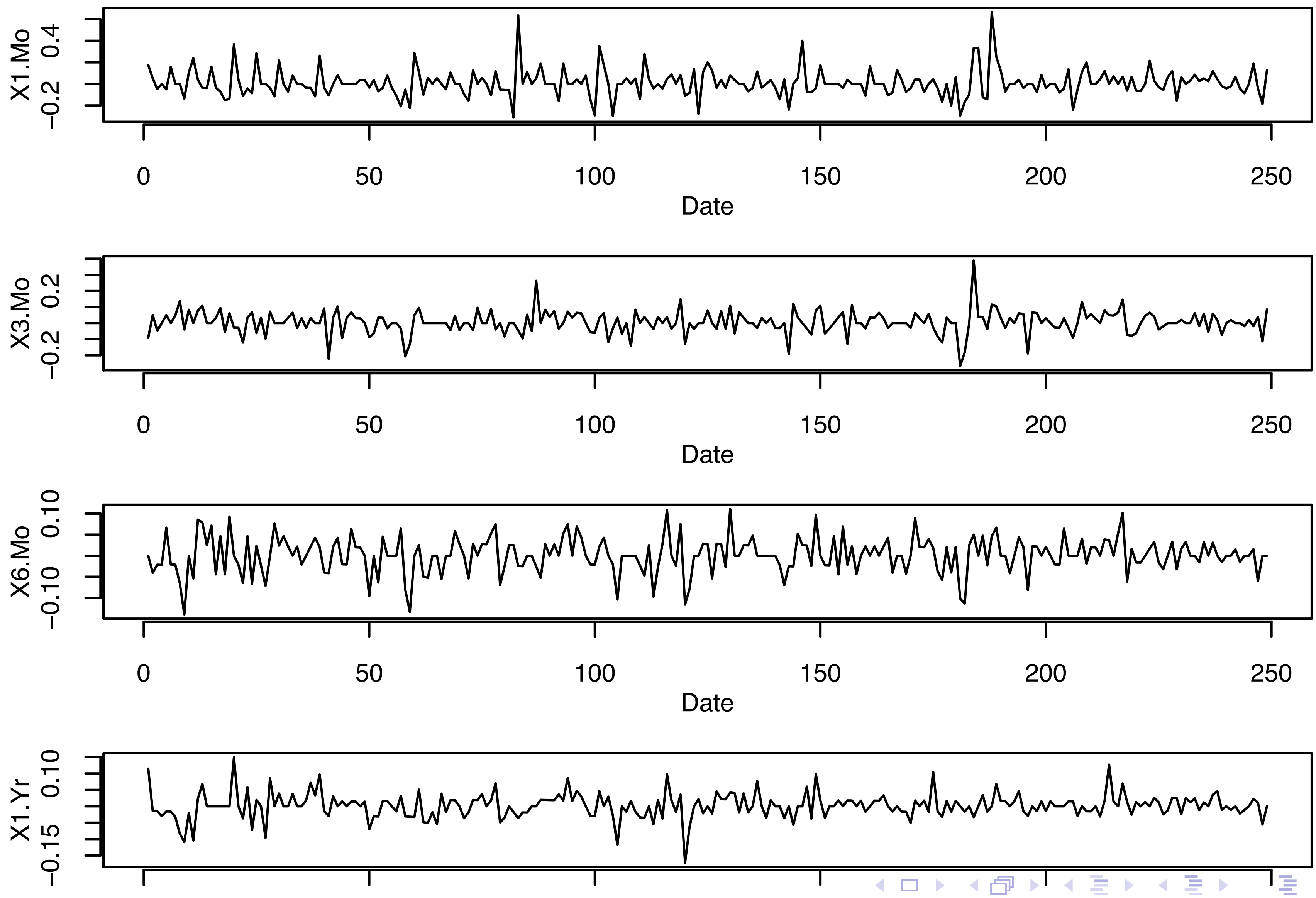
Stocks from Different Sectors: PCA

Variance of each principal component:

