PCA for Factor Model

Christopher Ting

http://cting.x10host.com/CUHKSZ/CUHKSZ.html

http://www.mysmu.edu/faculty/christophert/
✉: christophert@smu.edu.sg

July 12, 2018

Broad Lesson Plan

Correlation Matrix

- R_i: Return of stock i, $i = 1, 2, \ldots, N$.
- Cross-sectional correlation matrix Γ:
$$\Gamma_{ij} = \text{corr}(R_i, R_j)$$
- Estimation of correlation matrix from data requires selecting a sample size, or estimation period, T.
- If the universe of assets is large, then $T \ll N$ (e.g., T=252 vs. N=1,500)
- The correlation matrix is not full rank in general since we expect that the stocks are "driven" by m factors, where $m \ll N$.

$$R_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} F_k + \epsilon_i \tag{1}$$

Two Challenges and PCA Solution

- Degeneracy: rank of $\Gamma <$ dimension N
- Model selection: What is the right number of factors and are they?

1. Perform PCA on the correlation matrix, going back for T periods (days). The analysis is on a T by N matrix.
2. Estimate the number of significant components.
3. Analyze the corresponding eigenvectors and eigenportfolios (factors).
4. Associate the factors to features of the market (e.g. sectors, market cap, etc).

Standardized Returns

- Historical share-price data on a cross-section of N stocks going back $M+1$ days in history. Accordingly, we have simple return R_{ik}.
  - Stock $i = 1, 2, \ldots, N$
  - Day $k = 1, 2, \ldots, M$

- Since some stocks are more volatile than others, it is convenient to work with standardized returns

$$Y_{ik} = \frac{R_{ik} - \widehat{R}_i}{\widehat{\sigma}_i},$$

where $\widehat{R}_i$ is the sample average and $\widehat{\sigma}_i^2$ is the unbiased sample variance of Stock i.

Empirical Correlation Matrix

- The empirical correlation matrix of the data is defined by

$$\rho_{ij} = \frac{1}{M-1}\sum_{k=1}^{M} Y_{ik}Y_{jk},$$

  which is symmetric and non-negative definite.

- Notice that, for any index $i = 1, 2, \ldots, N$, we have

$$\rho_{ii} = \frac{1}{M-1}\sum_{k=1}^{M}(Y_{ik})^2 = \frac{1}{M-1}\frac{\sum_{k=1}^{M}(R_{ik} - \widehat{R}_i)^2}{\widehat{\sigma}_i^2} = 1$$

- Estimation window for the correlation matrix is a year (e.g., 252 trading days) prior to the trading date t_0.

PCA of Correlation Matrix

- Rank the eigenvalues of the empirical correlation matrix in decreasing order

$$N \geqq \lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_N \geqq 0.$$

- The corresponding eigenvectors are denoted as

$$\boldsymbol{v}^{(j)} = \begin{pmatrix} v_1^{(1)} & v_2^{(2)} & \cdots & v_N^{(N)} \end{pmatrix}$$

- Explained variance by first m eigenvectors $\frac{1}{N}\sum_{k=1}^{m} \lambda_k$.

