Risk and Portfolio Management
Spring 2010

Principal Components Analysis and
Factors explaining stock returns
Principal components analysis for equity markets

-- Define a universe, or collection of stocks corresponding to the market of interest (e.g. US Equities, Nasdaq-100, Brazilian equities components of S&P 500)

-- Collect as much data as possible

-- On any given date, perform PCA on the correlation matrix, going back for T periods (days). The analysis is on a T by N matrix

-- Estimate the number of significant components

-- Analyze the corresponding eigenvectors and eigenportfolios (factors)

-- Associate the factors to features of the market (e.g. sectors, market cap, etc)
Stocks of more than 1BB cap in January 2007

<table>
<thead>
<tr>
<th>Sector</th>
<th>ETF</th>
<th>Num of Stocks</th>
<th>Market Cap</th>
<th>unit: 1M/usd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average</td>
<td>Max</td>
</tr>
<tr>
<td>Internet</td>
<td>HHH</td>
<td>22</td>
<td>10,350</td>
<td>104,500</td>
</tr>
<tr>
<td>Real Estate</td>
<td>IYR</td>
<td>87</td>
<td>4,789</td>
<td>47,030</td>
</tr>
<tr>
<td>Transportation</td>
<td>IYT</td>
<td>46</td>
<td>4,575</td>
<td>49,910</td>
</tr>
<tr>
<td>Oil Exploration</td>
<td>OIH</td>
<td>42</td>
<td>7,059</td>
<td>71,660</td>
</tr>
<tr>
<td>Regional Banks</td>
<td>RKH</td>
<td>69</td>
<td>23,080</td>
<td>271,500</td>
</tr>
<tr>
<td>Retail</td>
<td>RTH</td>
<td>60</td>
<td>13,290</td>
<td>198,200</td>
</tr>
<tr>
<td>Semiconductors</td>
<td>SMH</td>
<td>55</td>
<td>7,303</td>
<td>117,300</td>
</tr>
<tr>
<td>Utilities</td>
<td>UTH</td>
<td>75</td>
<td>7,320</td>
<td>41,890</td>
</tr>
<tr>
<td>Energy</td>
<td>XLE</td>
<td>75</td>
<td>17,800</td>
<td>432,200</td>
</tr>
<tr>
<td>Financial</td>
<td>XLF</td>
<td>210</td>
<td>9,960</td>
<td>187,600</td>
</tr>
<tr>
<td>Industrial</td>
<td>XLI</td>
<td>141</td>
<td>10,770</td>
<td>391,400</td>
</tr>
<tr>
<td>Technology</td>
<td>XLK</td>
<td>158</td>
<td>12,750</td>
<td>293,500</td>
</tr>
<tr>
<td>Consumer Staples</td>
<td>XLP</td>
<td>61</td>
<td>17,730</td>
<td>204,500</td>
</tr>
<tr>
<td>Healthcare</td>
<td>XLV</td>
<td>109</td>
<td>14,390</td>
<td>192,500</td>
</tr>
<tr>
<td>Consumer discretionary</td>
<td>XLY</td>
<td>207</td>
<td>8,204</td>
<td>104,500</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1417</td>
<td>11,291</td>
<td>432,200</td>
</tr>
</tbody>
</table>

January, 2007
Principal Components Analysis of Correlation Data

Consider a time window \( t=0,1,2,...,T, \) (days) a universe of \( N \) stocks. The returns data is represented by a \( T \) by \( N \) matrix \( (R_{it}) \)

\[
\sigma_i^2 = \frac{1}{T} \sum_{t=1}^{T} (R_{it} - \overline{R}_i)^2, \quad \overline{R}_i = \frac{1}{T} \sum_{t=1}^{T} R_{it}
\]

\[
Y_{it} = \frac{R_{it} - \overline{R}_i}{\sigma_i} \quad \text{Standardized returns}
\]

\[
\Gamma_{ij} = \frac{1}{T} \sum_{t=1}^{T} Y_{it} Y_{jt}
\]

Clearly,

\[
\text{Rank}(\Gamma) \leq \min(N, T)
\]
Regularized correlation matrix

\[ C_{ij} = \frac{1}{T} \sum_{t=1}^{T} Y_{it} Y_{jt} + \gamma \delta_{ij}, \quad \gamma = 10^{-9} \]

\[ \Gamma_{ij}^{\text{reg}} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} = \frac{C_{ij}}{1+\gamma} \]

This matrix is a correlation matrix and is positive definite. It is equivalent for all practical purposes to the original one but is numerically stable for inversion and eigenvector analysis (e.g. with Matlab).

