Principal Components Analysis and Factors explaining stock returns
Correlation matrices in finance

\[ \Gamma_{ij} = \text{Corr}(R_i, R_j) \quad R_i = \text{return of stock \#i, } i = 1, \ldots, N \]

Estimation of correlation matrix from data requires selecting a sample size, or estimation window, \( T \).

If the universe of assets is large, then \( T<<N \) (e.g. \( T=252, N=1500 \))

The correlation matrix is not “full rank” in general since we expect that the stocks are “driven” by a few components

\[ R_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} F_k + \epsilon_i, \quad m \ll N, \quad (e.g., \ m = 15) \]

- **Degeneracy issue** (rank<dimension)
- **Noise issue** (determine the “right” number of factors, avoid numerical error)
Principal components analysis of stock returns

-- 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)
``Clean’’ correlation matrix from factor model

\[ R_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} F_k + \epsilon_i, \quad m \ll N, \ (e.g., \ m = 15) \]

\[ \Gamma_{ij} = \sum_{kl=1,m} \beta_{ik} G_{kl} \beta_{jl} + D_{ij} = \Gamma_{ij}^{(s)} + \Gamma_{ij}^{(f)} \]

- Systematic component is N by N with rank m
- Idiosyncratic component is diagonal N by N

This approach decomposes the variance of any asset into a sum of ``market-explained” variance and a company-specific variance
## 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 (unit: 1M/USD)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Average</td>
<td>Max</td>
<td>Min</td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>HHH</td>
<td>22</td>
<td>10,350</td>
<td>104,500</td>
<td>1,047</td>
<td></td>
</tr>
<tr>
<td>Real Estate</td>
<td>IYR</td>
<td>87</td>
<td>4,789</td>
<td>47,030</td>
<td>1,059</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>IYT</td>
<td>46</td>
<td>4,575</td>
<td>49,910</td>
<td>1,089</td>
<td></td>
</tr>
<tr>
<td>Oil Exploration</td>
<td>OIH</td>
<td>42</td>
<td>7,059</td>
<td>71,660</td>
<td>1,010</td>
<td></td>
</tr>
<tr>
<td>Regional Banks</td>
<td>RKH</td>
<td>69</td>
<td>23,080</td>
<td>271,500</td>
<td>1,037</td>
<td></td>
</tr>
<tr>
<td>Retail</td>
<td>RTH</td>
<td>60</td>
<td>13,290</td>
<td>198,200</td>
<td>1,022</td>
<td></td>
</tr>
<tr>
<td>Semiconductors</td>
<td>SMH</td>
<td>55</td>
<td>7,303</td>
<td>117,300</td>
<td>1,033</td>
<td></td>
</tr>
<tr>
<td>Utilities</td>
<td>UTH</td>
<td>75</td>
<td>7,320</td>
<td>41,890</td>
<td>1,049</td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td>XLE</td>
<td>75</td>
<td>17,800</td>
<td>432,200</td>
<td>1,035</td>
<td></td>
</tr>
<tr>
<td>Financial</td>
<td>XLF</td>
<td>210</td>
<td>9,960</td>
<td>187,600</td>
<td>1,000</td>
<td></td>
</tr>
<tr>
<td>Industrial</td>
<td>XLI</td>
<td>141</td>
<td>10,770</td>
<td>391,400</td>
<td>1,034</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>XLK</td>
<td>158</td>
<td>12,750</td>
<td>293,500</td>
<td>1,008</td>
<td></td>
</tr>
<tr>
<td>Consumer Staples</td>
<td>XLP</td>
<td>61</td>
<td>17,730</td>
<td>204,500</td>
<td>1,016</td>
<td></td>
</tr>
<tr>
<td>Healthcare</td>
<td>XLV</td>
<td>109</td>
<td>14,390</td>
<td>192,500</td>
<td>1,025</td>
<td></td>
</tr>
<tr>
<td>Consumer discretionary</td>
<td>XLY</td>
<td>207</td>
<td>8,204</td>
<td>104,500</td>
<td>1,007</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>1417</strong></td>
<td><strong>11,291</strong></td>
<td><strong>432,200</strong></td>
<td><strong>1,000</strong></td>
<td></td>
</tr>
</tbody>
</table>

January, 2007
Principal Components Analysis of Correlation Data

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

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

\[
Y_{it} = \frac{R_{it} - \bar{R}_i}{\sigma_i}
\]

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

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

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

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

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

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

\[ F_{jt} = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{N} V_i^{(j)} Y_{it} = \sum_{i=1}^{N} \left( \frac{1}{\sqrt{\lambda_j}} \frac{V_i^{(j)}}{\sigma_i} \right) R_{it} \]

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)

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

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

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

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

Then
\[ R_i = \sum_{k=1}^{N} \beta_{ik} F_k + \alpha \quad \text{expresses the result of multiple-regression of returns on factors} \]
50 largest eigenvalues using the 1400 US stocks with cap >1BB cap (Jan 2007)

\[ N \approx 1400 \text{ stocks} \]
\[ T = 252 \text{ days} \]
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