- Definition: Density of States

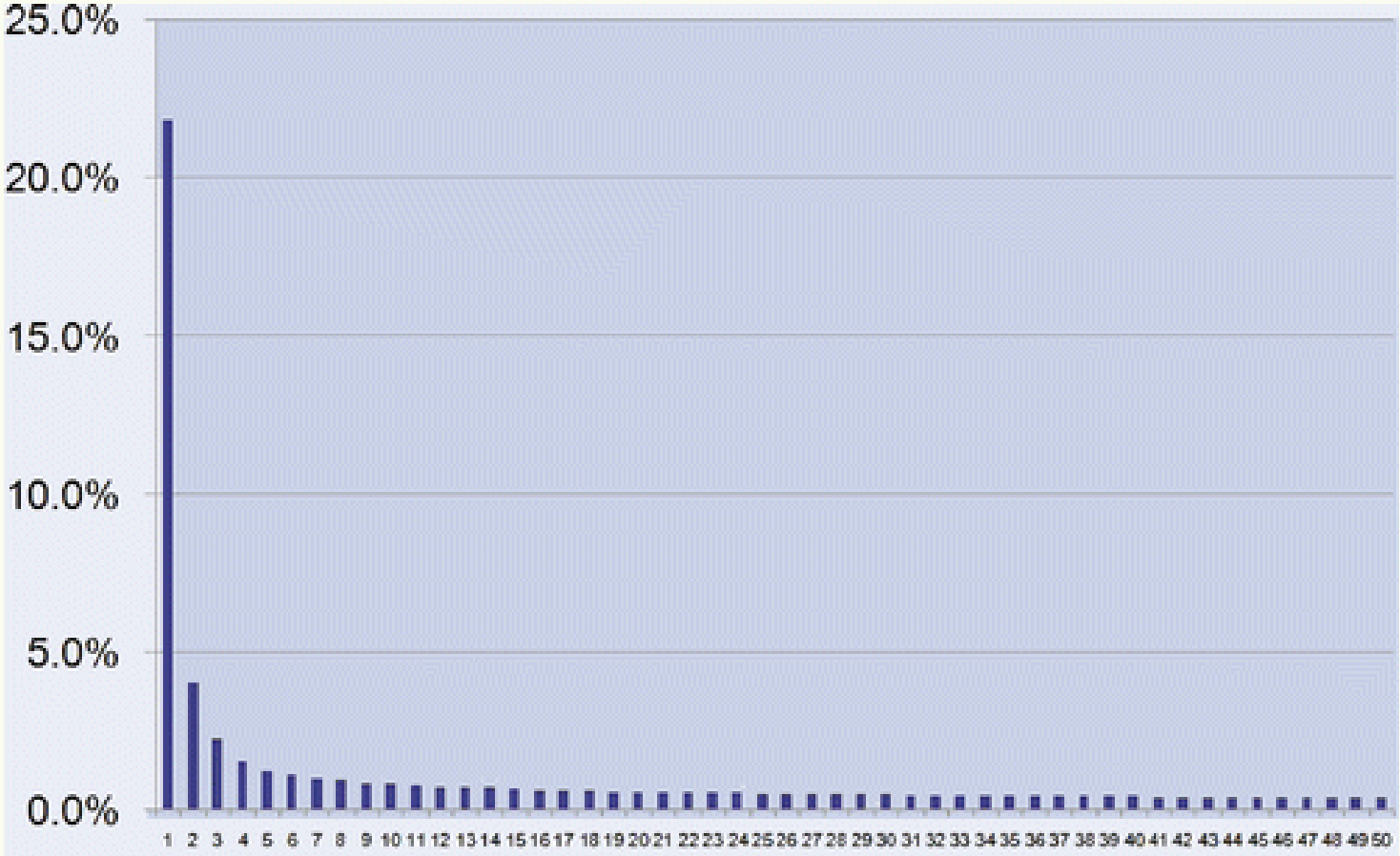
$$D(x, y) := \frac{\text{Number of eigenvalues between } x \text{ and } y}{N}$$

Data (January, 2007)

			Market cap unit: 1M/usd		
Sector	ETF	Num of stocks	Average	Max	Min
Internet	HHH	22	10,350	104,500	1,047
Real Estate	IYR	87	4,789	47,030	1,059
Transportation	IYT	46	4,575	49,910	1,089
Oil Exploration	OIH	42	7,059	71,660	1,010
Regional Banks	RKH	69	23,080	271,500	1,037
Retail	RTH	60	13,290	198,200	1,022
Semiconductors	SMH	55	7,303	117,300	1,033
Utility	UTH	75	7,320	41,890	1,049
Energy	XLE	75	17,800	432,200	1,035
Financial	XLF	210	9,960	187,600	1,000
Industrial	XU	141	10,770	391,400	1,034
Technology	XLK	158	12,750	293,500	1,008
Consumer Staples	XLP	61	17,730	204,500	1,016
Healthcare	XIV	109	14,390	192,500	1,025
Consumer discretionary	XLY	207	8,204	104,500	1,007
Total		1417	11,291	432,200	1,000

