Topics in Probability: Quantitative Investments Strategies

Marco Avellaneda

G63.2936.001

Spring Semester 2009

Syllabus

1. Risk Models, Factor Analysis and Correlation Structures

-- Statistical models of stock returns
-- The classics: CAPM, APT
-- Factor analysis
-- Dynamic PCA of correlation matrices
-- Economic significance of eigenvectors & eigenportfolios
-- Exchange-traded Funds (ETFs)
-- Factor analysis via ETFs
-- Random matrix theory
-- Examples: US equities, NASDAQ, EM bonds, Brazil, China, European stocks
-- Risk-functions and dynamic risk-management of equity portfolios

Syllabus

2. Statistical arbitrage for cash equities

-- Long-short market-neutral investment portfolios
-- Leverage & setting ex-ante performance targets
-- Performance measures
-- Back-testing concepts: in-sample/out-of-sample performance, survivorship biases
-- Time-series analysis of stock residuals
-- PCA-based residuals
-- ETF-based residuals
-- Extracting information from trading volume (subordination)

Syllabus

3. Statistical arbitrage in options markets

-- Option markets revisited
-- Volatility and options trading
-- Data issues with option markets, implied dividend
-- Modeling stock-ETF dynamics and ETF-stock dynamics
-- Weighted Monte-Carlo technique for model calibration
-- Relative-value analysis: options on single stocks
-- Relative-value analysis: options on indices and ETFs
-- Construction of risk-functions for option portfolios
-- Market-neutral option portfolios
-- Dispersion trading
-- Back-testing option portfolio strategies

Course Requirements

-- **Three projects**, or assignments, associated with the different parts of the course. Projects will be approved by instructor.

-- Projects will deal with **real data**. They will involve **programming** and **quantitative financial analysis** as well as your contribution to and interpretation of the theory presented.

-- Programming will involve the management of large (real) datasets, the use of Matlab but also other programming languages and software needed to ``get the job done’’.

-- The grade will be based on the three projects and on **class participation**.

-- Pre-requisites: knowledge of applied statistics, proficiency in at least one programming environment, knowledge of basic finance concepts (e.g., interest rates, present value, stocks, Markowitz, Black Scholes).

-- Books and notes: provided after each lecture.

Statistical Models of Stock Returns

Consider a stock (e.g IBM). The return *R* over a specified period is the change in price, plus dividend payments, divided by the initial price.

$$R_t = \frac{S_{t+\Delta t} - S_t + D_{t,t+\Delta t}}{S_t}$$

How can we explain or predict stock returns?

-- Fundamental analysis (earnings, balance sheet, business analysis)
this will not be considered in this course!

-- ``Trends'' in the prices. (Not very effective)

-- Explanation of the returns/prices based on statistical factors

Factor models

$$R = \sum_{j=1}^{N_f} \beta_j F_j + \varepsilon$$

$F_j, \quad j = 1, \ldots, N_f,$	Explanatory factors
$\beta_j, \quad j = 1, \ldots, N_f,$	Factor loadings
$\sum_{j=1}^{N_f} \beta_j F_j$	Explained, or systematic portion
ε	Residual, or idiosyncratic portion

CAPM: a `minimalist’ approach

Single explanatory factor: the ``market’’, or ``market portfolio’’

$$R = \beta F + \varepsilon, \quad Cov(R, \varepsilon) = 0$$

F = usually taken to be the returns of a broad-market index (e.g., S&P 500)

Normative statement: $< \varepsilon >= 0$ or $< R >= \beta < F >$

Argument: if the market is ``efficient’’, or in ``equilibrium’’, investors cannot make money (systematically) by picking individual stocks and shorting the index or vice-versa (assuming uncorrelated residuals). (Lintner, Sharpe. 1964)

Counter-arguments: (i) the market is not ``efficient’’, (ii) residuals may be correlated (additional factors are needed).

Multi-factor models (APT)

$$R = \sum_{j=1}^{N_f} \beta_j F_j + \varepsilon, \quad Corr(F_j, \varepsilon) = 0$$

Factors represent industry returns (think sub-indices in different sectors, size, financial statement variables, etc).

