Risk and Portfolio Management
NYU Spring Semester 2013

3. Correlations and factor analysis

Marco Avellaneda

Explaining co-movements in stocks via factor analysis

Separate the **systematic** components of stock returns from the company-specific, or **idiosyncratic** components

$$R_i = \beta_i R_{Mkt} + \varepsilon_i$$

Project returns on single Market Factor (CAPM)

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \varepsilon_i$$

Project returns on Multiple (sector, size) Factors (APT)

In principle, market-neutral portfolios should have no exposure to market factors ("defactoring")

Defactoring: the Correlation Matrix Approach

$$R_{it} = \text{ daily stock returns in panel form}$$

$$i = 1,\ldots,N, \;\; t = 1,\ldots,T$$

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

$$\overline{\rho}_{ij} = \frac{1}{T-1}\sum_i \frac{\left(R_{it} - \overline{R_i}\right)\left(R_{jt} - \overline{R_j}\right)}{\overline{\sigma_i}\,\overline{\sigma_j}}$$

Principal Component Analysis

$$\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_N$$

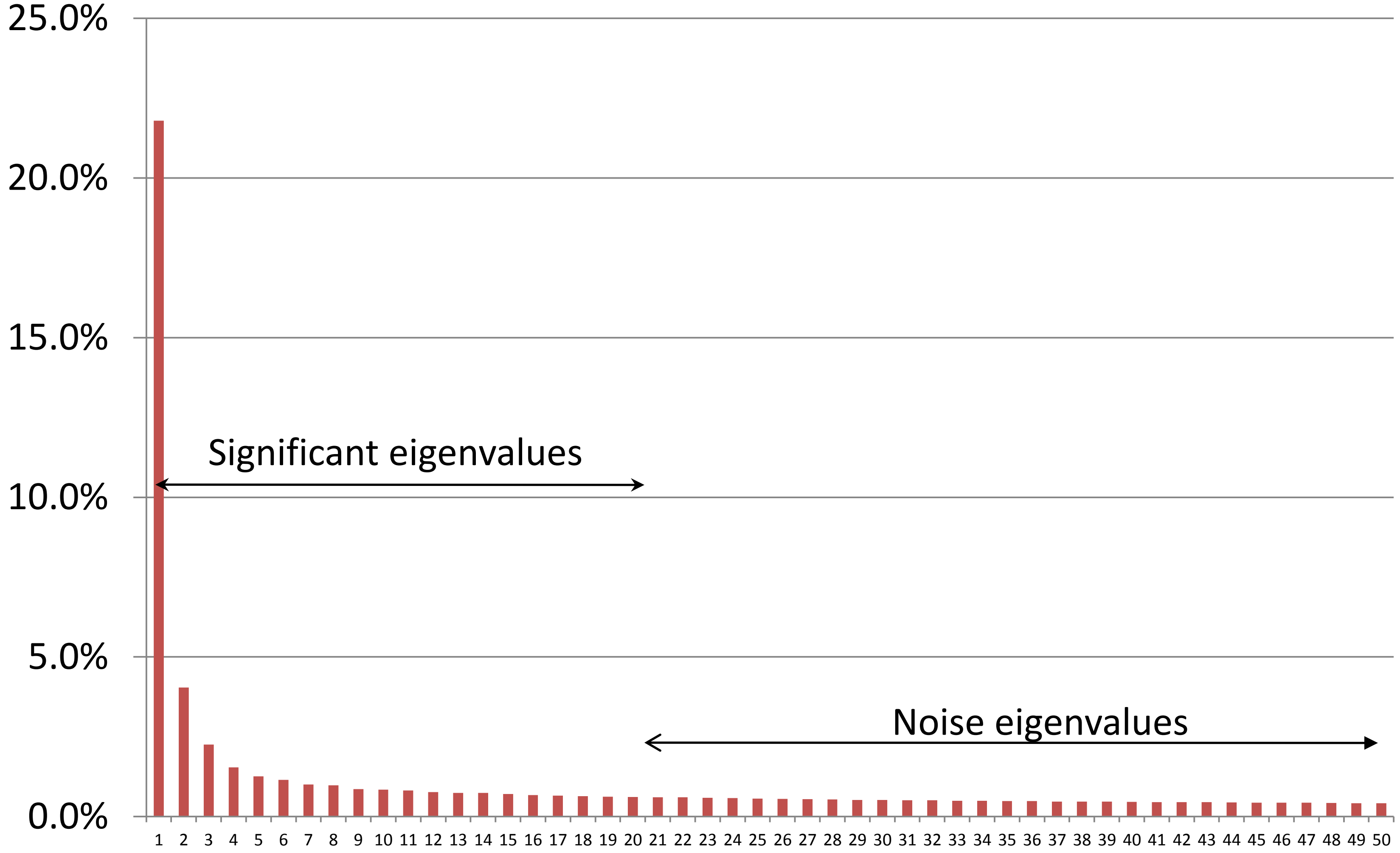
Eigenvalues are all non-negative

$$V^{(j)} = \left(v_1^{(j)}, v_2^{(j)}, \ldots, v_N^{(j)}\right)$$

Orthogonal eigenvectors

Stock market fluctuations can be characterized as moves along the eigenvector directions. We seek to extract mathematical factors from the PCA analysis.