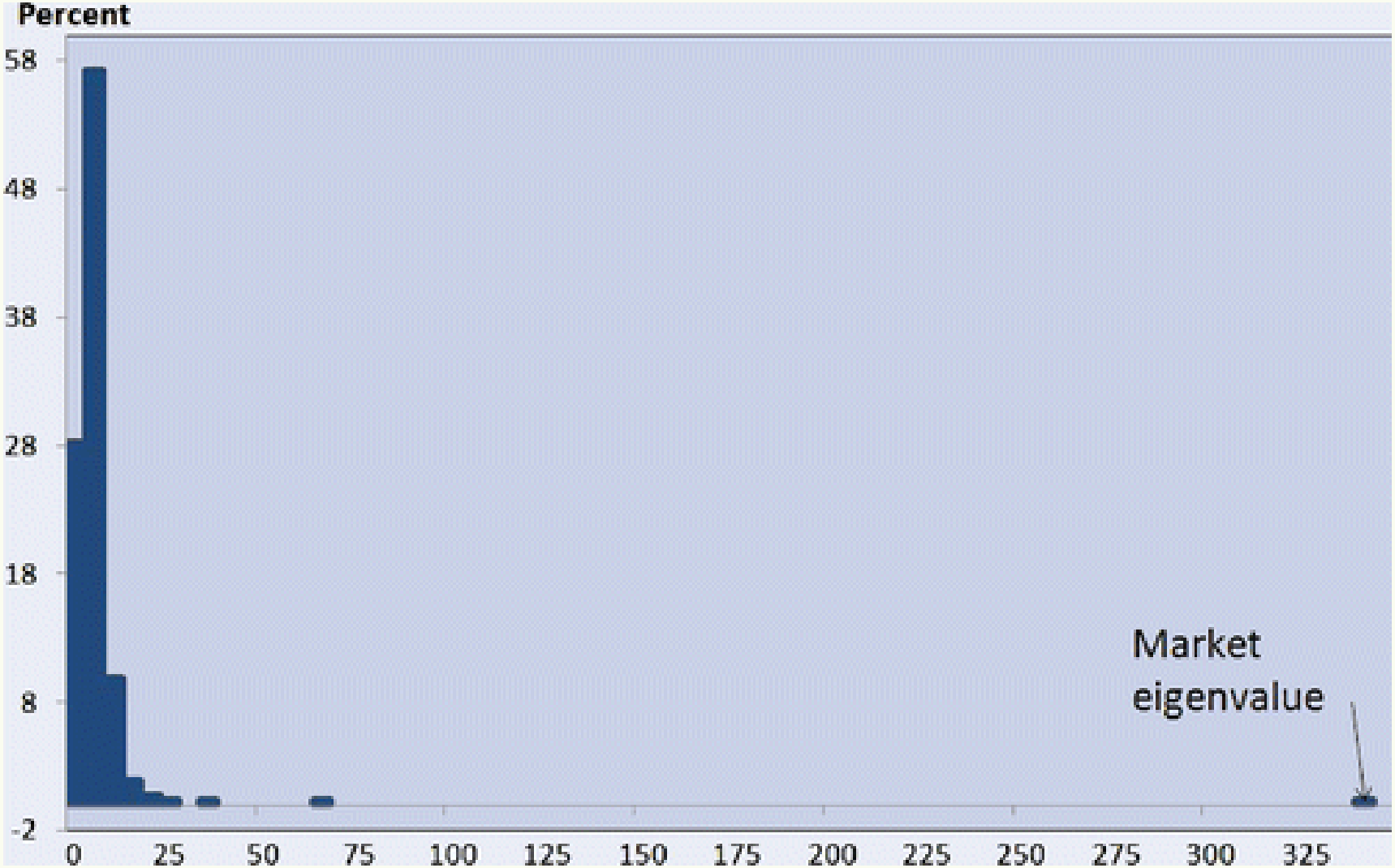
Top 50 Eigenvalues

- Computed on 1 May 1, 2007, using a one-year window and a universe of 1,417 stocks
- Eigenvalues are measured as percentage of explained variance.



