Risk and Portfolio Management with Econometrics

Marco Avellaneda

G63.2751.001

NYU Spring Semester 2010

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 forecast 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 <u>statistical models</u>

Factor models

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

$F_j, \;\; j = 1, \ldots, N_f,$	Explanatory factors
$\beta_j, \;\; j = 1, \ldots, N_f,$	Factor <u>loadings</u>
$\sum_{j=1}^{N_f} \beta_j F_j$	<u>Explained</u>, or <u>systematic</u> portion
ε	<u>Residual,</u> or <u>idiosyncratic</u> 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: $< \varepsilon >= 0 \;\; \text{or} \;\; < R >= \beta < F >$

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): $< \varepsilon >= 0$ 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)

What do we actually know about the statistics of stock returns?

Stock returns exhibit <u>heavy tails:</u>

-- Small moves are more frequent (likely) than predicted by Gaussian PDF

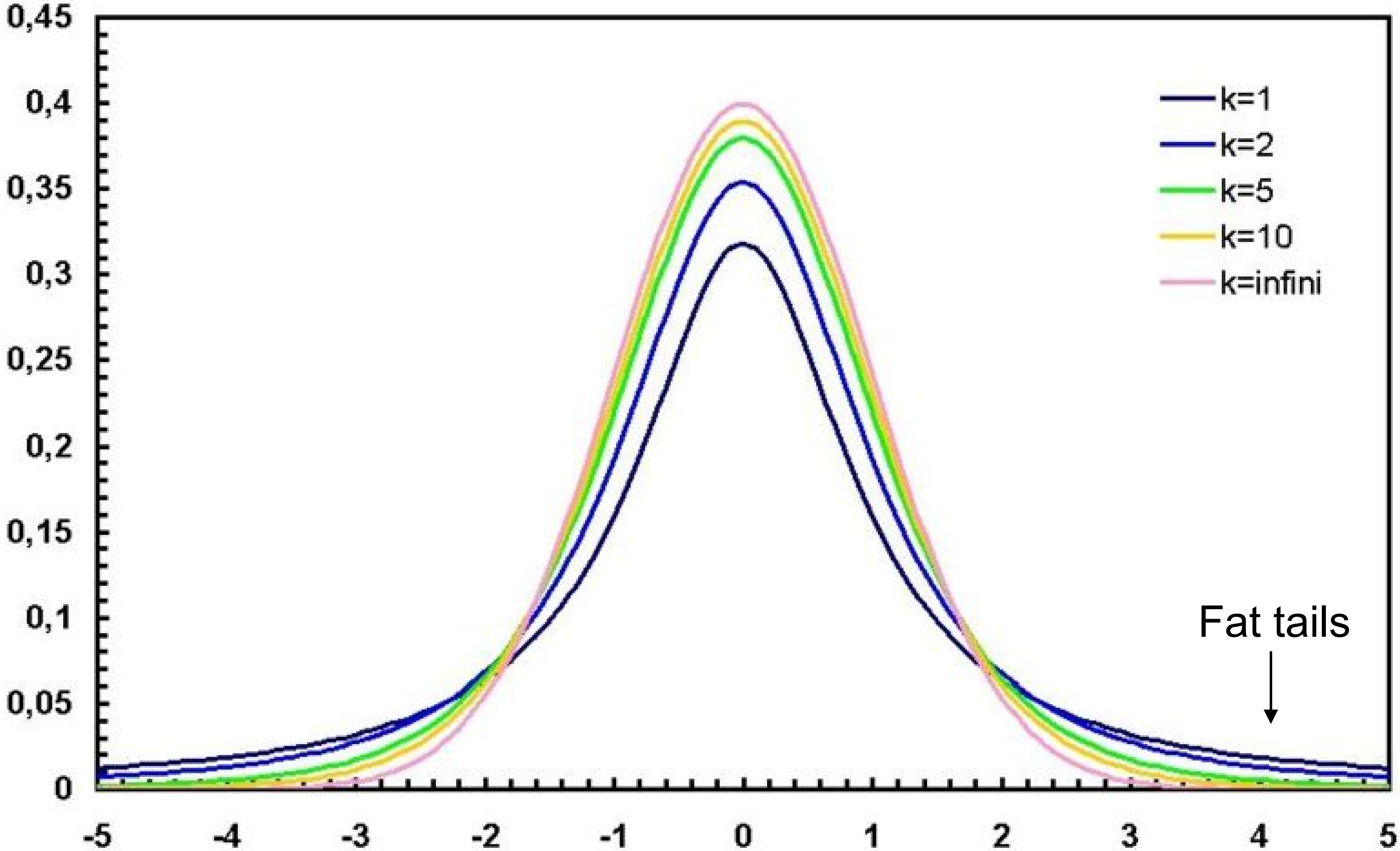
-- Large moves are more likely than predicted by Gaussian distribution

-- Consider a large cross-section of the US stock market (~3000 stocks)

-- Data consists of 1 year worth of data on ~ 3000 stocks

-- Standardized stock returns over T days (e.g. T=1) and fit to various probability distributions

The Student-t family of distributions



``Black swan''

``Orange swan''

Statistical model with fat tails to account for extreme moves

$$f(x) = \frac{C}{\left(1 + \frac{x^2}{k}\right)^{\frac{1+k}{2}}}$$

Student: Power-law tails

$$P\{X > r\} \sim 1/r^k$$

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Gauss: Exponential tails

(Gauss ~ k=infinity)