Normative statement (APT): $< \varepsilon >= 0 \text{ or } < R >= \sum_{j=1}^{N_f} \beta_j < F_j >$

Argument: Generalization of CAPM, based again on no-arbitrage. (Ross, 1976)

Counter-arguments: (i) How do we actually define the factors? (ii) Is the number of factors known? (iii) The structure of the stock market and risk-premia vary strongly (think pre & post WWW) (iv) The issue of correlation of residuals is intimately related to the number of factors.

Factor decomposition in practice

-- Putting aside normative theories (how stocks should behave), factor analysis can be quite useful in practice.

-- In risk-management: used to measure exposure of a portfolio to a particular industry of market feature.

-- Dimension-reduction technique for the study a system with a large number of degrees of freedom

-- Makes Portfolio Theory viable in practice. (Markowitz to Sharpe to Ross!)

-- Useful to analyze stock investments in a relative fashion (buy ABC, sell XYZ to eliminate exposure to an industry sector, for example).

-- New investment techniques arise from factor analysis. The technique is called *defactoring* (Pole, 2007, Avellaneda and Lee, 2008)

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$\left(R_{it}\right)$

$$\sigma_i^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(R_{it} - \overline{R_i}\right)^2, \qquad \overline{R_i} = \frac{1}{T}\sum_{t=1}^{T} R_{it}$$

$$Y_{it} = \frac{R_{it}}{\sigma_i}$$

$$\Gamma_{ij} = \frac{1}{T-1}\sum_{t=1}^{T} Y_{it} Y_{jt}$$

Clearly, $Rank(\Gamma) \le \min\left(N, T\right)$

Regularized correlation matrix

$$C_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\left(R_{it} - \overline{R_i}\right)\left(R_{jt} - \overline{R_j}\right) + \gamma\delta_{ij}, \quad \gamma = 10^{-9}$$