Note: this is especially useful when \( T<<N \).
Eigenvalues, Eigenvectors and Eigenportfolios

\[ \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_N > 0 \]
eigenvalues

\[ \mathbf{V}^{(j)} = (V_1^{(j)}, V_2^{(j)}, \ldots, V_N^{(j)}), \quad j = 1, 2, \ldots, N. \]
eigenvectors

\[ F_{jt} = \sum_{i=1}^{N} V_i^{(j)} Y_{it} = \sum_{i=1}^{N} \left( \frac{V_i^{(j)}}{\sigma_i} \right) R_{it} \]
returns of “eigenportfolios”

We shall use the coefficients of the eigenvectors and the volatilities of the stocks to build “portfolio weights”. These random variables \( F_j \) span same linear space as the original returns.
Expressing Stock Returns in terms of returns of Eigenportfolios (a bit of linear algebra)

\[ \text{Correl}(R_i, R_j) = C_{ij} = \sum_{k=1}^{N} \lambda_k V_i^{(k)} V_j^{(k)} \]

Define:
\[ F_k \equiv \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} \frac{V_i^{(k)}}{\sigma_i} R_i \]

\[ \text{Variance}(F_k) = 1, \quad \text{Cov}(F_k, F_l) = \delta_{kl} \]

Set:
\[ \beta_{ik} = \text{Cov}(R_i, F_k) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{N} V_j^{(k)} \sigma_i C_{ij} \]

\[ = \frac{1}{\sqrt{\lambda_k}} \sigma_i \lambda_k V_i^{(k)} = \sigma_i \sqrt{\lambda_k} V_i^{(k)} \]

Then
\[ R_i = \sum_{k=1}^{N} \beta_{ik} F_k + \alpha \]
50 largest eigenvalues using the 1400 US stocks with cap >1BB cap (Jan 2007)

N~1400 stocks
T=252 days
Top 50 eigenvalues for S&P 500 index components, May 1 2007, T=252
Nasdaq-100
Components of NDX/QQQQ

Data: Jan 30, 2007 to Jan 23, 2009
502 dates, 501 periods
99 Stocks (1 removed) MNST (Monster.com), now listed in NYSE
Bai and Ng 2002, *Econometrica*

Parsimonious approach for factor selection

\[
I(m) = \min_{\alpha, \beta} & \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( R_{it} - \sum_{k=1}^{m} \beta_{ik} F_{kt} - \alpha_{i} \right)^{2}
\]

Least squares penalty function

\[
m^* = \arg \min_{m} (I(m) + m \cdot g(N, T))
\]

\[
\lim_{N,T \to \infty} g(N, T) = 0, \quad \lim_{N,T \to \infty} \min(N,T) g(N, T) = \infty
\]

Under reasonable assumptions on the underlying model, Bai and Ng prove that under PCA estimation, \( m^* \) converges in probability to the true number of factors as \( N, T \to \infty \).
Connection with eigenvalues of correlation matrix

\[
J(m) \equiv \arg \min_{\alpha, \beta} \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left( R_{it} - \sum_{k=1}^{m} \beta_{ik} F_{kt} - \alpha \right)^2
\]

\[
J(m) = \sum_{k=m+1}^{N} \lambda_k \quad \text{also,} \quad I(m) = \sum_{k=m+1}^{N} \lambda_k \left( \sum_{i=1}^{N} \sigma_i^2 (V_i^{(k)})^2 \right)
\]

\[
m^* = \arg \min_{m} \left( \sum_{k=m+1}^{N} \lambda_k + mg(N,T) \right) \quad \text{Linear penalty function}
\]