Density of States

- 'Detached eigenvalues' (signal) versus 'bulk spectrum' (noise)



Eigenportfolios as Factors

- Let $\lambda_1, \lambda_2, \ldots, \lambda_m; m < N$ be the significant detached eigenvalues.
- For each j, we consider the corresponding "**eigenportfolio**", which is such that the amount invested in each of the stocks is given by

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\widehat{\sigma}_i}.$$

- The eigenportfolio returns are therefore

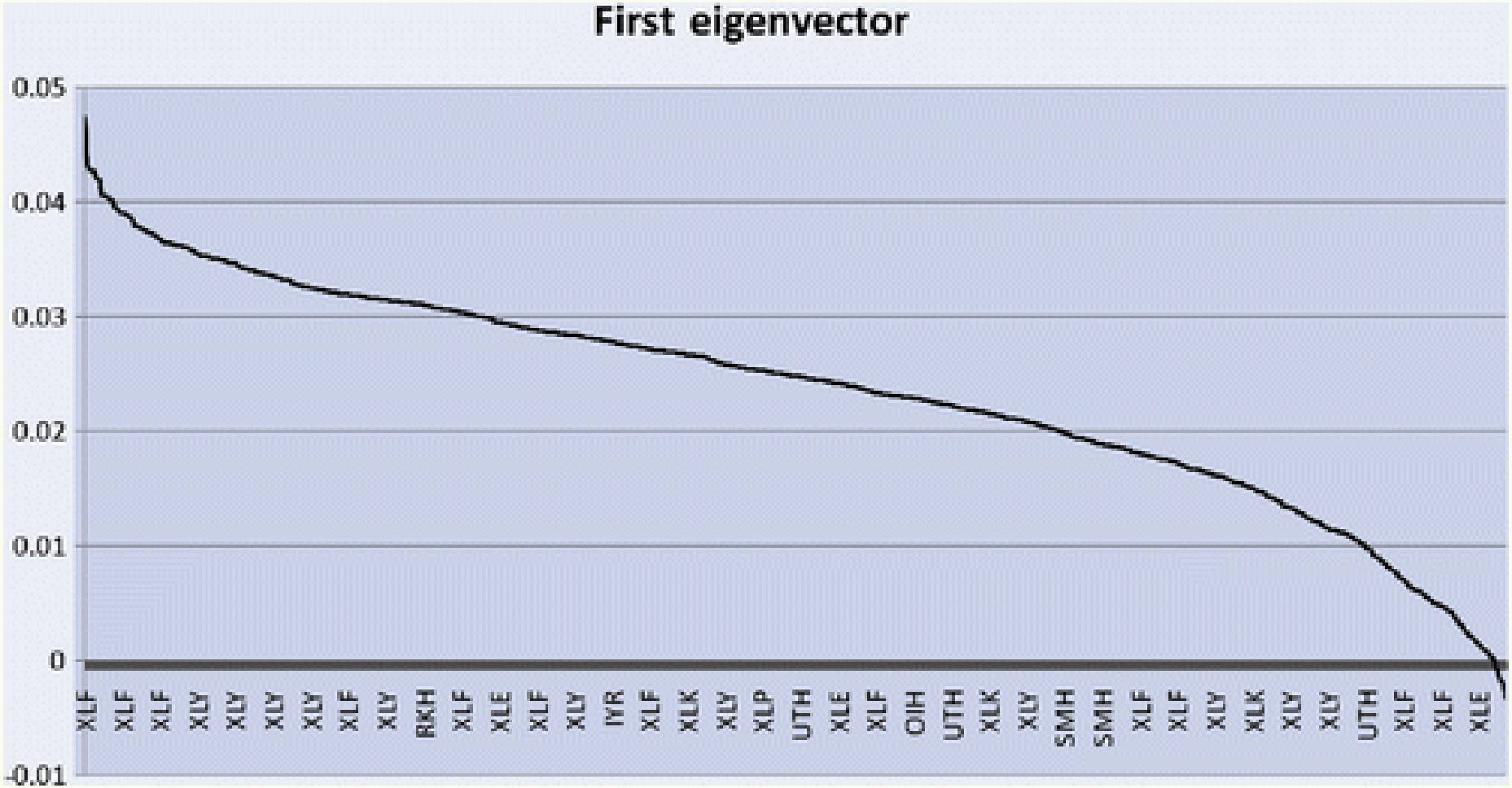
$$F_{jk} = \sum_{i=1}^{N} \frac{v_i^{(j)}}{\widehat{\sigma}_i} R_{ik}, \quad j = 1, 2, \ldots, m$$