$$\Gamma_{ij}^{reg} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

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

$$\lambda_1 > \lambda_2 \geq .... \geq \lambda_N > 0$$

eigenvalues

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

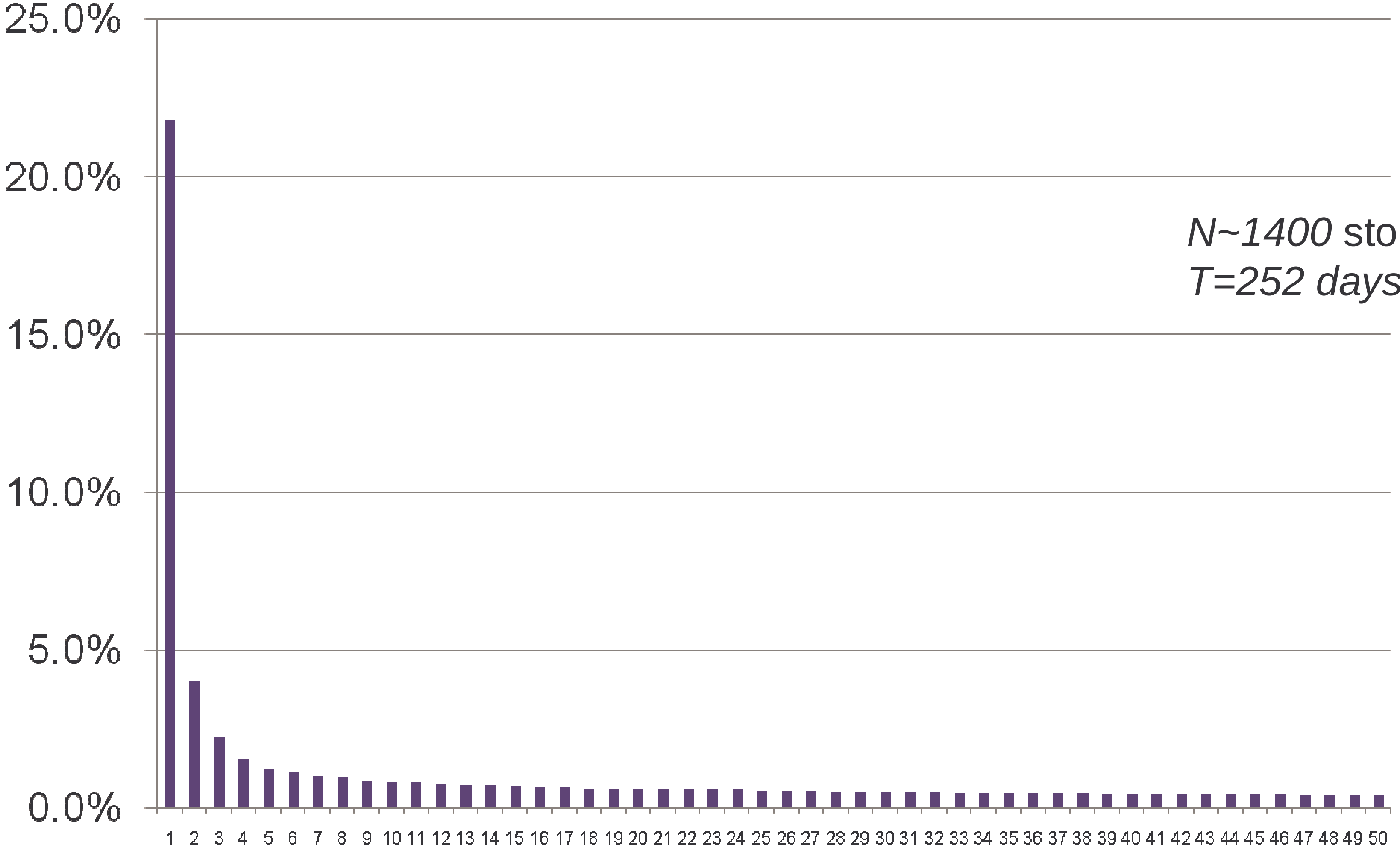
eigenvectors

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

returns of
“eigenportfolios”

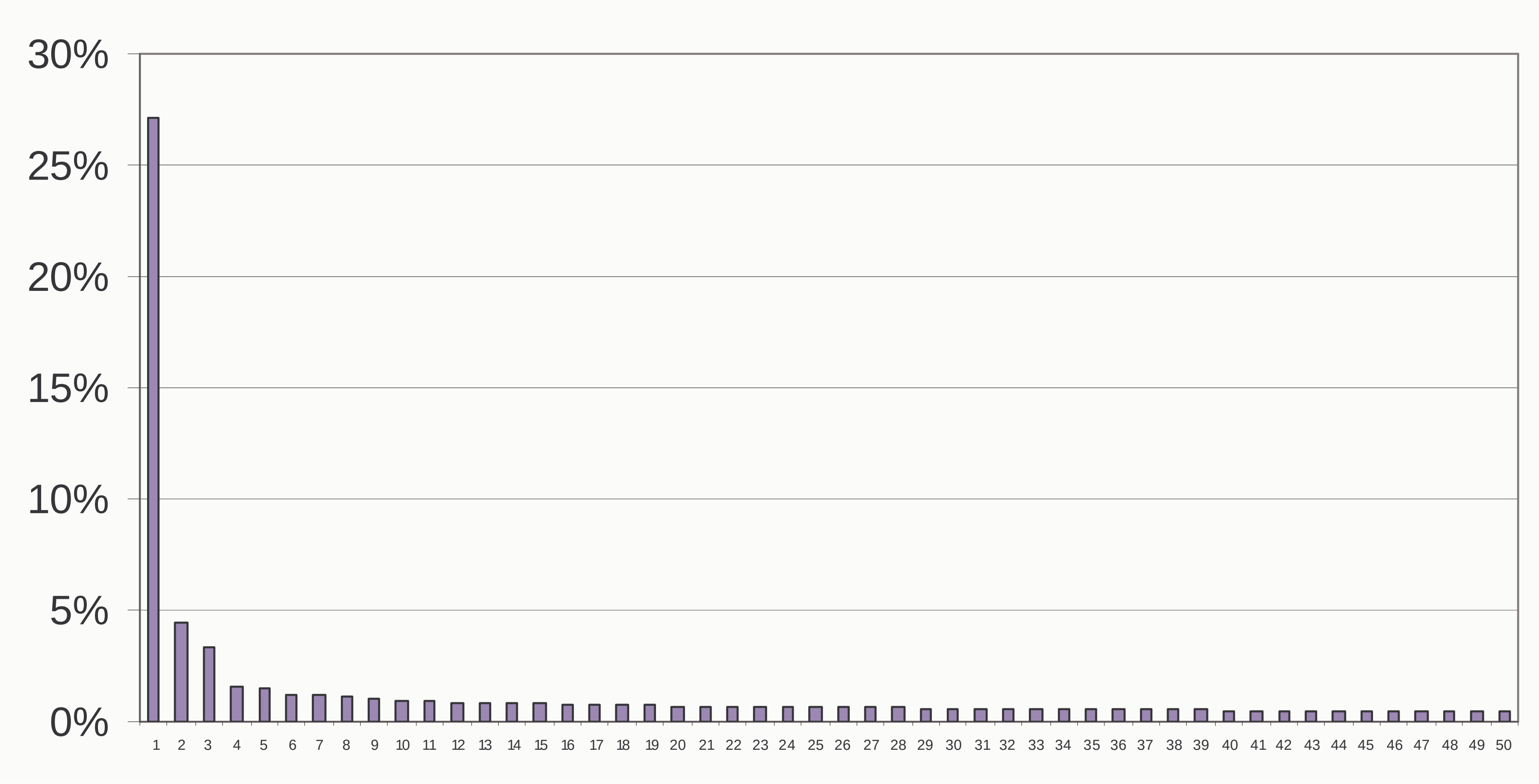
We use the coefficients of the eigenvectors and the volatilities of the stocks to build ``portfolio weights’’. These random variables span the same linear space as the original returns.