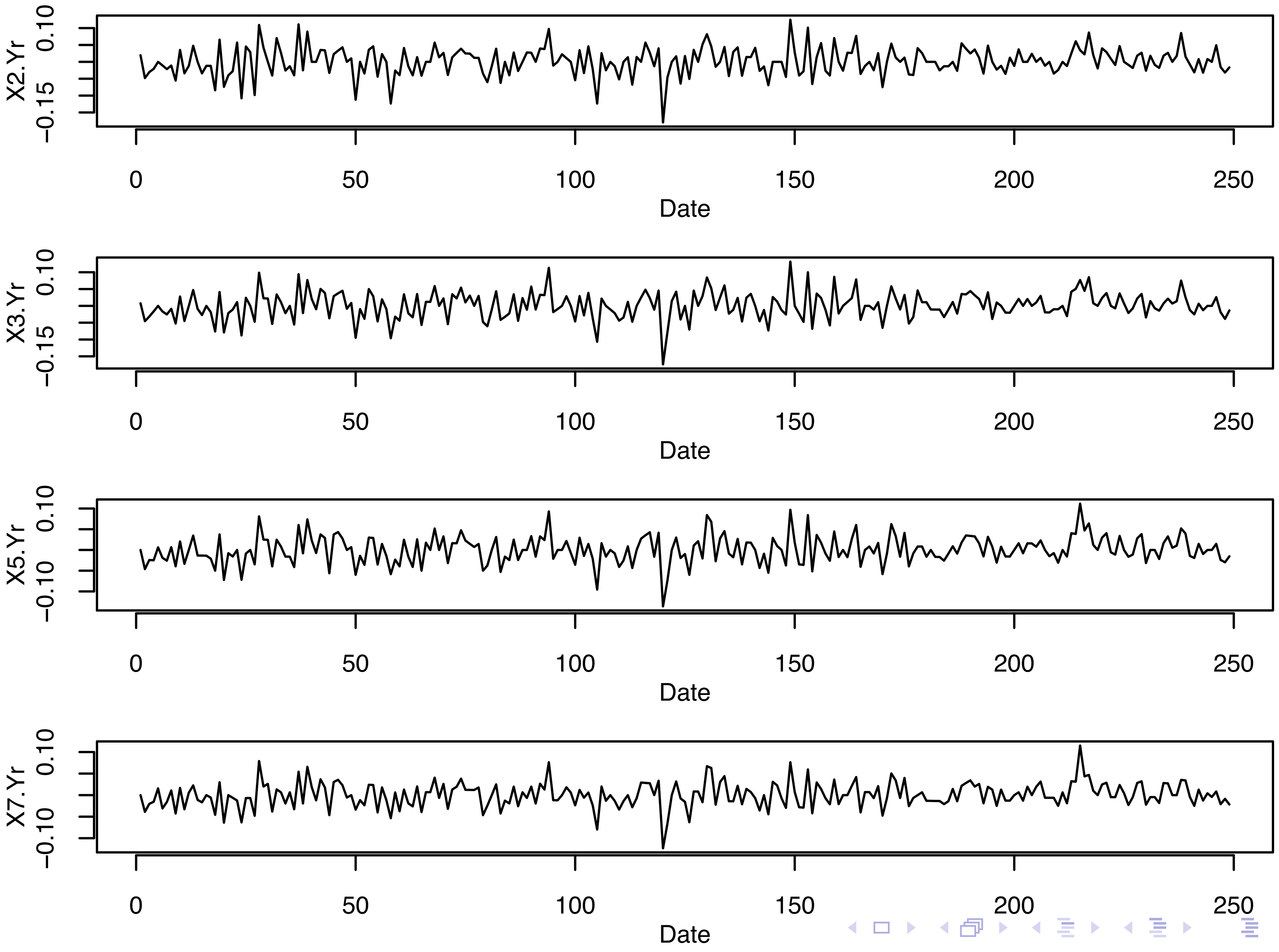


Returns of Constant Maturity Treasury (CMT) Rates

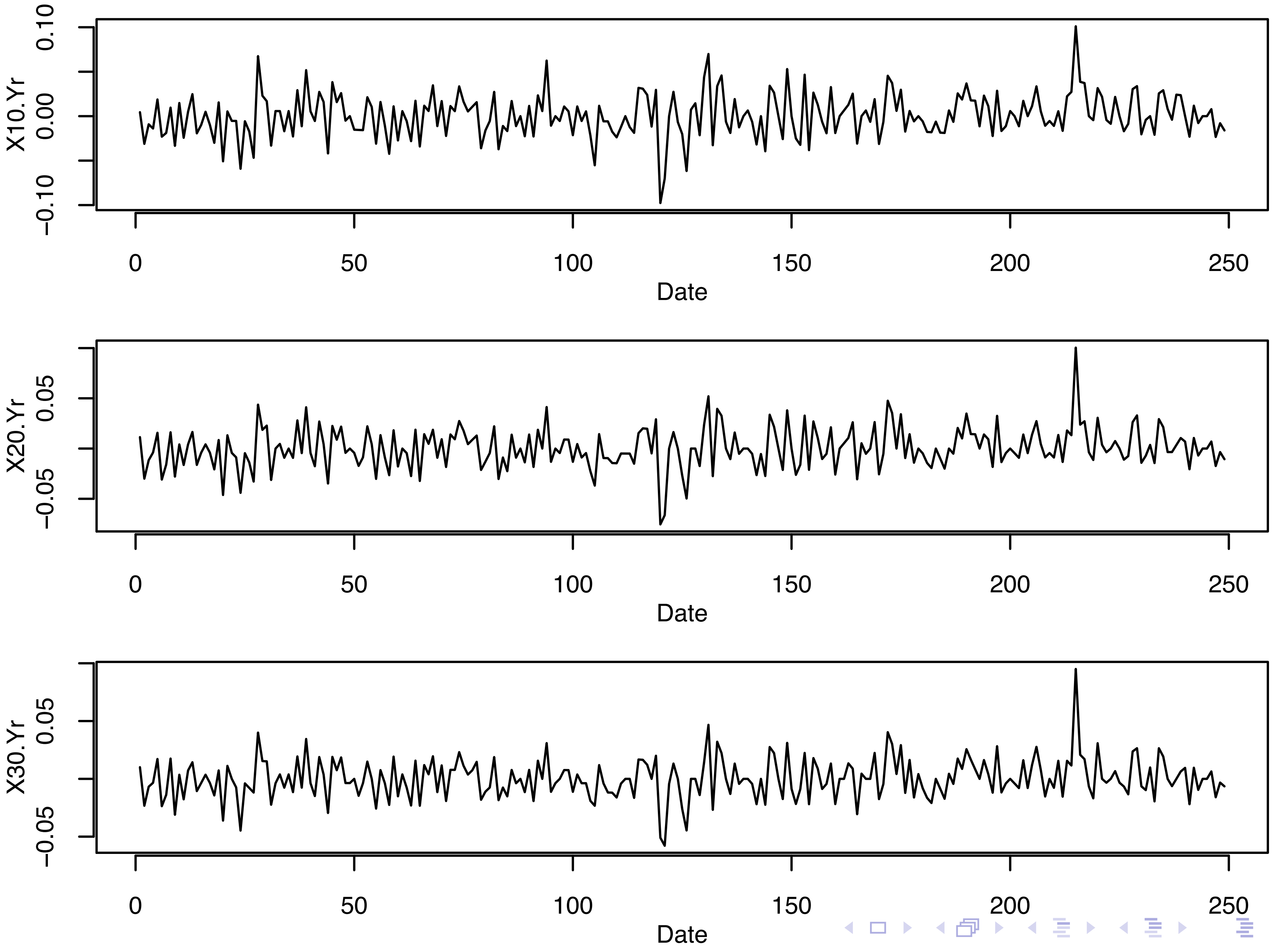
Date: 4 Jan 2016 - 30 Dec 2016



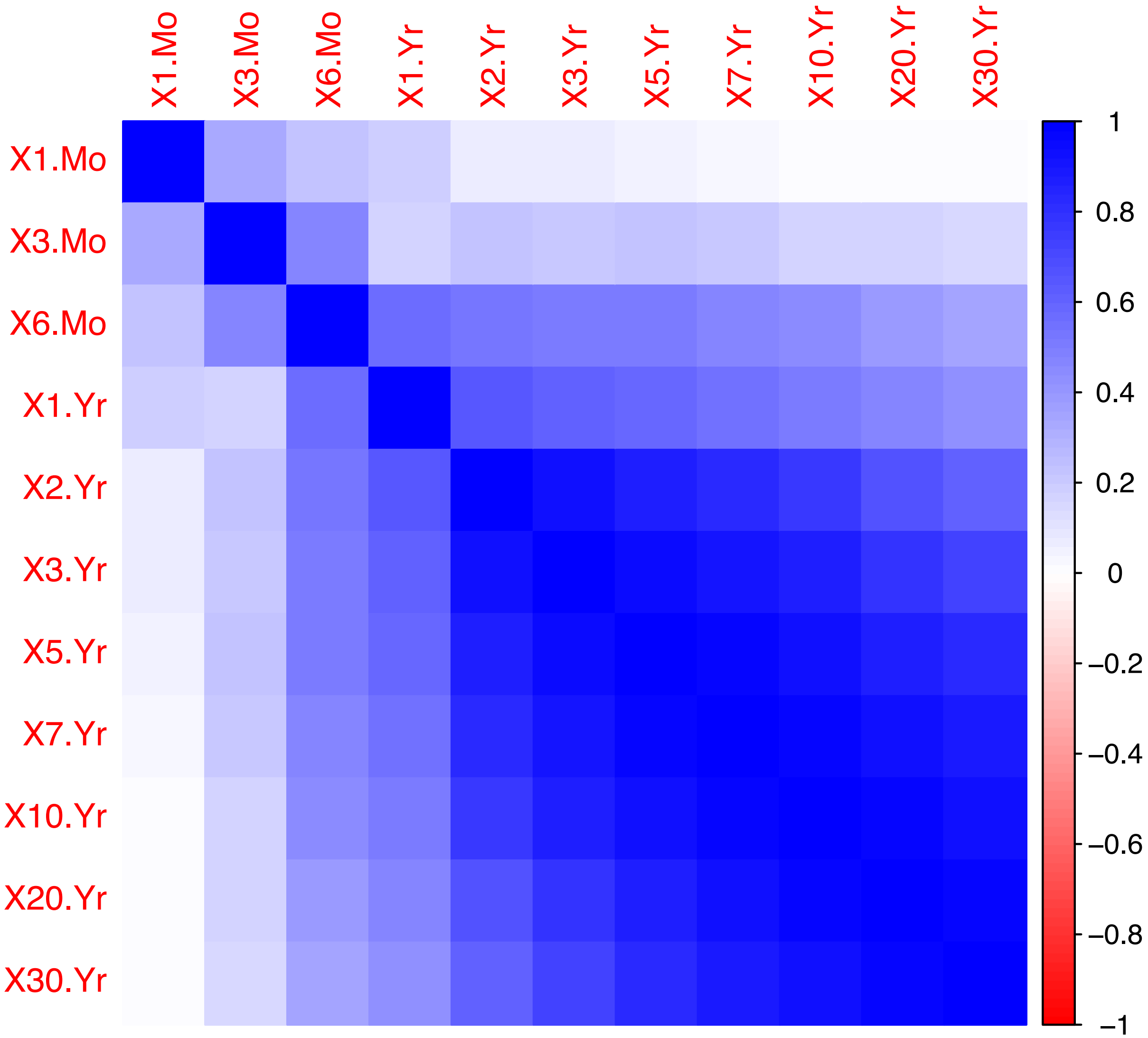
Returns of Treasury CMT Rates: Time Series Plots



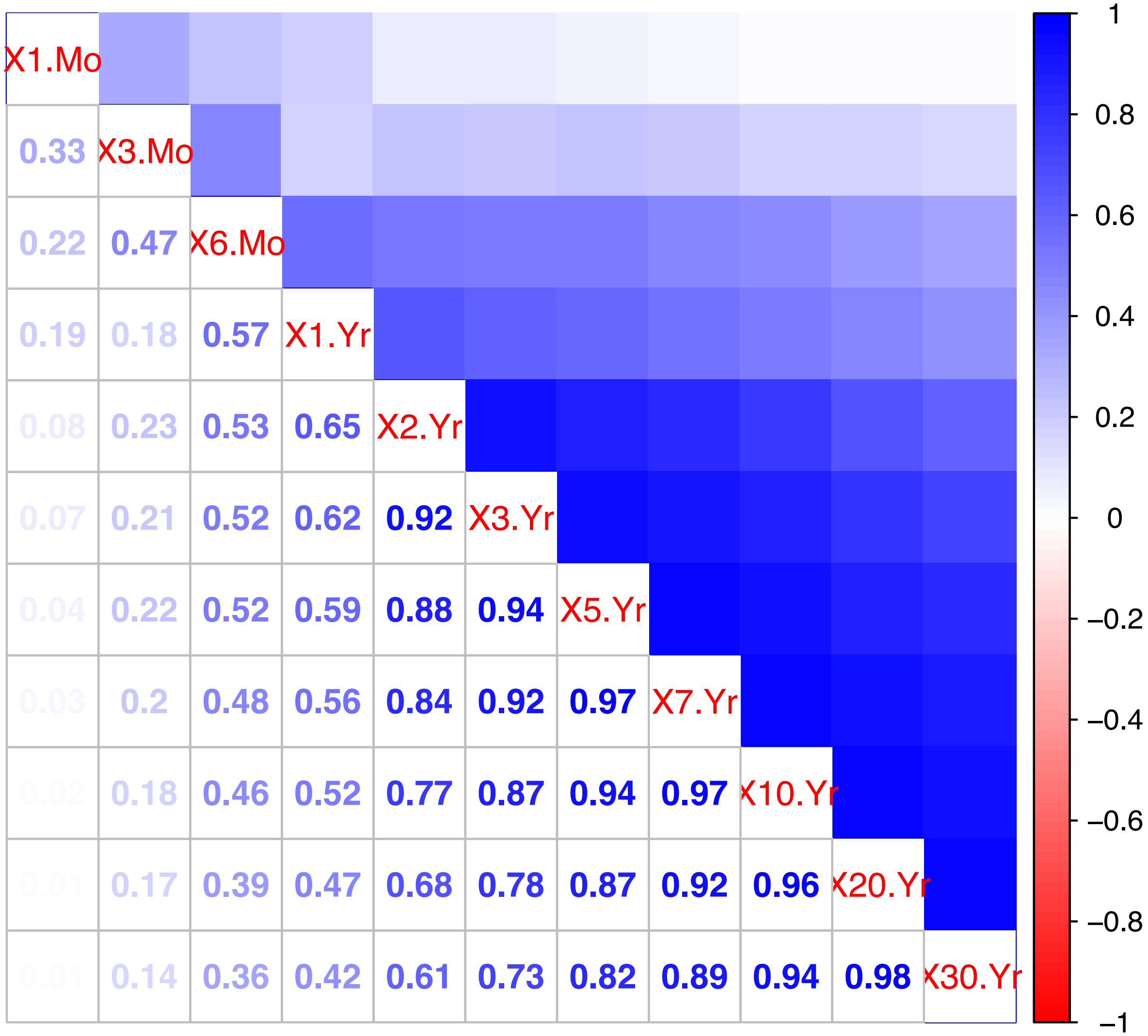
Returns of Treasury CMT Rates: Time Series Plots



Returns of Treasury CMT Rates: Heat map



Returns of Treasury CMT Rates: Heat map



X1.Mo										
0.33	X3.Mo									
0.22	0.47	X6.Mo								
0.19	0.18	0.57	X1.Yr							
0.08	0.23	0.53	0.65	X2.Yr						
0.07	0.21	0.52	0.62	0.92	X3.Yr					
0.04	0.22	0.52	0.59	0.88	0.94	X5.Yr				
0.03	0.2	0.48	0.56	0.84	0.92	0.97	X7.Yr			
0.02	0.18	0.46	0.52	0.77	0.87	0.94	0.97	X10.Yr		
0.01	0.17	0.39	0.47	0.68	0.78	0.87	0.92	0.96	X20.Yr	
0.01	0.14	0.36	0.42	0.61	0.73	0.82	0.89	0.94	0.98	X30.Yr