As a general rule, Gaussian or normal distributions are not suitable for financial data due to fat tails

QQ-plots

Generate a sample from the unknown distribution and <u>sort it</u> in increasing order

$$X_1, X_2, \ldots, X_N$$

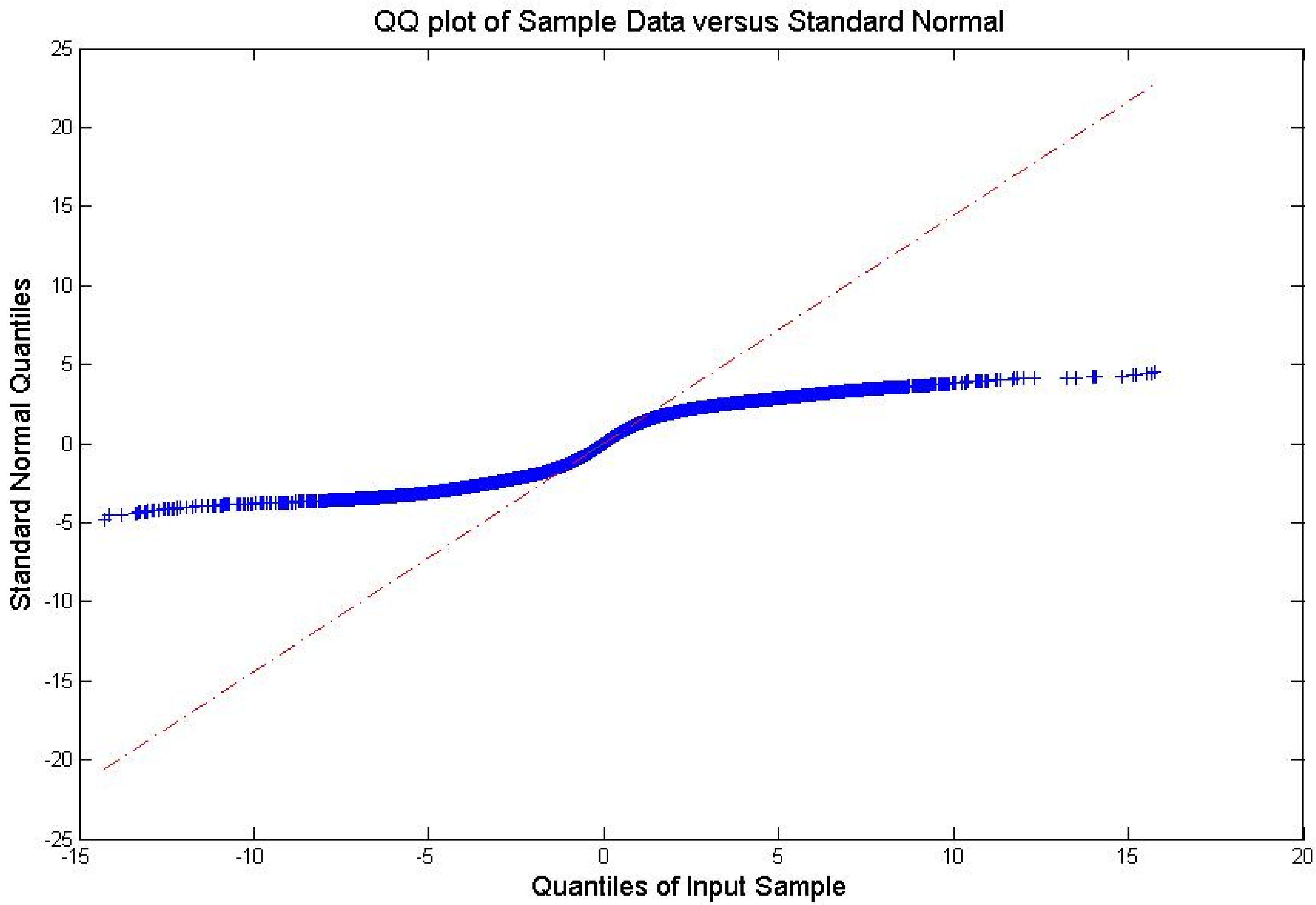
Generate a vector of a known distribution (e.g. Student t)

$$Y_1 = F_\alpha^{-1}\left(\frac{1}{N}\right), \ldots, Y_k = F_\alpha^{-1}\left(\frac{k}{N}\right), \ldots, Y_N = F_\alpha^{-1}(1) = \infty$$