Factor Model

- Define $F_k := \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} \left(\frac{v_i^{(k)}}{\sigma_i} \right) R_i$
- Then we can use F_k, $k = 1, 2, \ldots, m$ corresponding to the top m eignevalues as factors

$$R_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} F_k + \epsilon_i$$

Principal Eigenportfolio's Coefficient Order

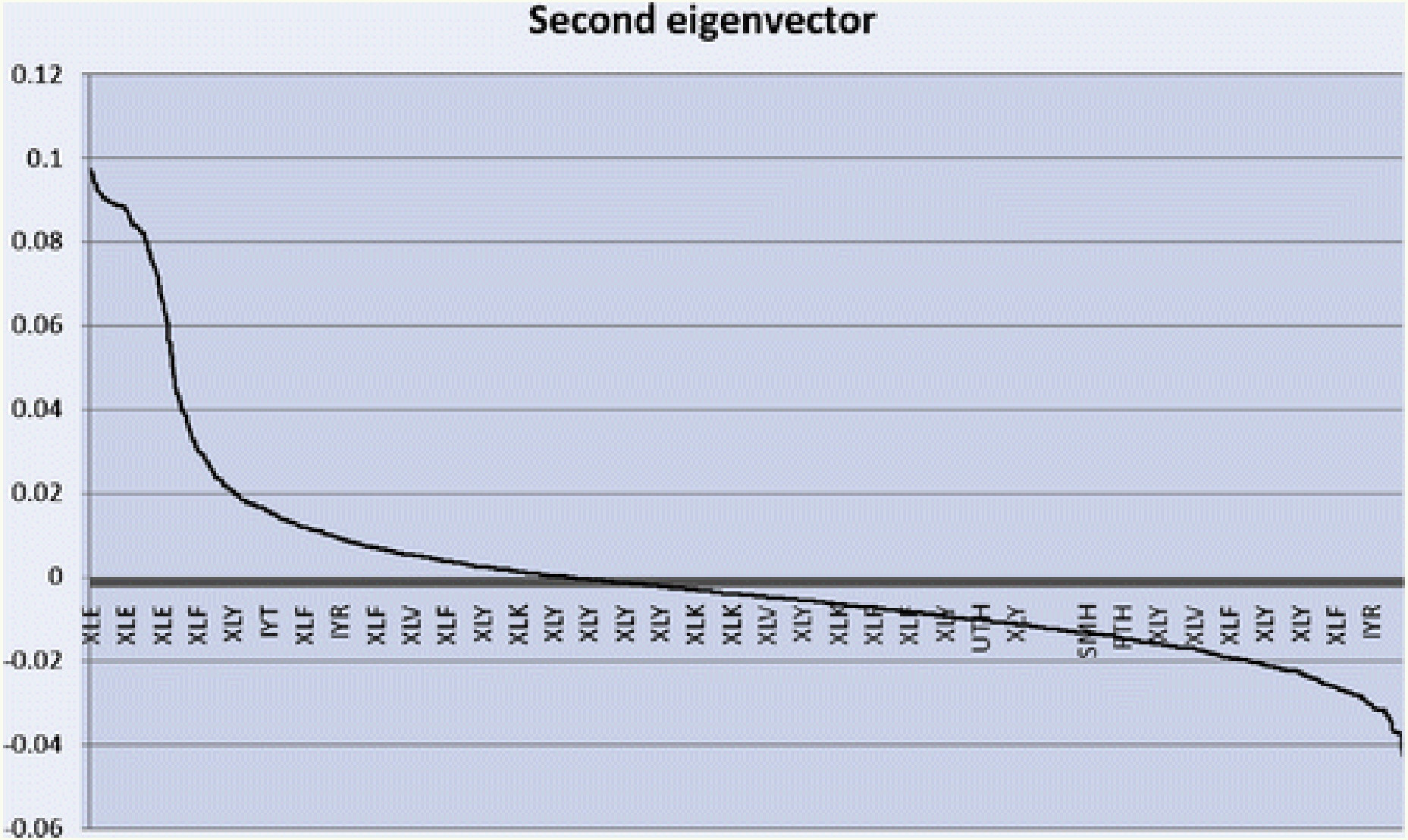


Coherence

- Definition
  If an eigenvector is such that stocks with a given property (size, industry sector) have entries with the same sign, then the eigenvector is said to be **coherent** (with respect to the given property).