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



Model Selection Problem: How many EV are significant?

Need to estimate the significant eigenportfolios which can be used as factors.

Assuming that the correlation matrix is invertible (regularize if necessary)

$$< R_i R_j >= C_{ij} = \sum_{k=1}^{N} \lambda_k V_i^{(k)} V_j^{(k)}$$

$$F_k \equiv \sum_{i=1}^{N} \frac{V_i^{(k)}}{\sigma_i} R_i , \qquad \tilde{F}_k \equiv \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} \frac{V_i^{(k)}}{\sigma_i} R_i$$

$$< F_k^2 >= \lambda_k , \qquad < \tilde{F}_k^2 >= 1, \qquad < \tilde{F}_k \ \tilde{F}_{k'} >= \delta_{kk'}$$

$$R_i = \sum_k \beta_{ik} F_k \quad \Rightarrow \quad \beta_{ik} = \sigma_i \sqrt{\lambda_k} \ V_i^{(k)}$$

Karhunen-Loeve Decomposition

$\mathbf{R} =$ vector of random variables with finite second moment, <.,>=correlation

$\mathbf{C} = < \mathbf{R} \otimes \mathbf{R} > = < \mathbf{RR}' >$ Covariance matrix

$\Omega = \mathbf{C}^{1/2}$ Symmetric square root of C

$\mathbf{F} = \Omega^{-1}\mathbf{R}, \quad \mathbf{R} = \Omega\mathbf{F}$ F has uncorrelated components

$\mathrm{B} = \Omega = \mathbf{C}^{1/2}$ Loadings= components of the square-root of C

Since the eigenvectors vanish or are very small in a real system, the modeling consists in defining a small number of factors and attribute the rest to ``noise’’