For finite samples, we need to adjust the slope \(g(N,T)\). Apparently, Bai and Ng (2002) tend to underestimate the number of factors in Nasdaq stocks considerably. (2 factors, T=60 monthly returns, N=8000 stocks)
Useful quantities

\[ \frac{1}{N} \sum_{k=1}^{m} \lambda_k = \text{Explained variance by first } m \text{ eigenvectors} \]

\[ \frac{1}{N} \sum_{k=m+1}^{N} \lambda_k = \text{Tail} \]

\[ \frac{1}{N} \sum_{k=m+1}^{N} \lambda_k + g \frac{m}{N} = \text{Objective Function} = U(m, g) \]

Convexity = \[ \frac{\partial^2 U(m^*(g), g)}{\partial g^2} \]
Objective function $U(m,g)$
Optimal value of $U(m,g)$ for different $g$
## Implementation of Bai & Ng on SP500 Data

<table>
<thead>
<tr>
<th>g</th>
<th>m*</th>
<th>Lambda_m*</th>
<th>Explained Variance</th>
<th>Tail</th>
<th>Objective Fun</th>
<th>Convexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>117</td>
<td>0.20%</td>
<td>87.88%</td>
<td>12.12%</td>
<td>0.355</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>0.39%</td>
<td>71.44%</td>
<td>28.56%</td>
<td>0.522</td>
<td>-0.085085</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>0.59%</td>
<td>57.11%</td>
<td>42.89%</td>
<td>0.603</td>
<td>-0.041266</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>0.76%</td>
<td>48.51%</td>
<td>51.49%</td>
<td>0.643</td>
<td>-0.018110</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.96%</td>
<td>43.52%</td>
<td>56.48%</td>
<td>0.665</td>
<td>-0.007000</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1.18%</td>
<td>40.43%</td>
<td>59.57%</td>
<td>0.680</td>
<td>-0.003096</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>1.22%</td>
<td>39.25%</td>
<td>60.75%</td>
<td>0.691</td>
<td>-0.004872</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>1.56%</td>
<td>36.56%</td>
<td>63.44%</td>
<td>0.698</td>
<td>0.001069</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>1.56%</td>
<td>36.56%</td>
<td>63.44%</td>
<td>0.706</td>
<td>0.000000</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>1.56%</td>
<td>36.56%</td>
<td>63.44%</td>
<td>0.714</td>
<td>0.000000</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>1.56%</td>
<td>36.56%</td>
<td>63.44%</td>
<td>0.722</td>
<td>0.000000</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>1.56%</td>
<td>36.56%</td>
<td>63.44%</td>
<td>0.730</td>
<td>0.000000</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>1.56%</td>
<td>36.56%</td>
<td>63.44%</td>
<td>0.738</td>
<td>-</td>
</tr>
</tbody>
</table>

If we choose the cutoff m* as the one for which the sensitivity to g is zero, then m*~5 to 7 seems appropriate. This would lead to the conclusion that the S&P 500 corresponds to a 5-factor model. The number is small in relation to industry sectors and to the amount of variance explained by industry factors.
The density of states: a useful formalism for finding significant EVs


\[ F(E) \equiv \frac{\# \{k : \lambda_k / N \leq E \}}{N} \quad \text{\(F(E)\) is increasing, \(F(1) = 1\)} \]

\[ f(E) = \frac{1}{N} \sum_k \delta \left( E - \frac{\lambda_k}{N} \right) \quad \therefore \quad F'(E) = f(E) \quad \text{D.O.S.} \]

One way to think about the DOS is as changing the x-axis for the y-axis, i.e. counting the number of eigenvalues in a neighborhood of any \(E\), \(0 < E < 1\).