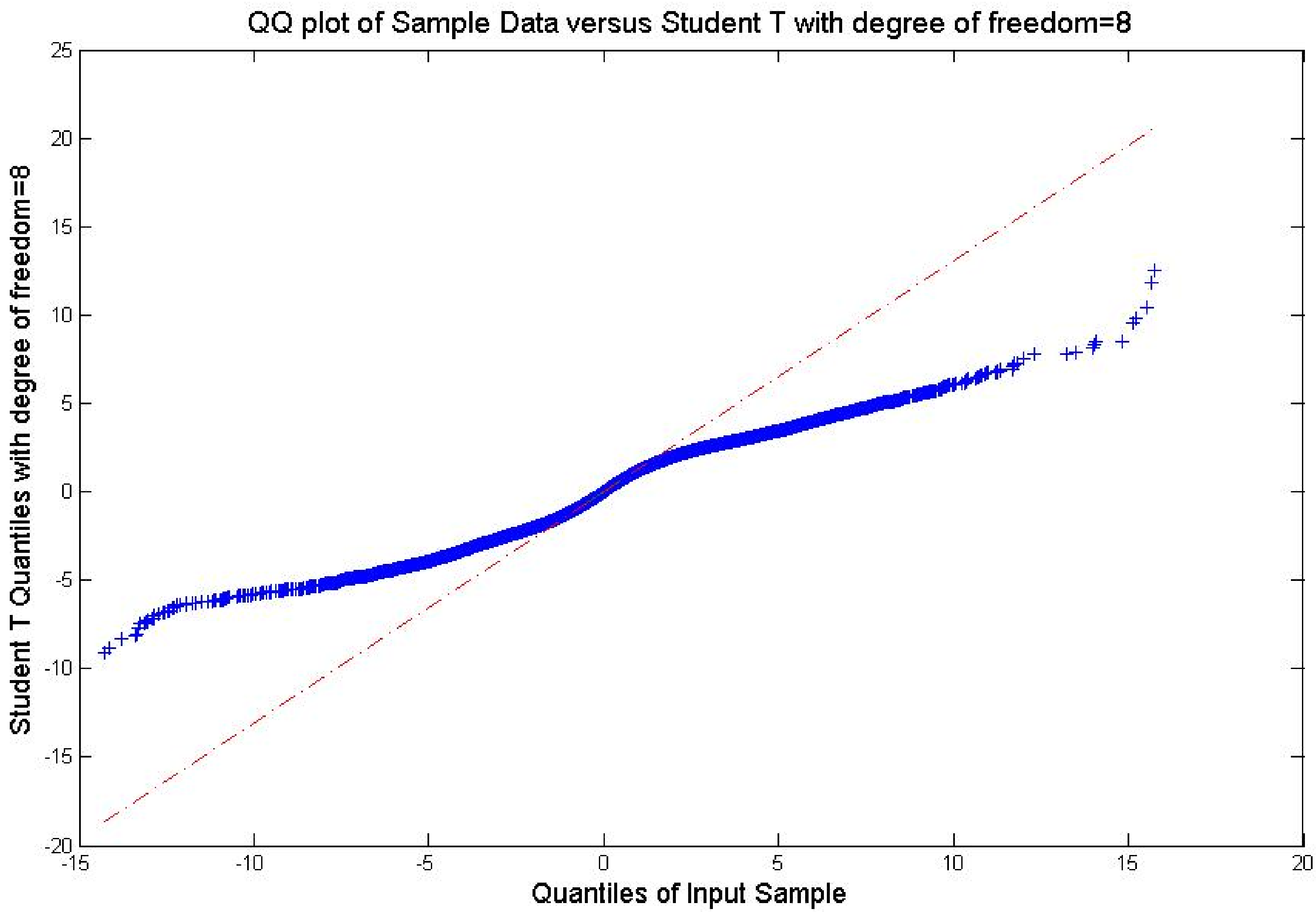
Draw an X-Y plot of the ``data'' $(X_k, Y_k)_{k=1}^{N-1}$

If the sample is drawn from the known distribution, then the points fall approximately on the X=Y line.

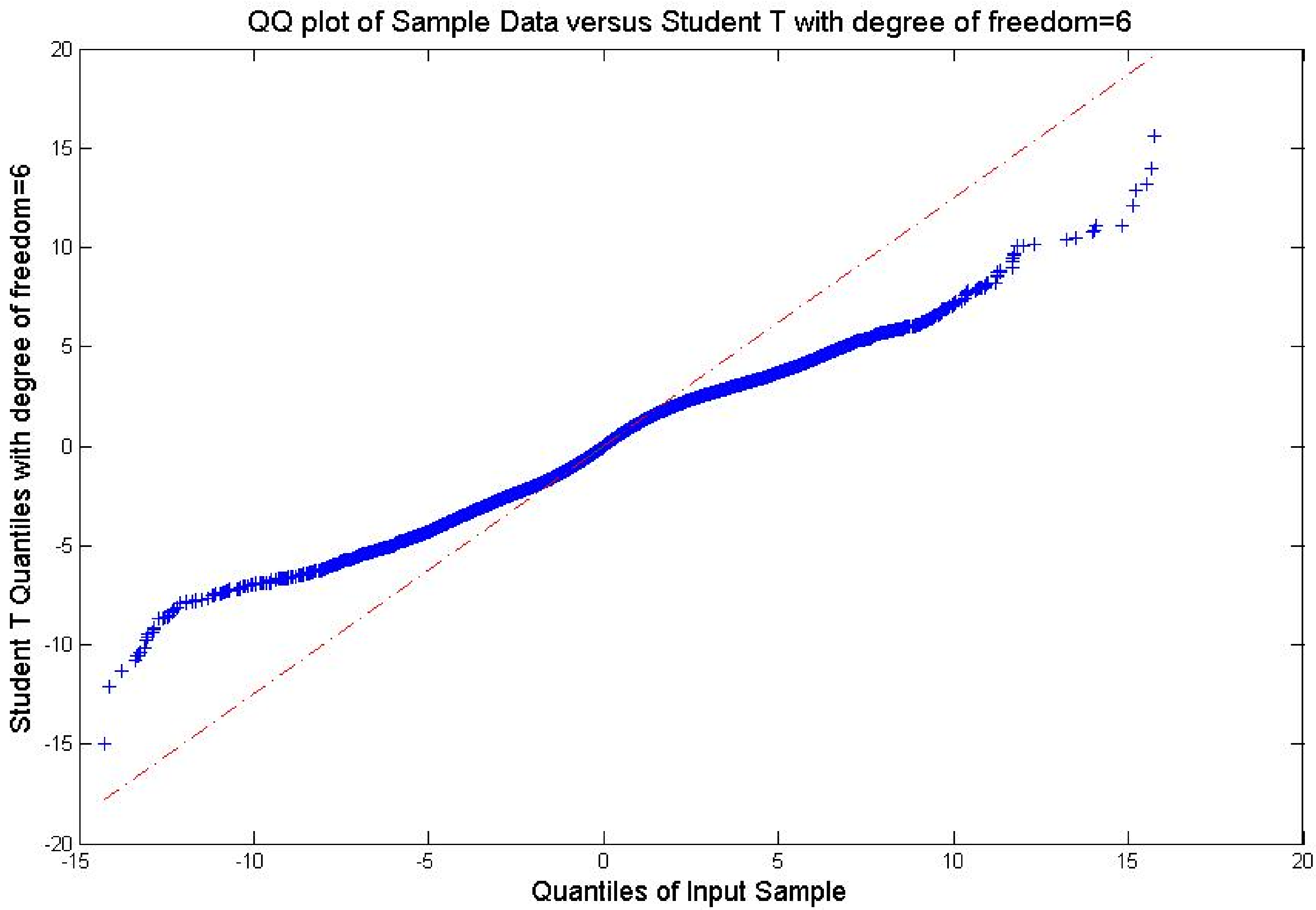
Tails (extreme values) for stock movements from T to T+2:Gaussian