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

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

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

\[ \text{Convexity} = \frac{\partial^2 U(m^*(g), g)}{\partial g^2} \]
Objective function $U(m,g)$
Optimal value of $U(m,g)$ for different $g$
If we choose the cutoff \( m^* \) as the one for which the sensitivity to \( g \) is zero, then \( m^* \approx 5 \) 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 : \frac{\lambda_k}{N} \leq E\}}{N} \quad \text{\text{\( \bullet \)}} \quad F(E) \text{ is increasing,} \quad F(1) = 1 \]

\[ f(E) = \frac{1}{N} \sum_k \delta\left( E - \frac{\lambda_k}{N} \right) \quad \text{\text{\( \bullet \)}} \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 showing the integrated DOS with the y-axis labeled F(E) and the x-axis labeled E. The graph demonstrates an increasing curve from E=0 to E=1, indicating the cumulative distribution of energy levels.](image-url)
In the DOS language...

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

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

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

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

\[ X_{tn}, \ t = 1,2,\ldots,T, \quad n = 1,2,\ldots,N \]

\[ X \sim N(0,1) \]

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

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

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 Distribution for the DOS of a Random Correlation Matrix

**Theorem:** Let $X$ be a $T$ by $N$ matrix of standardized normal random variables and let $C=X'X$. Then, the DOS of $C$ approaches the Marcenko Pastur distribution as $N,T$ tend to infinity with the ratio $N/T$ held constant.

$$\gamma = \frac{N}{T} \quad \lambda_+ = \left(1 + \sqrt{\gamma}\right)^2 \quad \lambda_- = \left(1 - \sqrt{\gamma}\right)^2$$

$$MP(\lambda) = \left(1 - \frac{1}{\gamma}\right)^+ \delta(\lambda) + \frac{1}{2\pi\gamma} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}$$
Marcenko-Pastur Distribution
\[ \text{gamma} = \frac{500}{269} = 1.858736 \]
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!)
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
Returns of First PCA eigenportfolio (S&P 500) compared with S&P 500 returns (1/5/2009 to 1/29/2010)

\[ Y = 0.9856 \times X + 0.0223 \]

\[ R^2 = 0.9716 \]
Second Eigenvector: Biotech vs. Chips

Sorted by coefficients V

Sorted by weights

Top 10

<table>
<thead>
<tr>
<th>Company</th>
<th>Top 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>

Bottom 10

<table>
<thead>
<tr>
<th>Company</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>Company</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENZ</td>
<td>MRVL</td>
</tr>
<tr>
<td>BRCM</td>
<td>MRVL</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>CTAS</td>
<td>LRCX</td>
</tr>
<tr>
<td>ESRX</td>
<td>ALTR</td>
</tr>
<tr>
<td>AMGN</td>
<td>AMAT</td>
</tr>
</tbody>
</table>

Bottom 10

<table>
<thead>
<tr>
<th>Company</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>
<tr>
<td>AMGN</td>
<td>AMAT</td>
</tr>
</tbody>
</table>
Third Eigenvector: Manufacturing vs. Chips

Sorted by coefficients V

Sorted by weights

Top 10
- FWLT
- LEAP
- JOYG
- STLD
- TEVA
- DISCA
- DISH
- CEPEH
- ATVI
- GOOG

Bottom 10
- KLAC
- ALTR
- BBBY
- PETM
- LRCX
- AMAT
- LLTC
- SHLD
- XLNX
- BRCM

Top 10
- LEAP
- FWLT
- DISCA
- FMCN
- NIHD
- ATVI
- DISC
- LRCX
- SHLD
- AMAT

Bottom 10
- BBBY
- LLTC
- SHLD
- AMAT
- ALTR
- PETM
- MRVL
- LRCX
- BRCM
- 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 of sector.
Identification of the Eigenportfolios via ETFs

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

- EFA: Europe & Far East
- HHH: Internet
- IBB: Biotechnology
- IYT: Transportation
- QQQQ: Nasdaq 100 Index Tracker
- RTH: Retail
- SMH: Semiconductors
- XLB: Materials
- XLI: Industrials
- XLK: Technology
- XLP: Consumer Staples
- XLV: Health Care
- XLY: Consumer Discretionary
First Eigenportfolio (NDX)

Nasdaq Tracker QQQQ
Scaled returns QQQQ vs 1st EV

\[ y = 0.8174x - 3 \times 10^{-5} \]

\[ R^2 = 0.6681 \]
Second Eigenportfolio (NDX)

- EFA
- HHH
- IBB
- IYT
- QQQQ
- RTH
- SMH
- XLB
- XLI
- XLK
- XLP
- XLV
- XLY

- Consumer Staples
- Industrials
- Retail
- Non-US
- Transportation
- Biotech
- internet