Intuition: if \(N\) is large, the eigenvalues of the insignificant portion of the spectrum will \``bunch up’’ into a continuous distribution \(f(E)\).
Integrated DOS

![Graph of Integrated DOS](image)
In the DOS language...

\[
\frac{1}{N} \sum_{k=m+1}^{N} \lambda_k = \int_0^E E \ f(E) dE, \quad \frac{m}{N} = 1 - F(\lambda_m)
\]

\[
U(E, g) = \int_0^E x \ f(x) dx + g(1 - F(E))
\]

\[
\frac{\partial U(E, g)}{\partial E} = E \ f(E) - g f(E) = (E - g) f(E)
\]

If \( f(g) \neq 0 \), then \( E^*(g) = g \).
Dependence of the problem on $g$

$$V(g) = U(E^*(g), g) = \int_0^g xf(x)dx + g(1 - F(g))$$

$$= gF(g) - \int_0^g F(x)dx + g - gF(g)$$

$$= g - \int_0^g F(x)dx$$

$$V'(g) = 1 - F(g)$$

$$V''(g) = -f(g)$$

According to this calculation, the best cutoff is the level $E$ where the DOE vanishes (or nearly vanishes) coming from the left, i.e. from the smallest eigenvalues.
Density of States (from previous data for Nasdaq 100)

One large mass at 0.44,
Some masses near 0.025
Nearly continuous density for lower levels
Zoom of the DOS for low eigenvalues

“Edge of DOS”
Random Matrix Theory

\[ X_{tn}, \quad t = 1,2,..,T, \quad n = 1,2,...,N \]
\[ X \sim N(0,1) \]

\[ W_{mn} = \sum_{t=1}^{T} X_{tm} X_{tn}, \quad W = X^t X \]

\[ \lambda_n, \quad n = 1,2,...N \quad \text{eigenvalues of } W \]

What are the statistical properties of the eigenvalues as \( N, T \) tend to infinity?

What are the fluctuations of the eigenvalues for large \( N, T \)?
Marcenko-Pastur Theorem
(hard)

\[ F(E) = \lim_{N \to \infty} \frac{\#\{k : \frac{\lambda_k}{N} \leq E\}}{N} \quad \text{integrated DOE} \]

\[ f(E) = \frac{dF(E)}{dE} \]

\[ f(E) = \frac{1}{2\pi\gamma} \frac{\sqrt{(E_+ - E)(E - E_-)}}{E} \quad E_\pm = \left(1 + \sqrt{\gamma}\right)^2 \]
Marcenko Pastur (N/T=2)

DOS for a random Gaussian ensemble
Bouchaud, Cizeau, Laloux, Potters  
(PRL, 1999)

The bulk distribution for spectrum of S&P 500 is described approximately by Marcenko-Pastur properly normalized but there are detached eigenvalues (the significant ones!)

Bulk spectrum  

significant eigenvalues
Results of PCA with DOS analysis for Nasdaq 100

-- 4 significant eigenvectors/eigenvalues

-- first Eigen-state explains about 44% of the correlation

-- total explained variance= 51%

-- Now we need to indentify the eigenportfolios in terms of real market factors (industry, size, etc, etc).
First Eigenvector: Market

Sorted Eigenvector

Sorted Weights

Ticker

BBBY
LLTC
DELL
MRVL
INFY
ISRG
CMCSA
CTXS
QCOM
CSCO

Ticker

CSCO
INTC
ORCL
ADBE
DTV
SIAL
MSFT
SPLS
INFY
PAYX
Second Eigenvector: Biotech vs. Chips

Sorted by coefficients V

Genentech (GENZ) and Bausch & Lomb (BRCM) are among the top performers.

Sorted by weights

AMN Genetics (GENZ) and Micron Technology (MTCH) are leading in terms of weight.