Bai and Ng 2002, *Econometrica*

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

$$m^* = \arg\min_{m} \left(I(m) + m \cdot g(N,T) \right)$$

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

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

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

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

$$m^* = \arg\min_{m} \left(\sum_{k=m+1}^{N} \lambda_k + mg(N,T) \right)$$

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

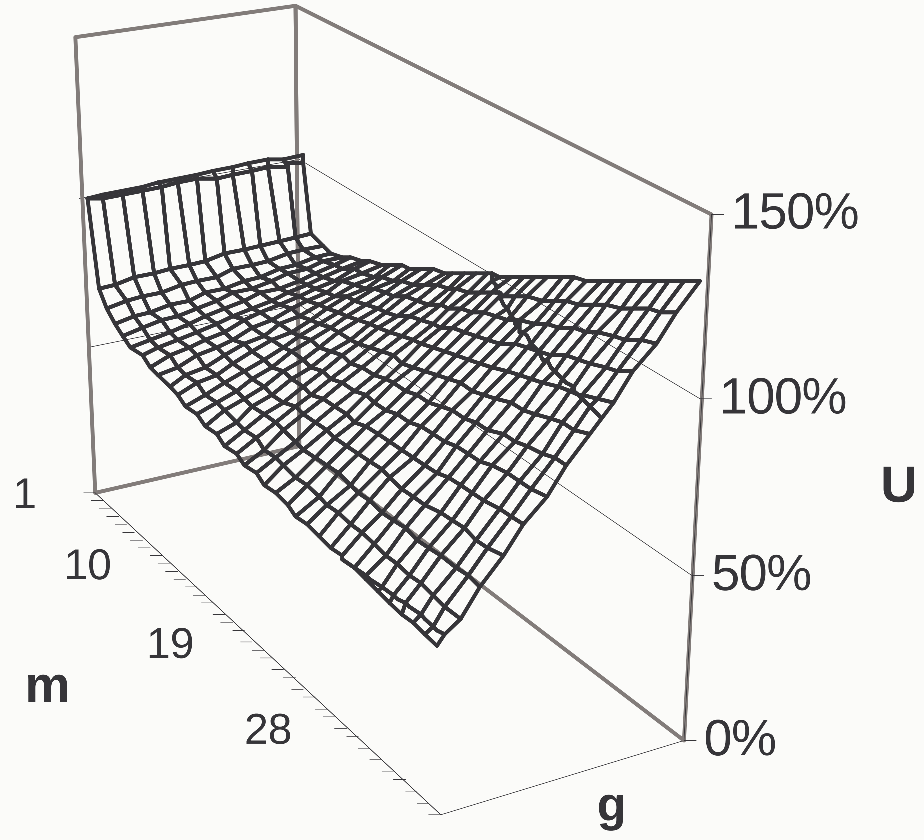
$$\frac{1}{N}\sum_{k=1}^{m}\lambda_k = \text{ Explained variance by first } m \text{ eigenvectors}$$