PCA: Explained variance from the viewpoint of eigenvalues



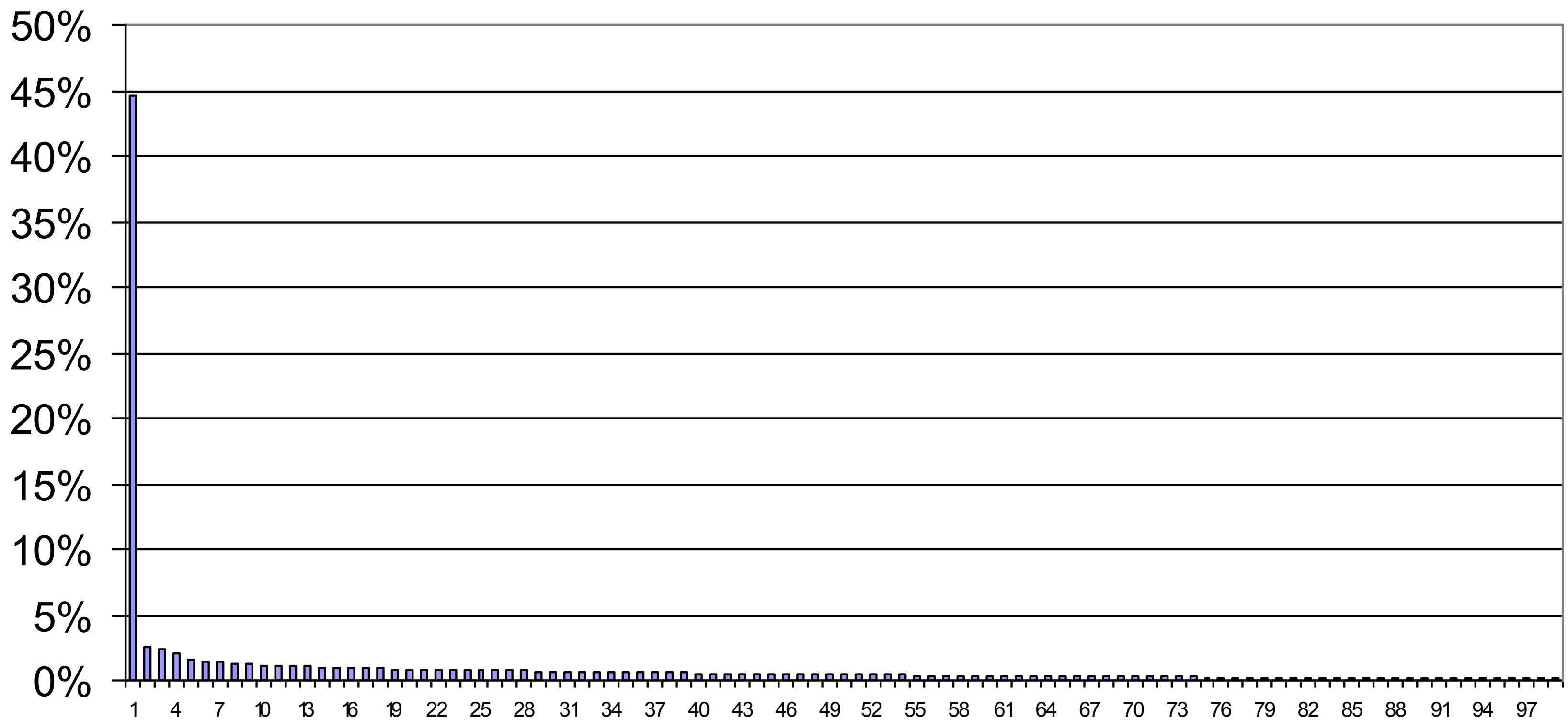
Big universe: Jan 2007-Dec 2007

Nasdaq-100
Components of NDX/QQQQ

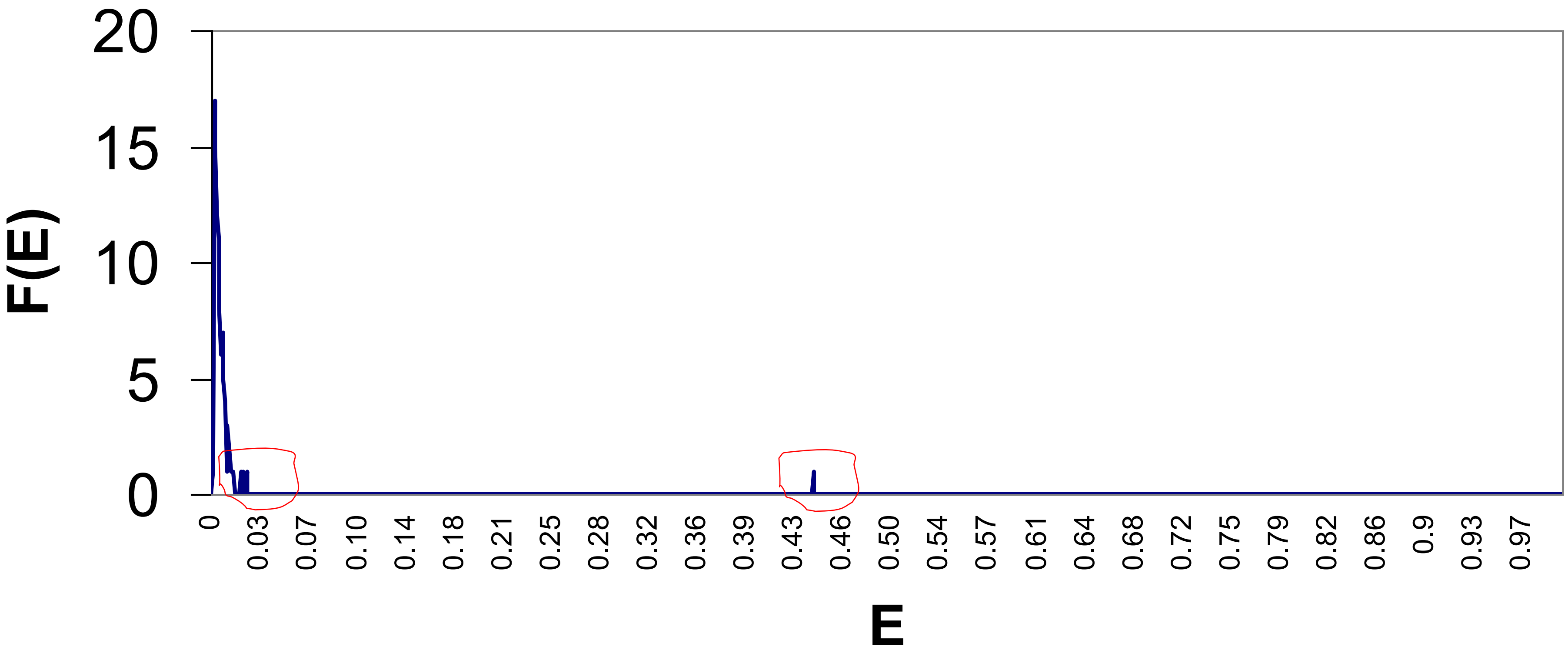
Data: Jan 30, 2007 to Jan 23, 2009

502 dates, 501 periods

99 Stocks (1 removed) MNST (Monster.com), now listed in NYSE

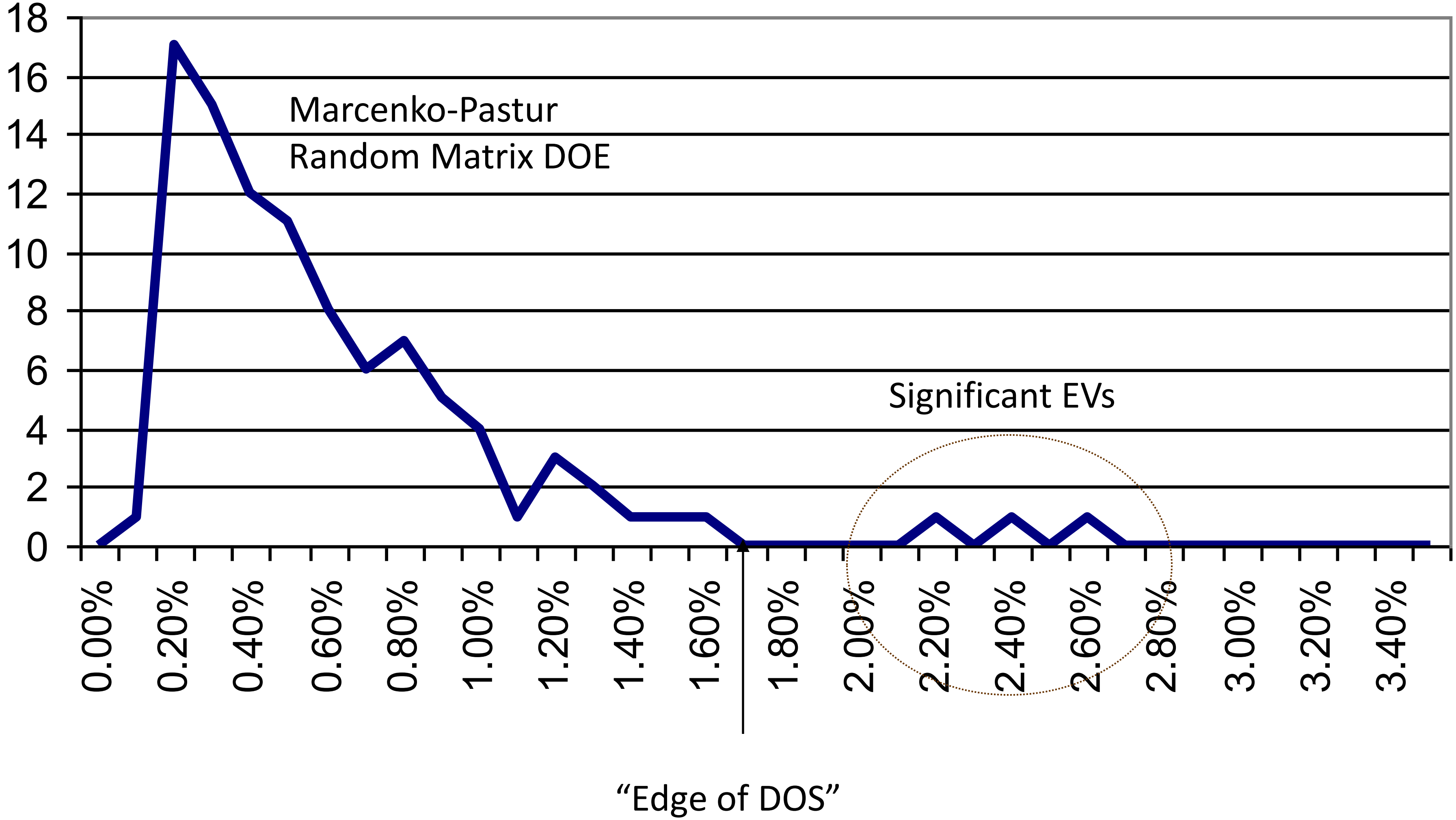


Density of States (from previous data)



One large mass at 0.44,
Some masses near 0.025
Nearly continuous density for lower levels

Zoom of the DOS for low eigenvalues



De-noising the correlation matrix

A reasonable assumption is that the empirical correlation matrix consists of a few ``significant'' eigenvalues/eigenvectors and other eigenvalues/eigenvectors

The latter come from either idiosyncratic risk (specific risk) or from estimation noise.

The idea is to think of the correlation matrix as

$$R = \underbrace{\sum_{k=1}^{n} \lambda_k v^{(k)} \otimes v^{(k)}}_{\text{significant}} + \underbrace{\sum_{k=m+1}^{N} \lambda_k v^{(k)} \otimes v^{(k)}}_{\text{Random matrix}}$$