- Conjecture: The significant eigenvectors are coherent with respect to either size of sector

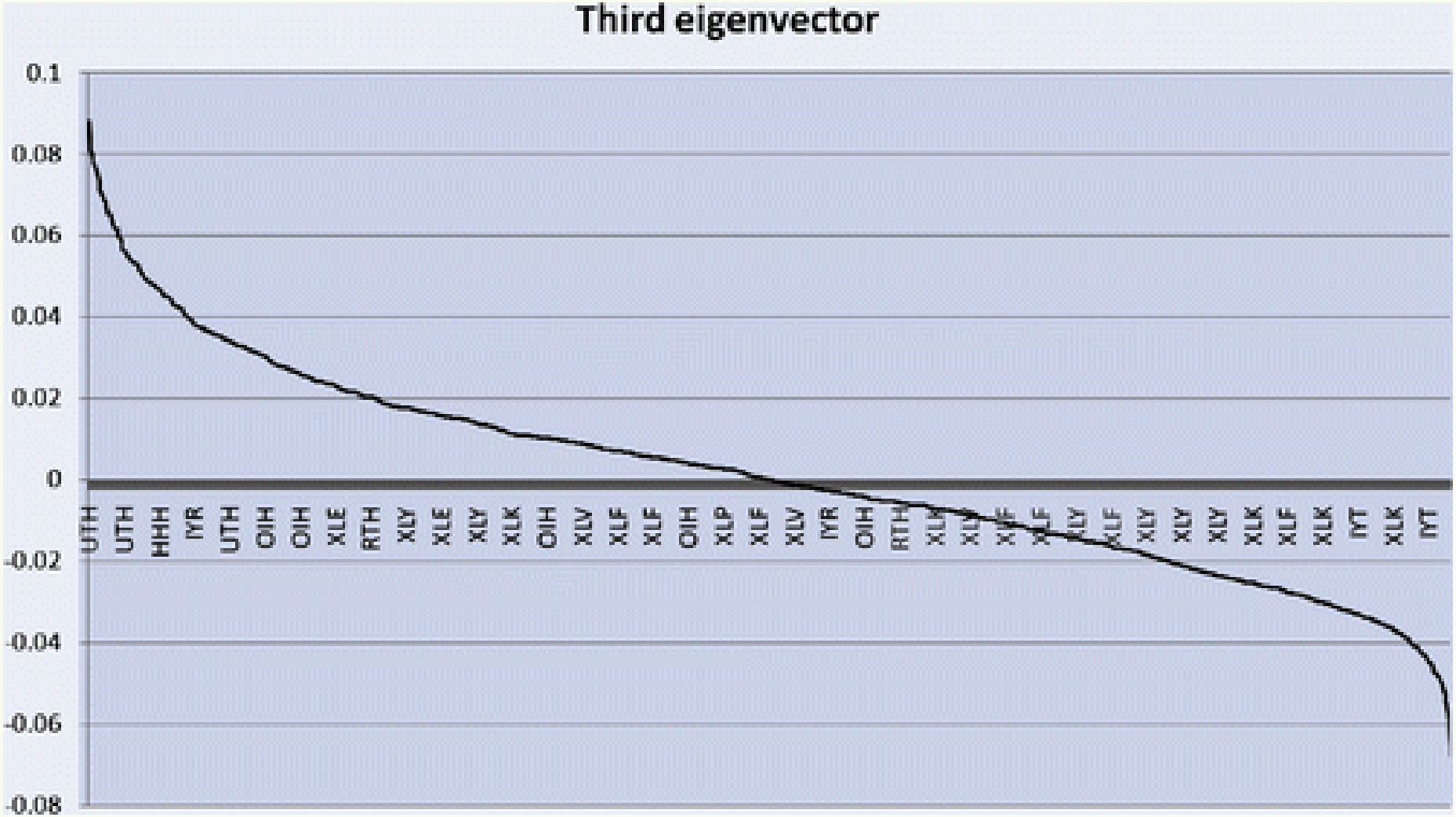
Second Eigenportfolio's Coefficient Order



Top and Bottom Ten in Second Eigenvector

Top 10 stocks	Bottom 10 stocks
energy, oil and gas	Real estate, financials, airlines
Suncor Energy Inc.	American Airlines
Quicksilver Res.	United Airlines
XTO Energy	Marshall & Isley
Unit Corp.	Fifth Third Bancorp
Range Resources	BBT Corp.
Apache Corp.	Continental Airlines
Schlumberger	M & T Bank
Denbury Resources Inc.	Colgate-Palmolive Company
Marathon Oil Corp.	Target Corporation
Cabot Oil & Gas Corporation	Alaska Air Group, Inc.