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

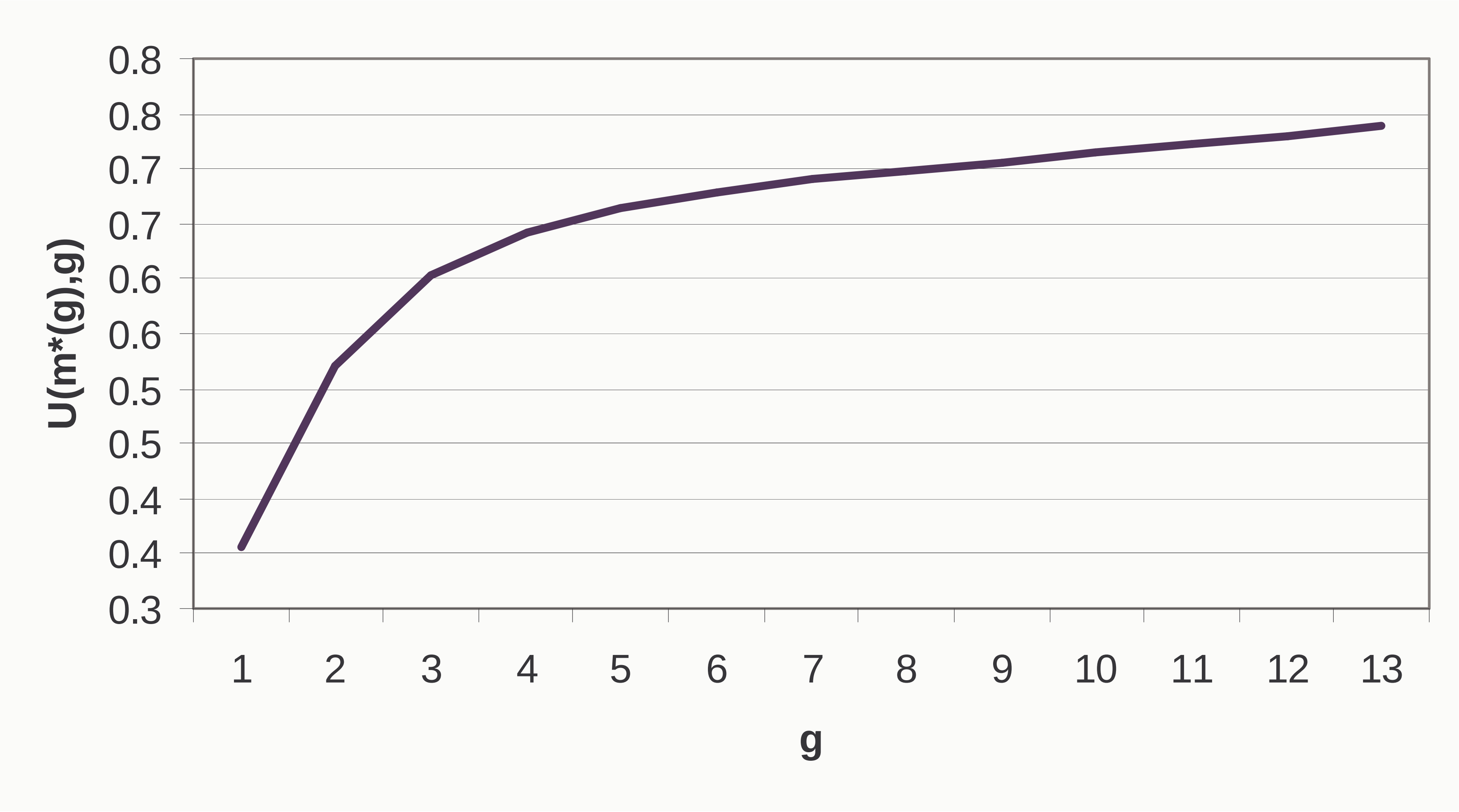
$$\frac{1}{N}\sum_{k=m+1}^{N}\lambda_k + g\,\frac{m}{N} = \text{Objective Function} = U(m,g)$$

$$\text{Convexity} = \frac{\partial^2 U(m^*(g),g)}{\partial g^2}$$

Objective function U(m,g)



Optimal value of U(m,g) for different g



Implementation of Bai & Ng on SP500 Data

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

If we choose the cutoff m* as the one for which the sensitivity to g is zero, then m*~5 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

Spectral theory as seen by physicists – origins in Quantum Mechanics and High Energy Physics.