Returns of Treasury CMT Rates: PCA

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	2.64	1.25	0.90	0.85	0.67	0.57	0.28	0.21	0.15	0.13	0.12
Proportion of Variance	0.63	0.14	0.07	0.07	0.04	0.03	0.01	0.00	0.00	0.00	0.00
Cumulative Proportion	0.63	0.77	0.85	0.91	0.95	0.98	0.99	1.00	1.00	1.00	1.00

Returns of Treasury CMT Rates: PCA

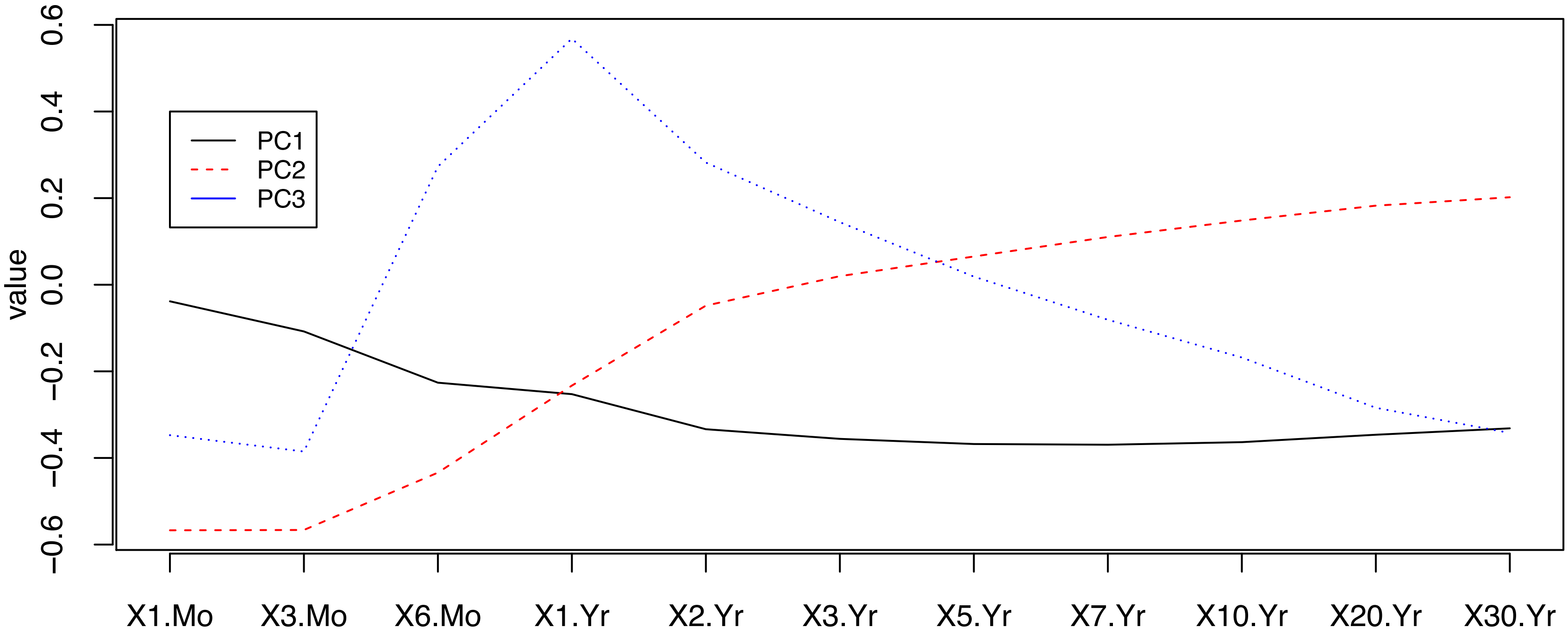
Loadings:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
X1.Mo	-0.04	-0.57	-0.35	0.73	-0.03	0.15	0.00	-0.03	0.00	-0.01	0.01
X3.Mo	-0.11	-0.57	-0.39	-0.56	-0.24	-0.39	0.02	0.02	0.01	-0.01	-0.02
X6.Mo	-0.23	-0.43	0.27	-0.32	0.49	0.59	-0.05	0.02	-0.01	0.02	0.02
X1.Yr	-0.25	-0.23	0.57	0.22	0.29	-0.65	0.08	-0.00	0.01	0.00	-0.02
X2.Yr	-0.33	-0.05	0.28	0.03	-0.50	0.10	-0.69	-0.25	0.02	0.00	-0.02
X3.Yr	-0.36	0.02	0.14	0.05	-0.37	0.14	0.22	0.80	0.03	-0.01	0.03
X5.Yr	-0.37	0.07	0.02	-0.00	-0.19	0.10	0.45	-0.38	-0.57	-0.38	-0.01
X7.Yr	-0.37	0.11	-0.08	0.01	-0.08	0.05	0.31	-0.27	0.18	0.79	0.02
X10.Yr	-0.36	0.15	-0.17	0.01	0.10	0.02	0.11	-0.15	0.71	-0.45	-0.26
X20.Yr	-0.35	0.18	-0.28	0.01	0.26	-0.11	-0.21	0.06	-0.03	-0.09	0.79
X30.Yr	-0.33	0.20	-0.34	0.03	0.34	-0.12	-0.33	0.21	-0.37	0.12	-0.55

Interpretation: 1st PC is the trend component, 2st PC is the tilt component, 3rd PC is the convexity component. (see the plot in the next slide)