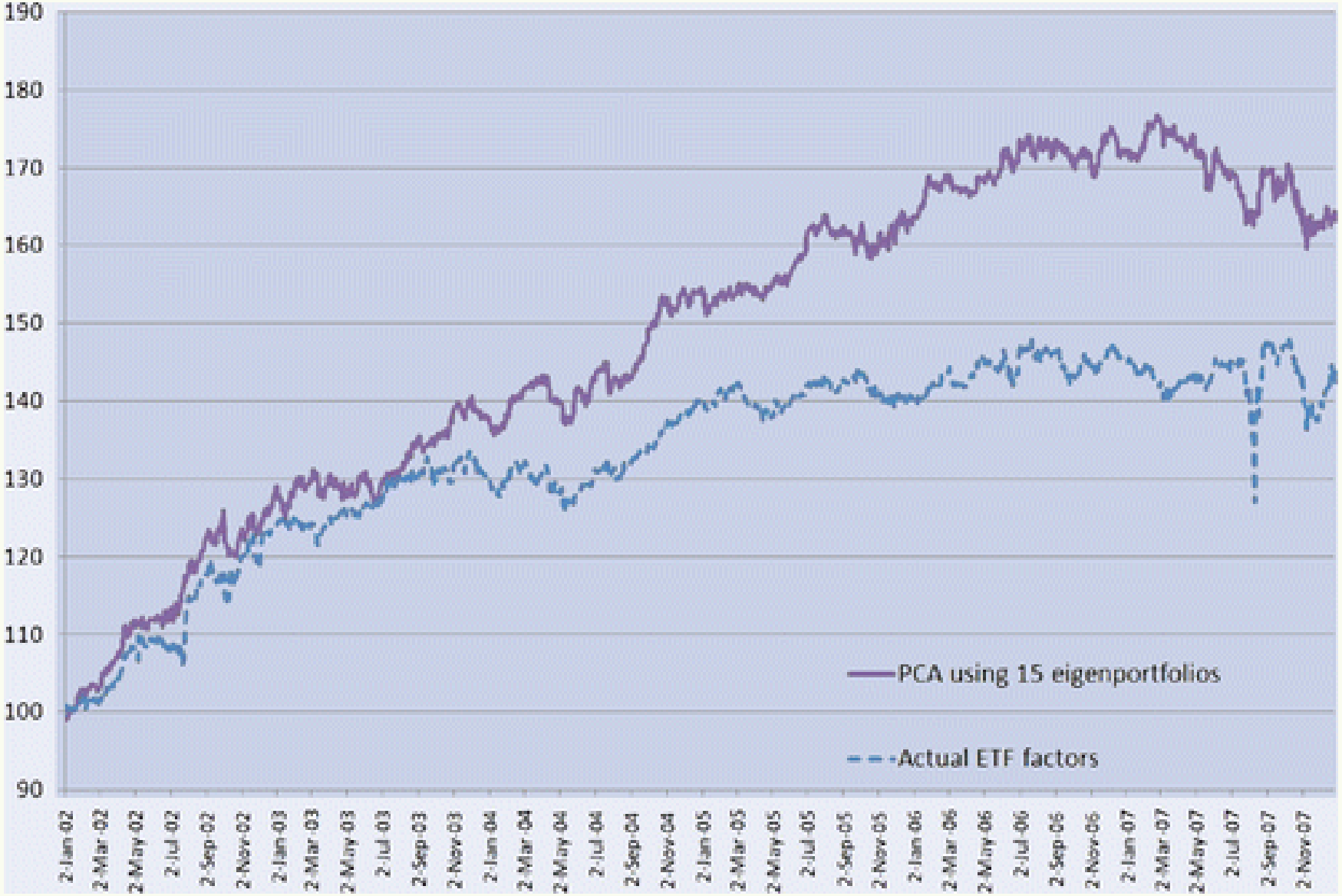
Third Eigenportfolio's Coefficient Order