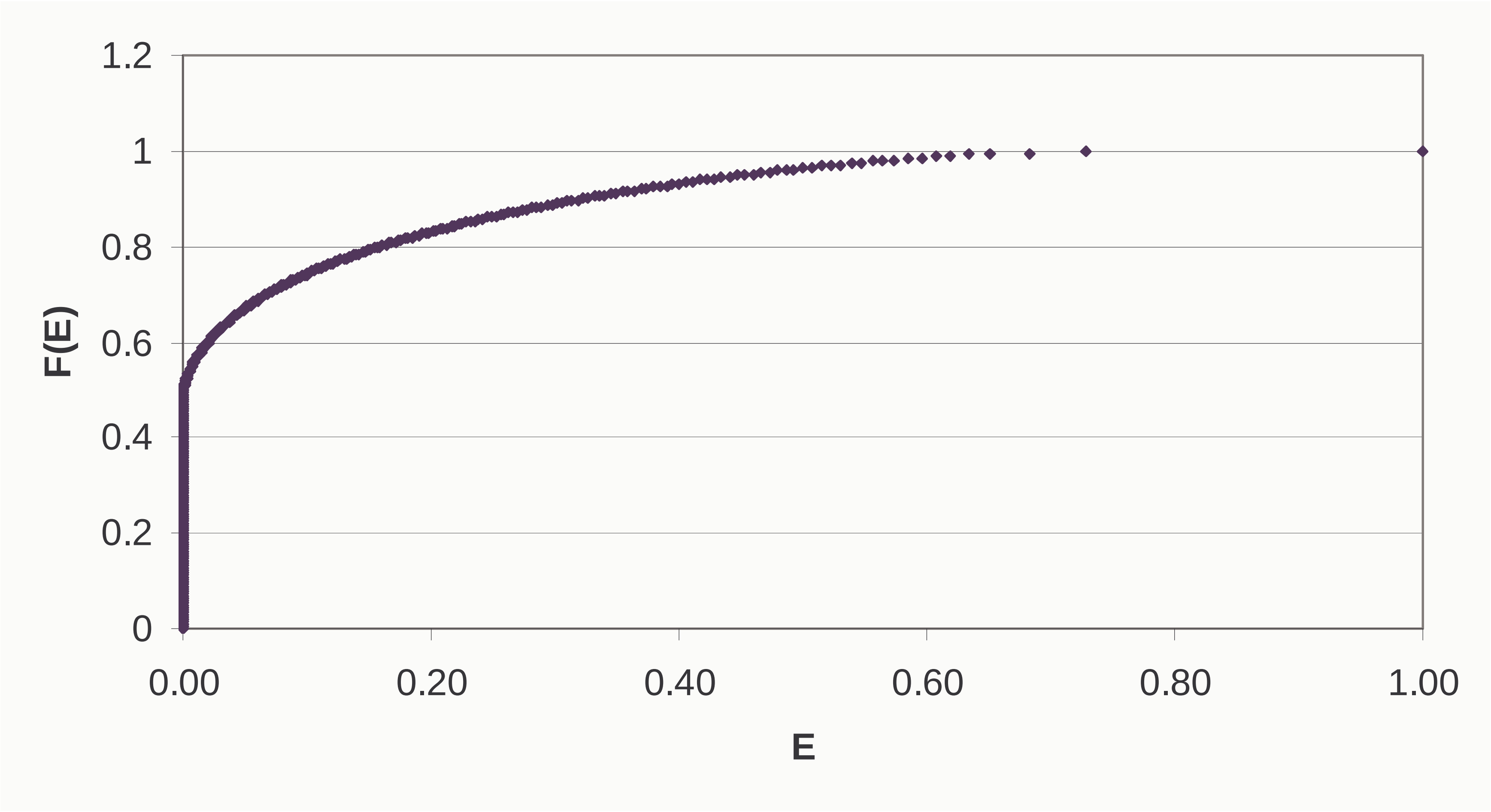
$$F\left(E\right)\equiv\frac{\#\{k:\lambda_k/N\le E\}}{N}\qquad F(E)\ \text{is increasing,}\quad F\left(1\right)=1$$

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

One way to think about the DOE 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 DOE



In the DOE language…

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

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

$$\frac{\partial U(E,g)}{\partial E} = E\ f(E) - gf(E) = (E-g)f(E)$$

If $f(g) \neq 0$, then $E^*(g) = g.$

Dependence of the problem on g

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

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

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

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

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

According to this calculation, the best cutoff is the level E
where the DOE vanishes (or nearly vanishes) coming from the right.

A closer look at equities

-- There is information in equities markets related to different activities of listed companies

-- Industry sectors

-- Market capitalization

-- Regression on industry sector indexes explain often no more than 50% of returns

-- Since there exist at least 15 distinct sectors that we can identify in US/ G7 economies, we conclude that we probably require **at least 15 factors** to explain asset returns.

-- Temporal market fluctuations are important as well. In order for factor models to be useful, they need to adapt to economic cycles.

Stocks of more than 1BB cap in 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
Utilities	**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	**XLI**	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	**XLV**	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**

January, 2007