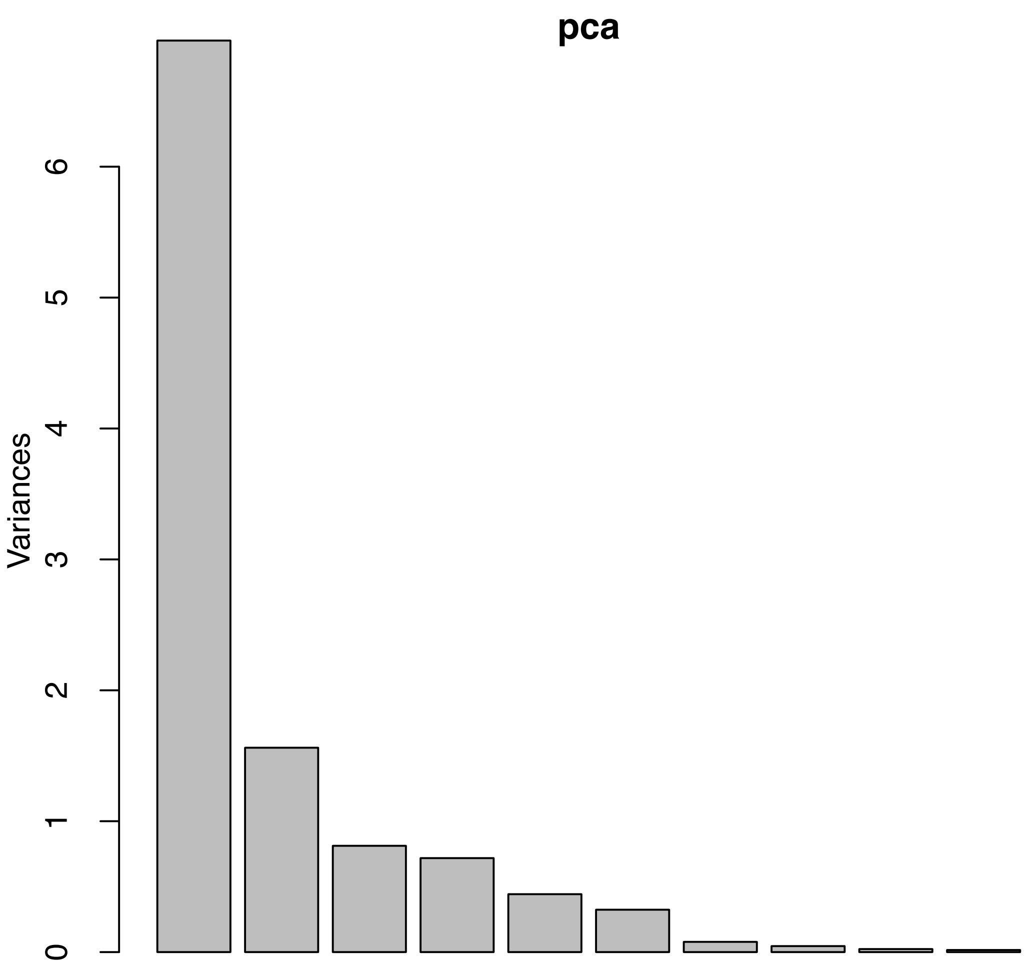
Returns of Treasury CMT Rates: PCA

Interpretation: 1st PC is the trend component, 2st PC is the tilt component, 3rd PC is the convexity component.



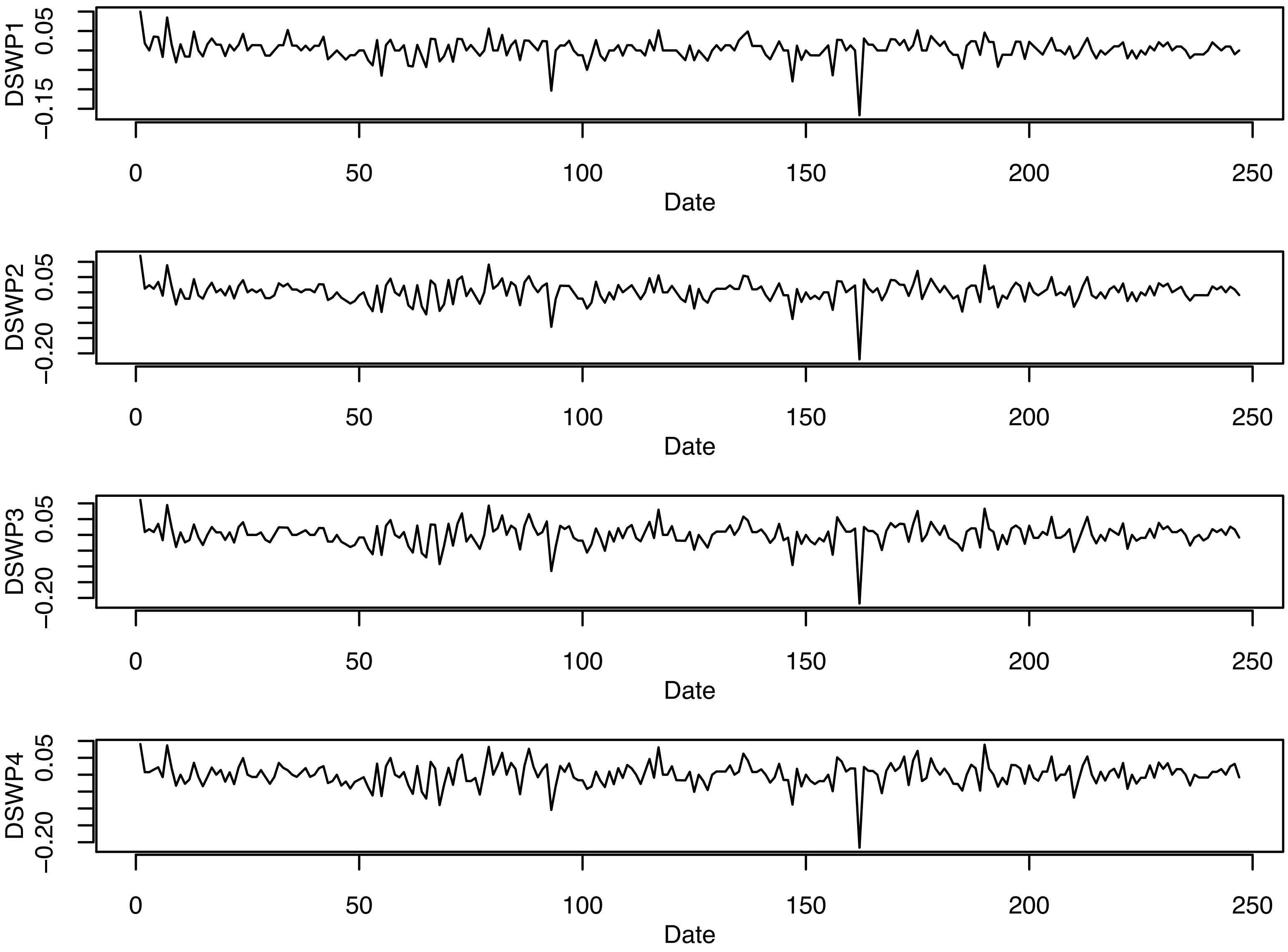
Returns of Treasury CMT Rates: PCA

Variance of each principal component:

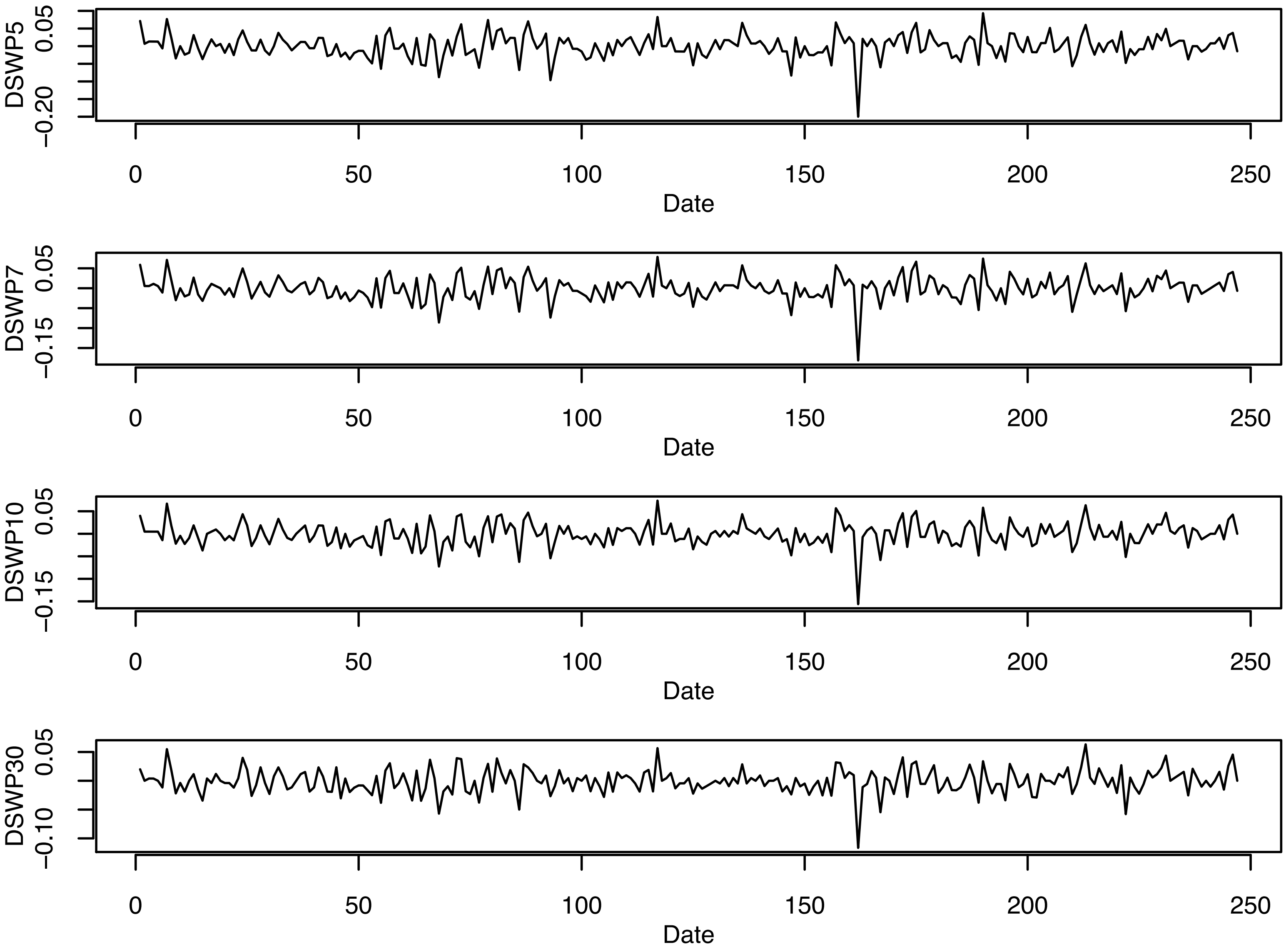


Returns of Interest Rate Swap: Time Series Plots

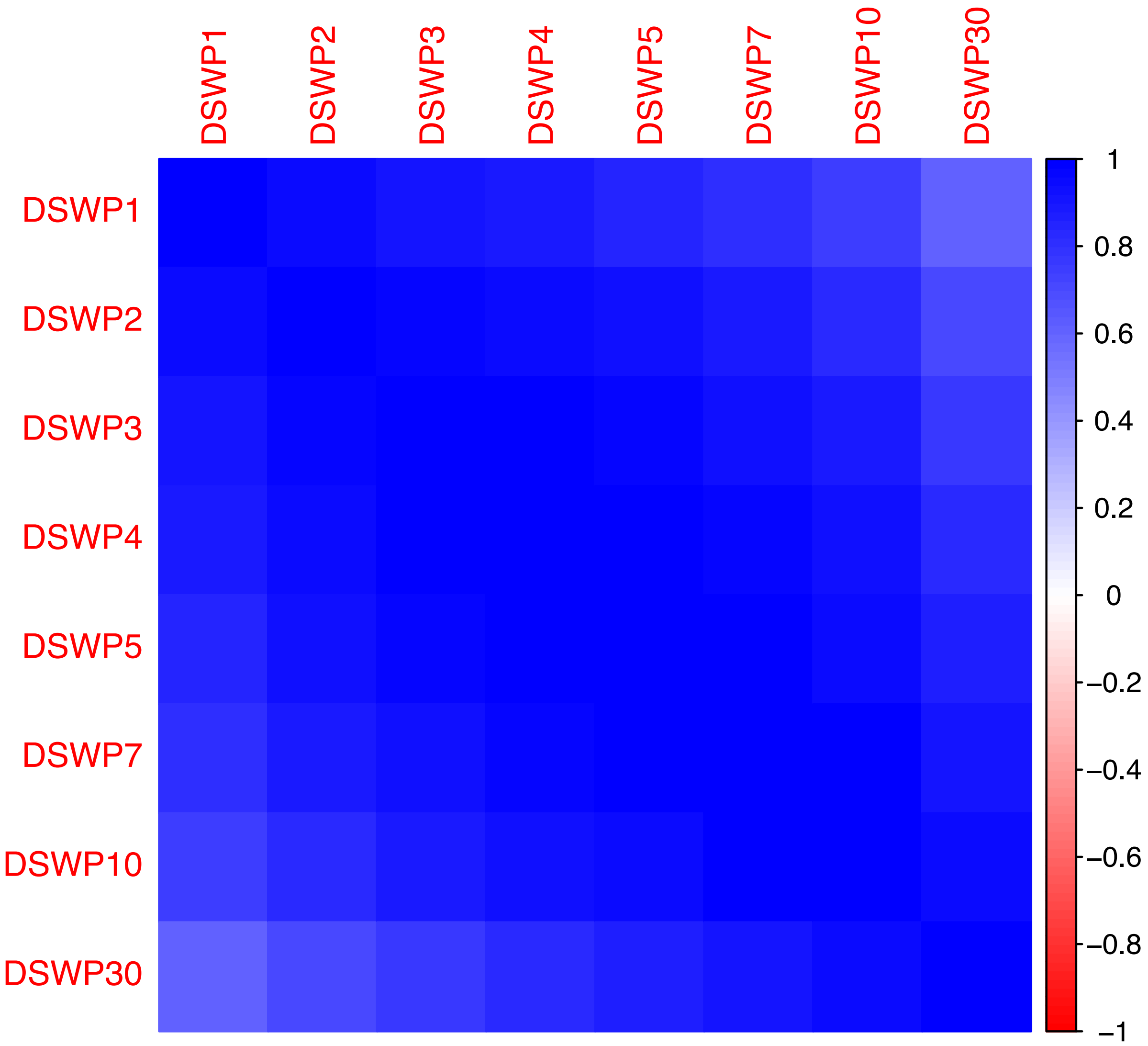
Date: Oct 28, 2015 - Oct 28, 2016



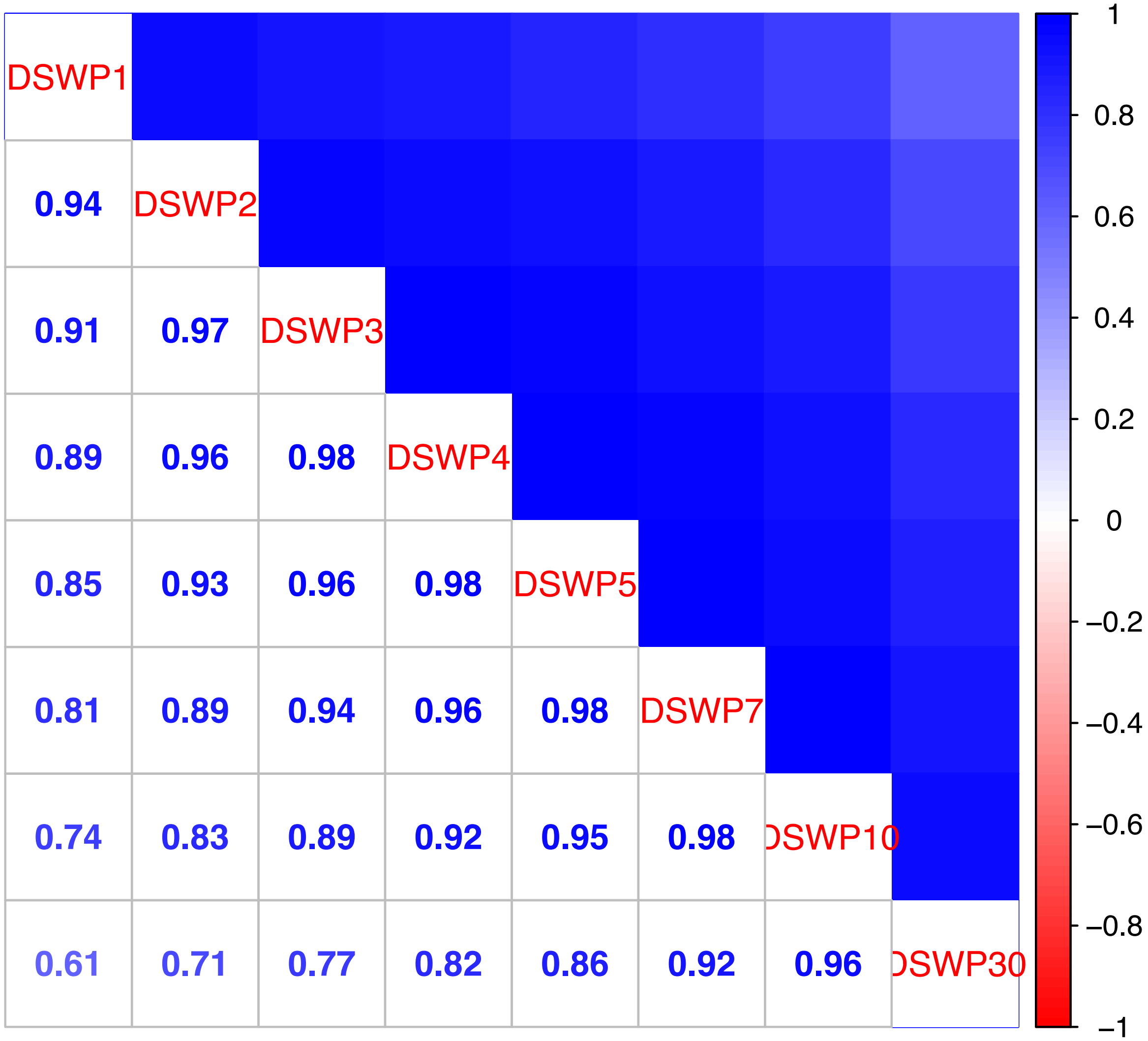
Returns of Interest Rate Swap: Time Series Plots



Returns of Interest Rate Swap: Heat map



Returns of Interest Rate Swap: Heat map



DSWP1							
0.94	DSWP2						
0.91	0.97	DSWP3					
0.89	0.96	0.98	DSWP4				
0.85	0.93	0.96	0.98	DSWP5			
0.81	0.89	0.94	0.96	0.98	DSWP7		
0.74	0.83	0.89	0.92	0.95	0.98	DSWP10	
0.61	0.71	0.77	0.82	0.86	0.92	0.96	DSWP30

Returns of Interest Rate Swap: PCA

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.69	0.76	0.30	0.18	0.13	0.12	0.11	0.10
Proportion of Variance	0.91	0.07	0.01	0.00	0.00	0.00	0.00	0.00
Cumulative Proportion	0.91	0.98	0.99	0.99	1.00	1.00	1.00	1.00