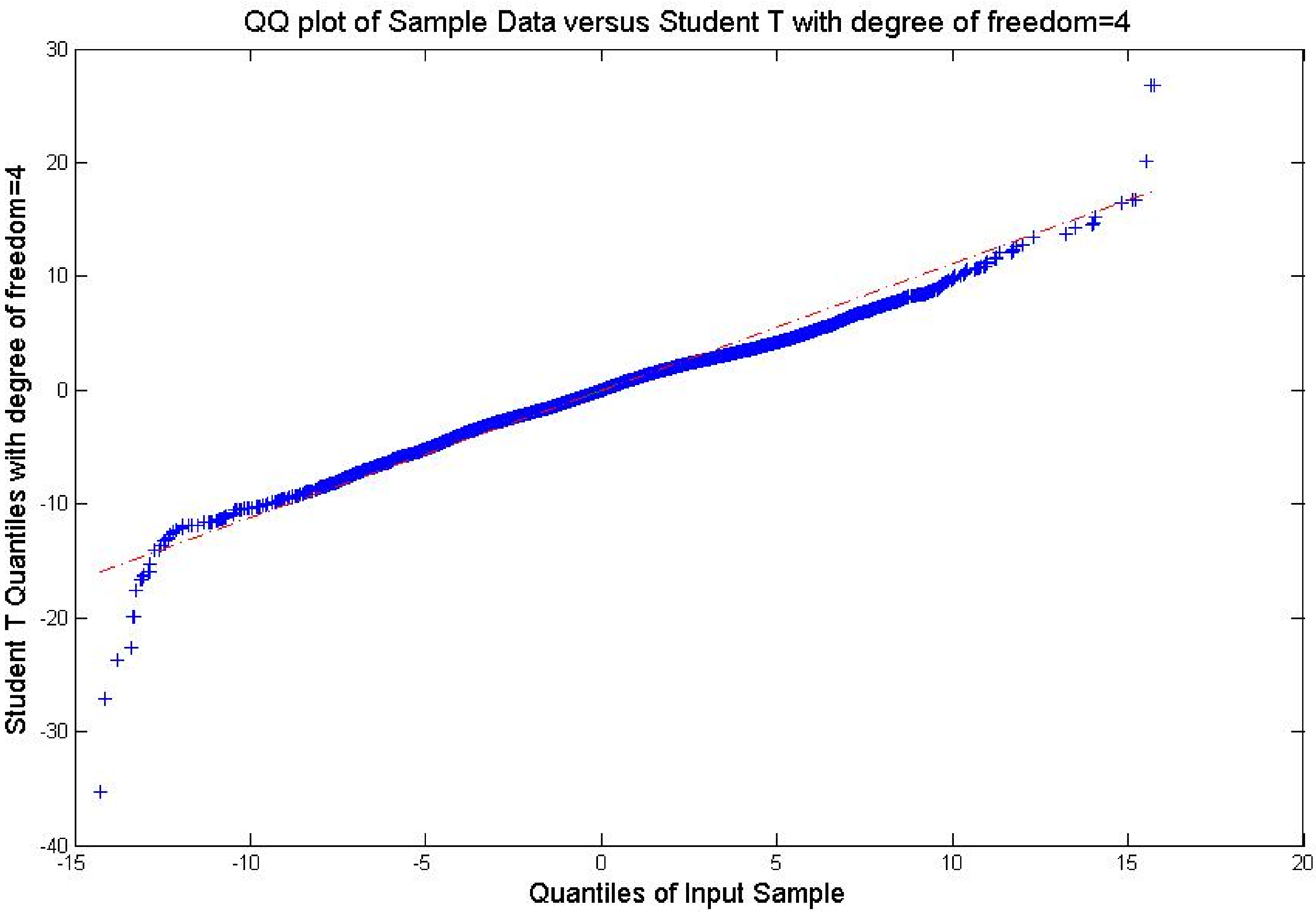
Student-t with 8 df



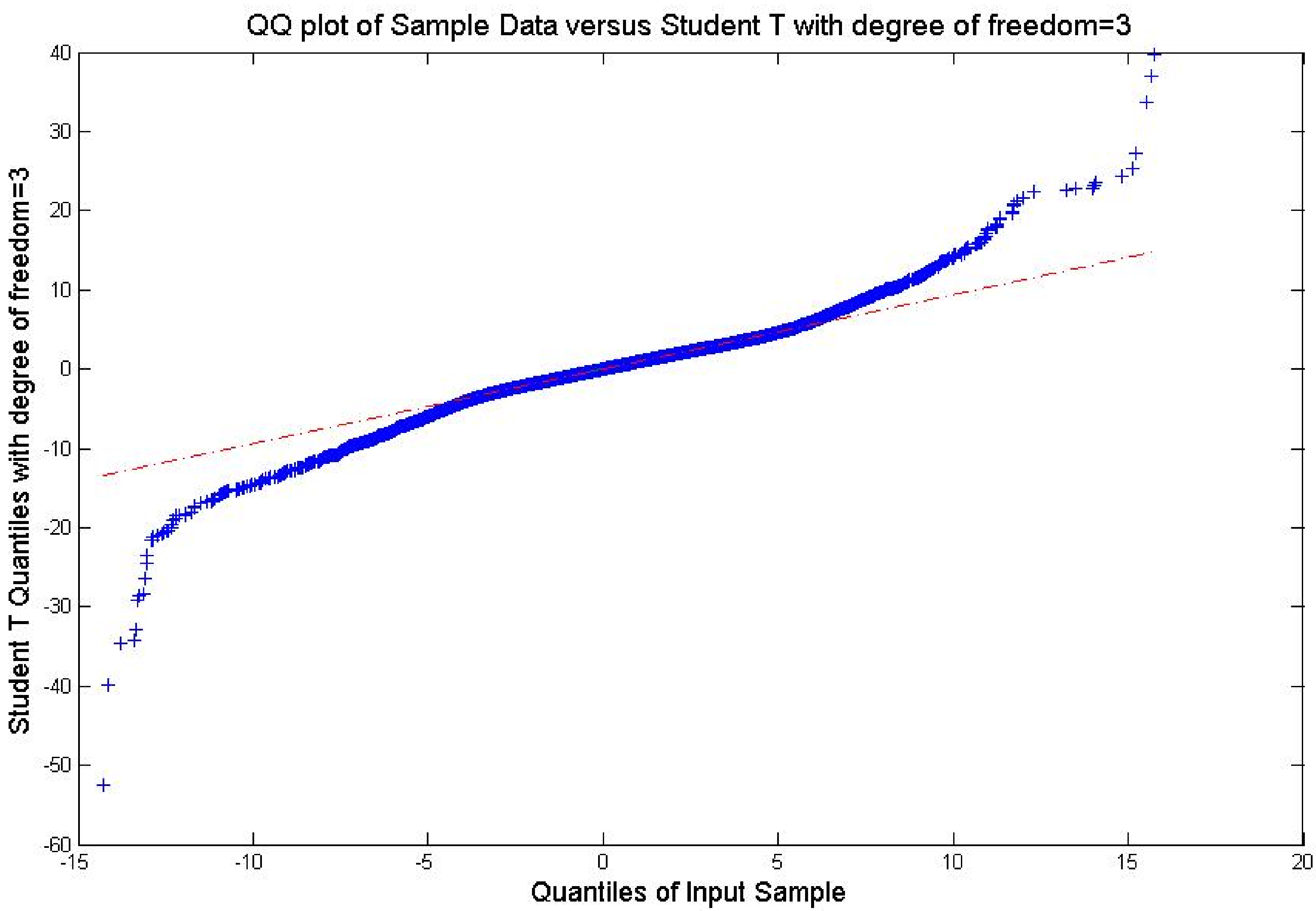
Student-t with 6 df



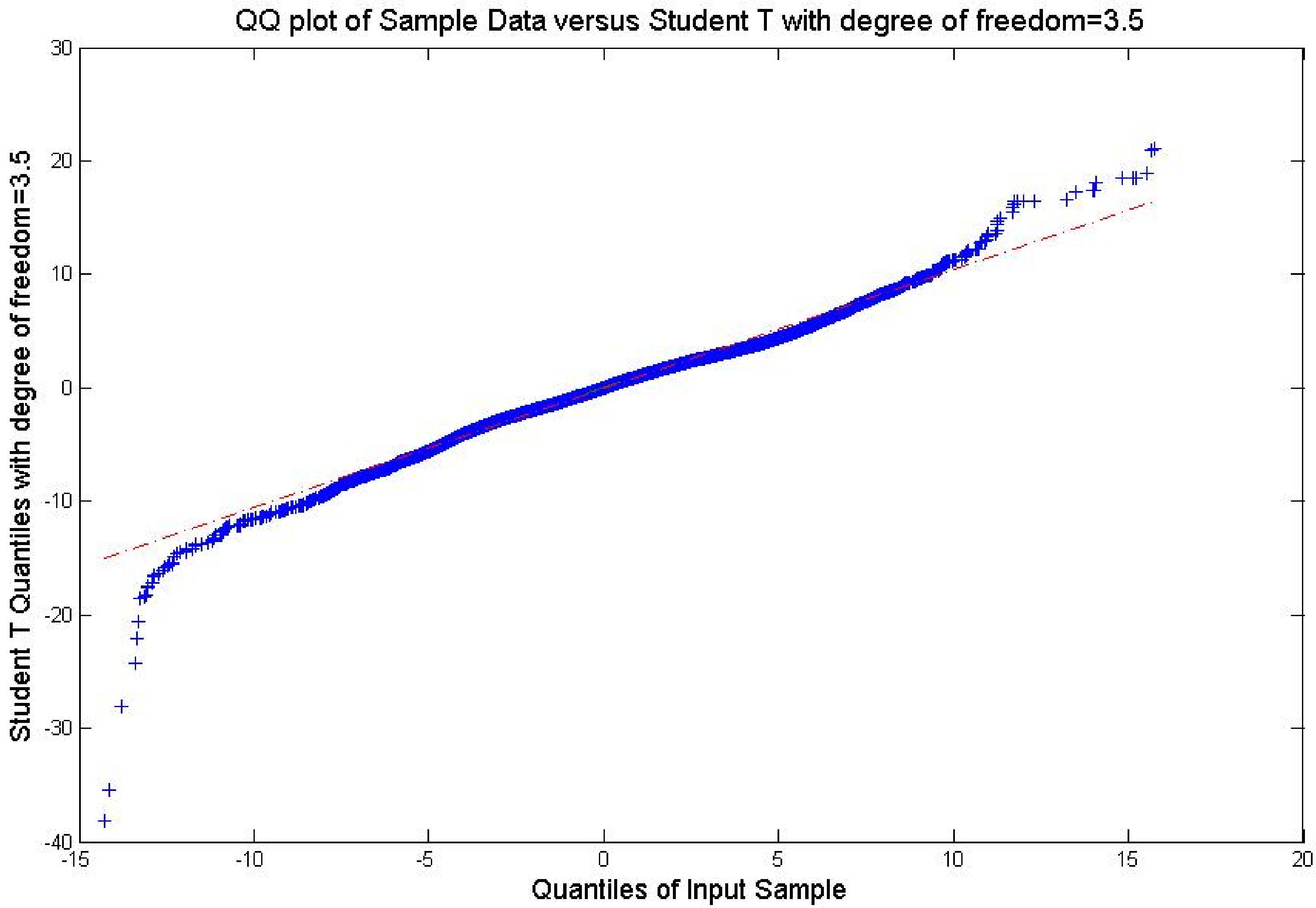
Student-t with 4 df



Student-t with 3 df



Student-t with 3.5 df

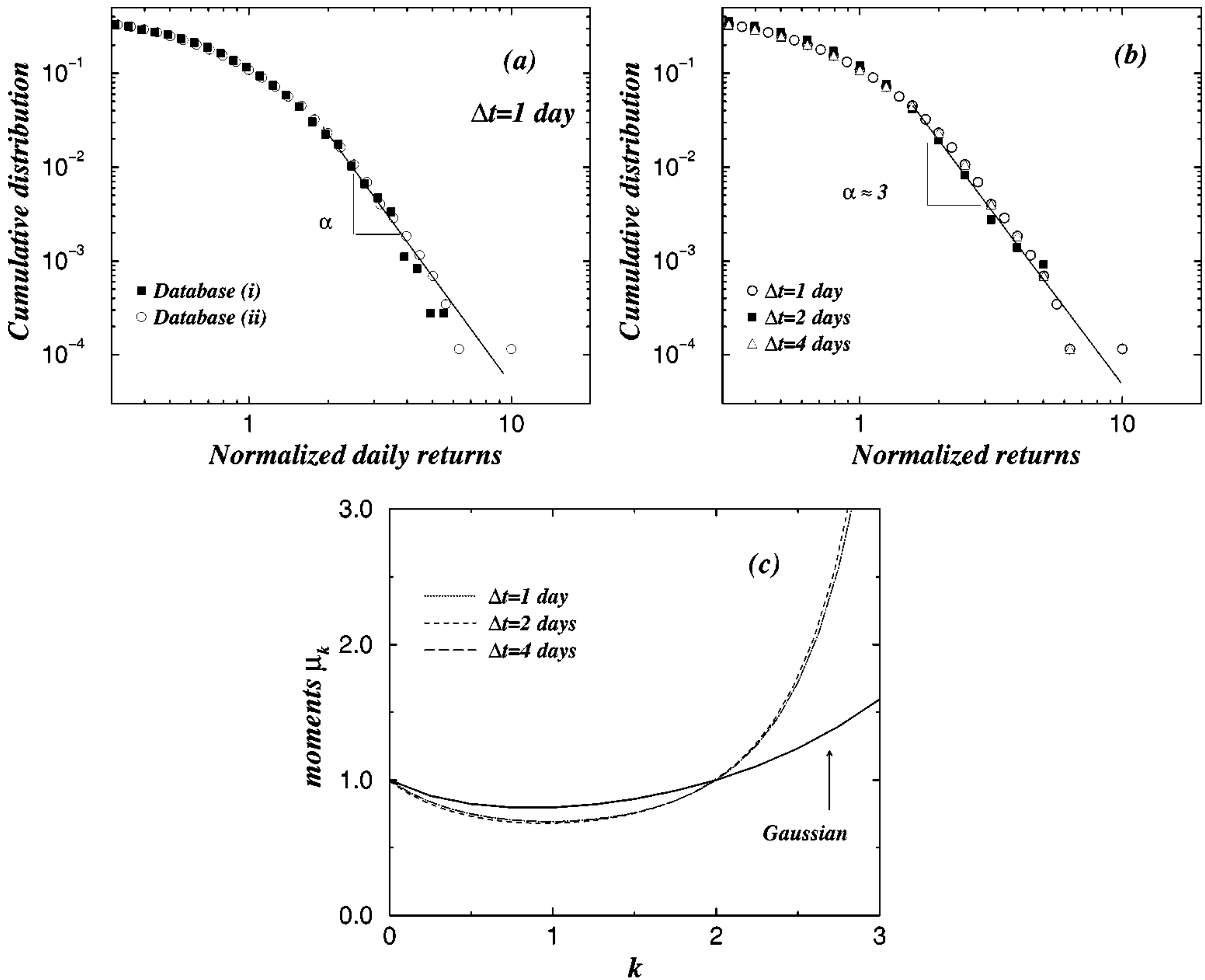


Consistent with classical result from Econophysics

Phys. Rev. E 60, 5305–5316
Scaling of the distribution of fluctuations of financial market indices

Parameswaran Gopikrishnan, Vasiliki Plerou, Luís A. Nunes Amaral, Martin Meyer, and H. Eugene Stanley

This paper studies equity returns over different time-horizons (1 minute- 1 day) and finds scaling behavior in the probability of large moves (up or down).



(a)
Δt=1 day
Cumulative distribution
α
Database (i)
Database (ii)
Normalized daily returns
(b)
Cumulative distribution
α ≈ 3
Δt=1 day
Δt=2 days
Δt=4 days
Normalized returns
(c)
Δt=1 day
Δt=2 days
Δt=4 days
moments μ_k
Gaussian
k

QQ-plots

Generate a sample from the unknown distribution and <u>sort it</u> in increasing order

$$X_1, X_2, \ldots, X_N$$

Generate a vector of a known distribution (e.g. Student t)

$$Y_1 = F_\alpha^{-1}\left(\frac{1}{N}\right), \ldots, Y_k = F_\alpha^{-1}\left(\frac{k}{N}\right), \ldots, Y_N = F_\alpha^{-1}(1) = \infty$$

Draw an X-Y plot of the ``data'' $\left(X_k, Y_k\right)_{k=1}^{N-1}$

If the sample is drawn from the known distribution, then the points fall approximately on the X=Y line.

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

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

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

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

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

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 $T \ll N$.