Top and Bottom Ten in Third Eigenvector

Top 10 stocks	Bottom 10 stocks
Utility	Semiconductor
Energy Corp.	Arkansas Best Corp.
FPL Group, Inc.	National Semiconductor Corp.
DTE Energy Company	Lam Research Corp.
Pinnacle West Capital Corp.	Cymer, Inc.
The Southern Company	Intersil Corp.
Consolidated Edison, Inc.	KLA-Tencor Corp.
Allegheny Energy, Inc.	Fairchild Semiconductor International
Progress Energy, Inc.	Broadcom Corp.
PG&E Corporation	Cellcom Israel Ltd.
FirstEnergy Corp.	Leggett & Platt, Inc.

Horse Race


190
180
170
160
150
140
130
120
110
100
90
2-Jan-02
2-Mar-02
2-May-02
2-Jul-02
2-Sep-02
2-Nov-02
2-Jan-03
2-Mar-03
2-May-03
2-Jul-03
2-Sep-03
2-Nov-03
2-Jan-04
2-Mar-04
2-May-04
2-Jul-04
2-Sep-04
2-Nov-04
2-Jan-05
2-Mar-05
2-May-05
2-Jul-05
2-Sep-05
2-Nov-05
2-Jan-06
2-Mar-06
2-May-06
2-Jul-06
2-Sep-06
2-Nov-06
2-Jan-07
2-Mar-07
2-May-07
2-Jul-07
2-Sep-07
2-Nov-07
PCA using 15 eigenportfolios
Actual ETF factors

Performance Measured by the Sharpe's Ratio

- Sharpe ratios for actual 15 ETFs as factors: 2002—2007.
- Industry Sharpe ratios assume beta-neutrality with respect to the corresponding ETF.
- We observe that the eigenportfolio achieved the Sharpe ratios above 1.0 in 2002 and 2004.

	HHH	IYR	IYT	OIH	RKH	RTH	SMH	UTH	XLE	XLF	XLI	XLK	XLP	XLV	XLY	Portfolio
2002	1.9	2.1	1.4	0.6	2.4	1.5	-0.7	-0.2	-0.2	1.8	0.7	1.5	1.8	-0.1	2.4	2.7
2003	-0.2	0.8	-0.3	-0.5	1.4	1.1	-1.0	-0.1	0.5	0.6	-0.6	2.6	0.3	-0.4	-0.4	0.8
2004	0.9	1.6	-0.7	0.4	0.5	0.1	0.2	-0.4	0.6	0.6	1.4	1.9	0.5	-0.6	0.3	1.6
2005	0.3	-1.5	0.8	-0.6	0.3	0.5	0.5	-1.1	-0.1	0.9	0.6	1.3	-0.7	0.2	0.0	0.1
2006	-0.2	-1.3	0.0	-0.2	0.9	-0.1	0.5	1.7	-0.5	-0.6	1.7	1.7	0.0	-0.4	2.0	0.7
2007	-0.4	-0.3	0.0	-1.3	-1.2	-0.7	0.9	-0.7	-1.0	-0.6	1.1	0.6	0.4	-0.5	1.3	-0.2
Incept	0.4	0.2	0.2	-0.3	0.7	0.4	0.1	-0.1	-0.1	0.5	0.8	1.6	0.4	-0.3	0.9	0.9

Summary

- Each stock return in the investment universe can be decomposed into its projection on the m factors and a residual.
- The PCA approach corresponds to modelling the correlation matrix of stock returns as a sum of a rank-m matrix corresponding to the significant spectrum and a diagonal matrix of full rank:

$$\overline{\rho}_{ij} = \sum_{k=1}^{m} \lambda_k v_i^{(k)} v_j^{(k)} + \epsilon_{ij}^2 \delta_{ij},$$

where δ_{ij} is the Kronecker delta and ϵ_{ii}^2 is given by

$$\epsilon_{ii}^2 = 1 - \sum_{k=1}^{m} \lambda_k v_i^{(k)} v_i^{(k)},$$

so that $\overline{\rho}_{ii} = 1$.