Returns of Interest Rate Swap: PCA

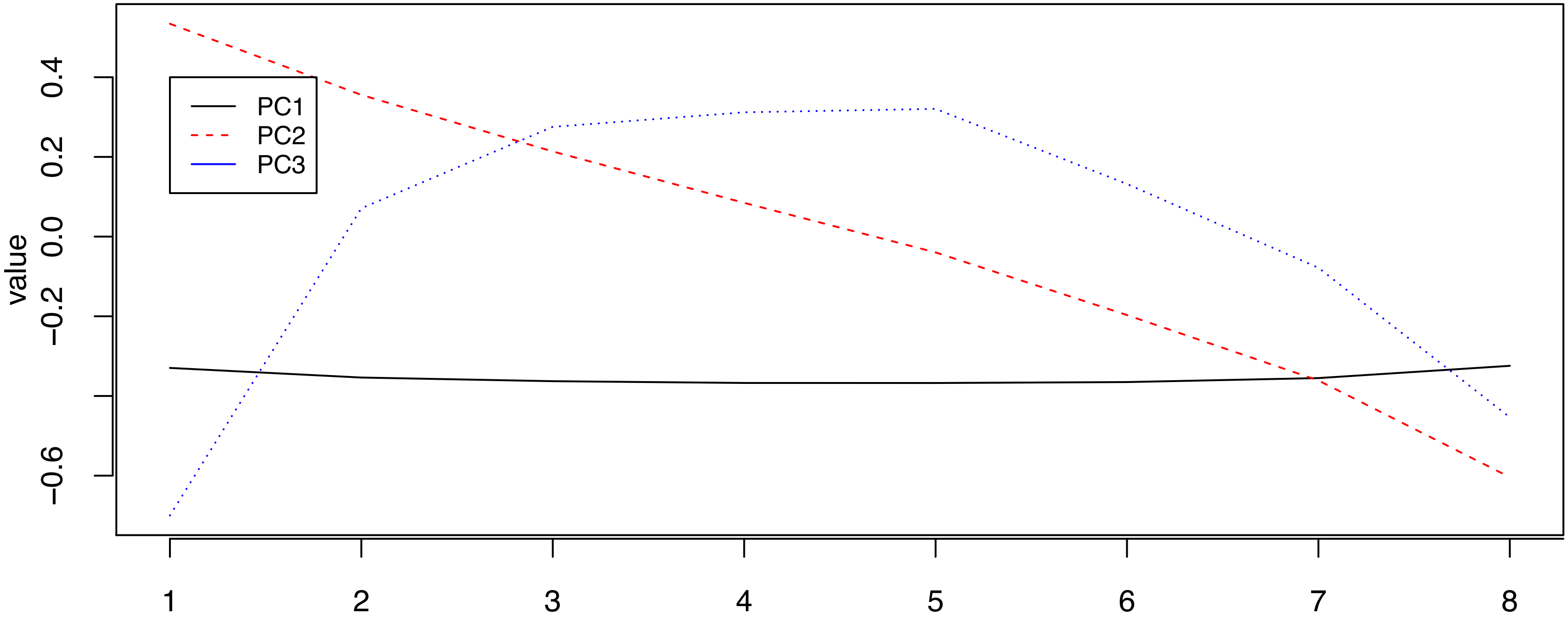
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1yr	-0.33	0.53	-0.70	0.32	0.04	-0.10	0.00	0.02
2yr	-0.35	0.36	0.07	-0.66	-0.49	0.26	0.03	-0.05
3yr	-0.36	0.21	0.28	-0.17	0.71	0.10	-0.45	0.03
4yr	-0.37	0.08	0.31	0.08	0.22	-0.19	0.82	-0.05
5yr	-0.37	-0.04	0.32	0.26	-0.37	-0.58	-0.36	-0.30
7yr	-0.37	-0.20	0.13	0.28	-0.21	0.21	-0.07	0.80
10yr	-0.36	-0.36	-0.08	0.30	-0.03	0.62	0.00	-0.51
30yr	-0.32	-0.60	-0.45	-0.44	0.14	-0.33	0.02	0.05

Interpretation: 1st PC is the trend component, 2st PC is the tilt component, 3rd PC is the convexity component. (see the plot in the next slide)

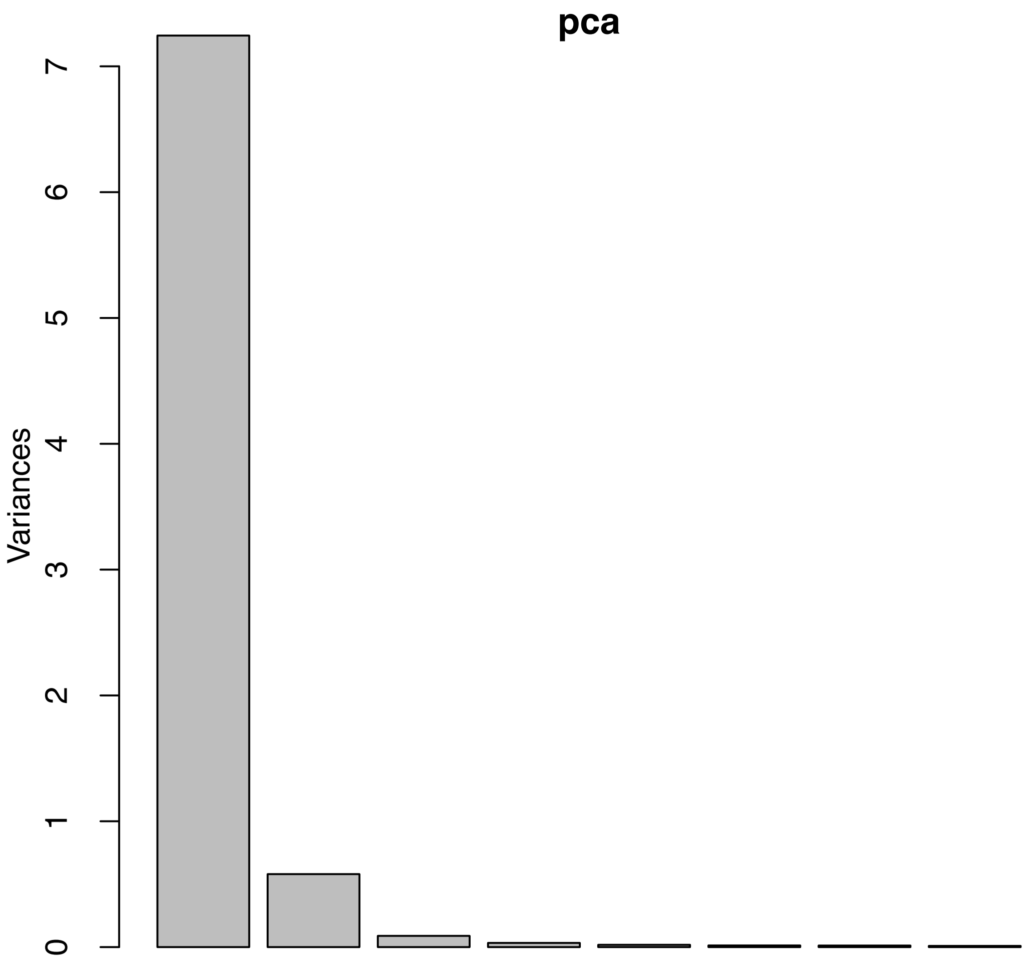
Returns of Interest Rate Swap: PCA

Interpretation: 1st PC is the trend component, 2st PC is the tilt component, 3rd PC is the convexity component.



Returns of Interest Rate Swap: PCA

Variance of each principal component:


pca
Variances