Top 10

<table>
<thead>
<tr>
<th>Top 10</th>
<th>Bottom 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENZ</td>
<td>MRVL</td>
</tr>
<tr>
<td>CEPH</td>
<td>NVDA</td>
</tr>
<tr>
<td>HSIC</td>
<td>FWLT</td>
</tr>
<tr>
<td>CELG</td>
<td>BRCM</td>
</tr>
<tr>
<td>GILD</td>
<td>SNDK</td>
</tr>
<tr>
<td>BIIB</td>
<td>JOYG</td>
</tr>
<tr>
<td>XRAY</td>
<td>RIMM</td>
</tr>
<tr>
<td>AMLN</td>
<td>BIDU</td>
</tr>
<tr>
<td>CTAS</td>
<td>LRCX</td>
</tr>
<tr>
<td>ESRX</td>
<td>ALTR</td>
</tr>
</tbody>
</table>

Top 10

<table>
<thead>
<tr>
<th>Top 10</th>
<th>Bottom 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENZ</td>
<td>BRCM</td>
</tr>
<tr>
<td>BBBY</td>
<td>FWLT</td>
</tr>
<tr>
<td>BIIB</td>
<td>DELL</td>
</tr>
<tr>
<td>GILD</td>
<td>FMCN</td>
</tr>
<tr>
<td>CEPH</td>
<td>AKAM</td>
</tr>
<tr>
<td>CELG</td>
<td>BIDU</td>
</tr>
<tr>
<td>ESRX</td>
<td>ALTR</td>
</tr>
<tr>
<td>CTAS</td>
<td>FLEX</td>
</tr>
<tr>
<td>AMGN</td>
<td>AMAT</td>
</tr>
</tbody>
</table>
Third Eigenvector: Manufacturing vs. Chips

Sorted by coefficients $V$

Top 10
- FWLT
- DISCA
- GOO
- GILD
- INFY
- DTV
- HOLX
- RIMM
- DELL
- JNPR

Bottom 10
- LEAP
- ALTR
- JOYG
- BBBY
- STLD
- PETM
- TEVA
- LRCX
- DISCA
- AMAT
- DISH
- LLTC
- CEPH
- SHLD
- ATVI
- XLNX
- LBTYA
- BRCM

Sorted by weights

Top 10
- LEAP
- BIDU
- XRAY
- BIIB
- CMCSA
- DELL
- HIS
- AOBE
- MSFT
- VRX
- NVDA
- SYMC
- SNDK
- WFMI
- COST
- XLN
- BRCM
- AMAT

Bottom 10
- FWLT
- KLAC
- LEAP
- ALTR
- JOYG
- BBBY
- STLD
- PETM
- TEVA
- LRCX
- DISCA
- AMAT
- DISH
- LLTC
- CEPH
- SHLD
- ATVI
- XLNX
- LBTYA
- BRCM
- JOYG
- KLAC
``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 or sector.
Identification of the Eigenportfolios via ETFs

Identify the Eigenportfolios by making multiple regressions or "greedy" regressions on the returns of Exchange Traded Funds

<table>
<thead>
<tr>
<th>ETF Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFA</td>
<td>Europe &amp; Far East</td>
</tr>
<tr>
<td>HHH</td>
<td>Internet</td>
</tr>
<tr>
<td>IBB</td>
<td>Biotechnology</td>
</tr>
<tr>
<td>IYT</td>
<td>Transportation</td>
</tr>
<tr>
<td>QQQQ</td>
<td>Nasdaq 100 Index Tracker</td>
</tr>
<tr>
<td>RTH</td>
<td>Retail</td>
</tr>
<tr>
<td>SMH</td>
<td>Semiconductors</td>
</tr>
<tr>
<td>XLB</td>
<td>Materials</td>
</tr>
<tr>
<td>XLI</td>
<td>Industrials</td>
</tr>
<tr>
<td>XLK</td>
<td>Technology</td>
</tr>
<tr>
<td>XLP</td>
<td>Consumer Staples</td>
</tr>
<tr>
<td>XLV</td>
<td>Health Care</td>
</tr>
<tr>
<td>XLY</td>
<td>Consumer Discretionary</td>
</tr>
</tbody>
</table>
First Eigenportfolio (NDX)
Second Eigenportfolio (NDX)
Third Eigenportfolio (NDX)