Eigenvalues, Eigenvectors and Eigenportfolios

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

eigenvalues

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

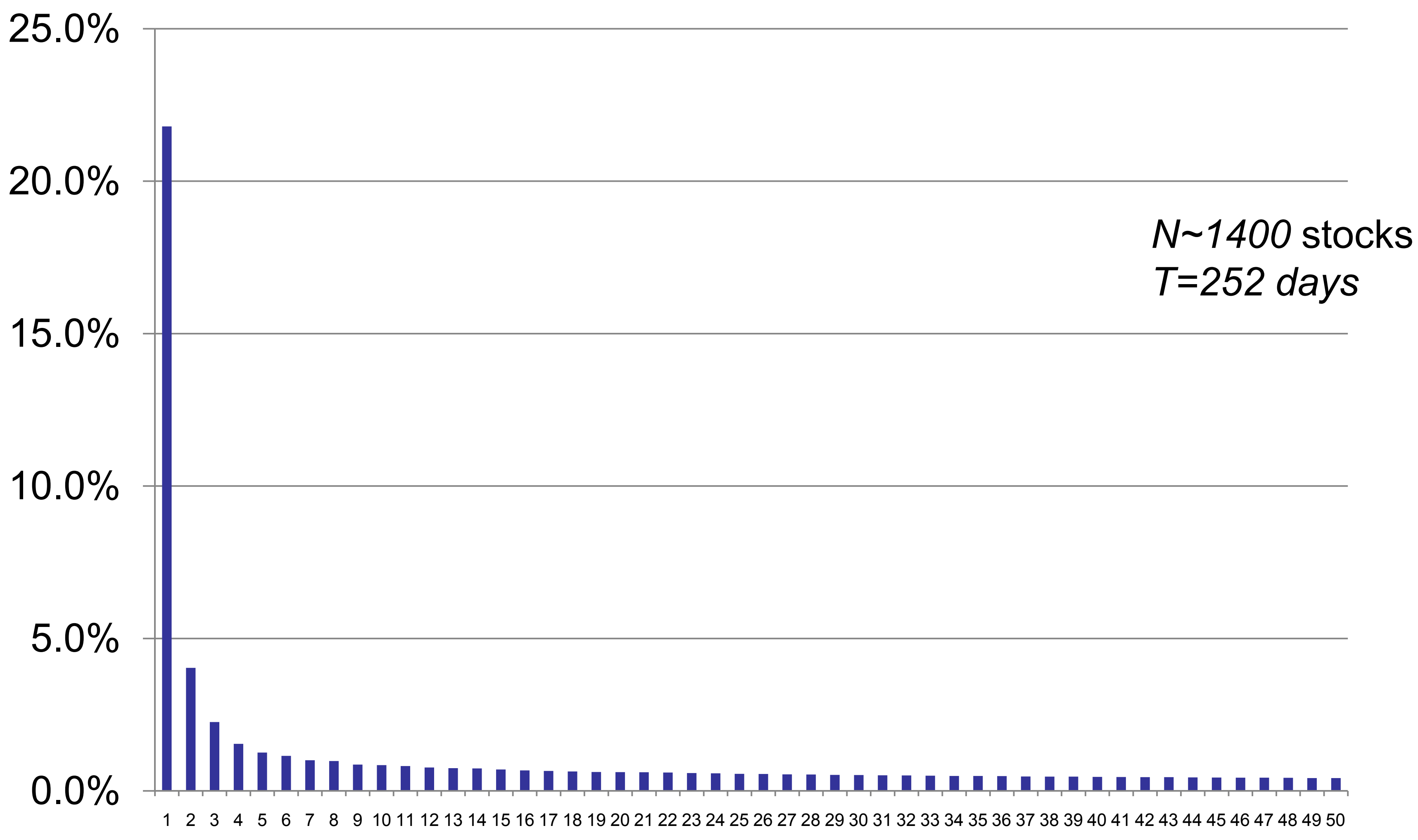
eigenvectors

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

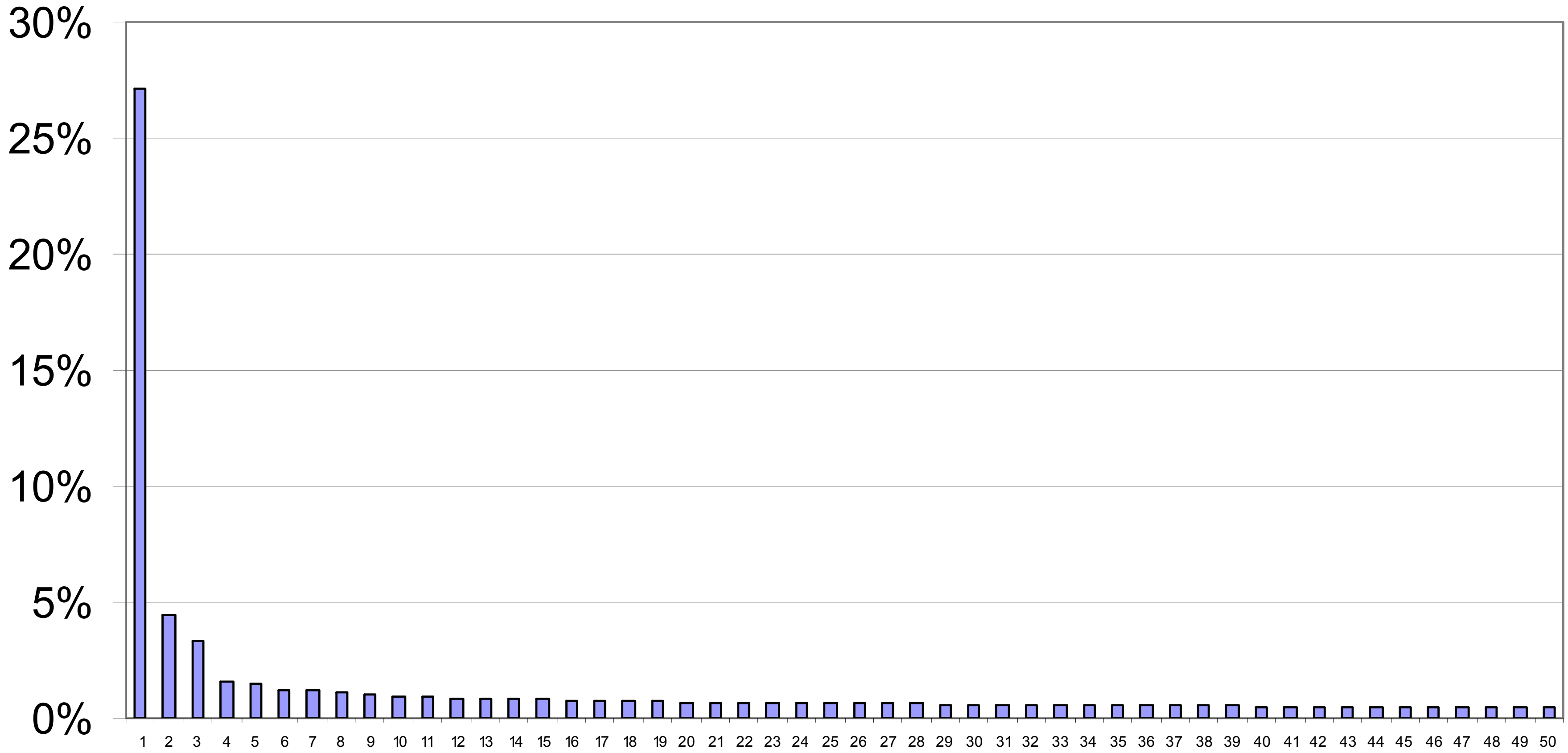
returns of "eigenportfolios"

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)



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

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

Covariance matrix

$$\Omega = \mathbf{C}^{1/2}$$

Symmetric square root of C

$$\mathbf{F} = \Omega^{-1}\mathbf{R}, \qquad \mathbf{R} = \Omega\mathbf{F}$$

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 $N,T \to \infty$

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.

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.