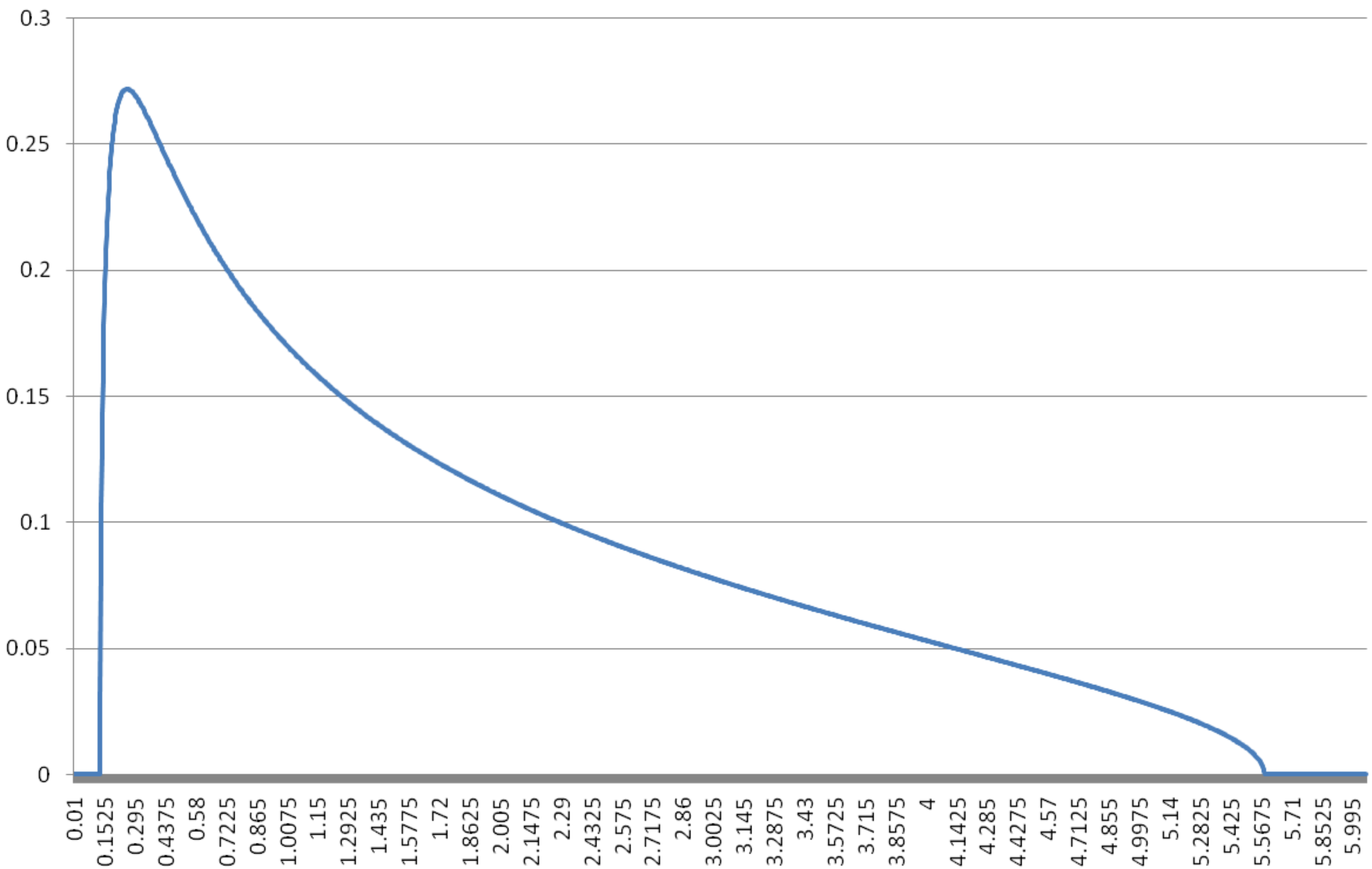
Marcenko-Pastur Distribution for the DOS of a Random Correlation Matrix

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.

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

$$MP(\lambda) = \left(1-\frac{1}{\gamma}\right)^+ \delta(\lambda) + \frac{1}{2\pi\gamma}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}$$

Marcenko-Pastur Distribution
gamma=500/269= 1.858736



Factors & Eigenportfolios

For each eigenvector, build a portfolio which is weighted proportionally to the coefficient of each stock and inversely proportionally to its volatility

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\overline{\sigma_i}}$$

Portfolio weight of j-th eigenportfolio

$$F_j = \sum_{i=1}^{N} Q_i^{(j)} R_i = \sum_{i=1}^{N} \left(\frac{v_i^{(j)}}{\overline{\sigma_i}} \right) R_i$$

J-th factor is the return of the j-th eigenportfolio

How many eigenportfolios are significant?

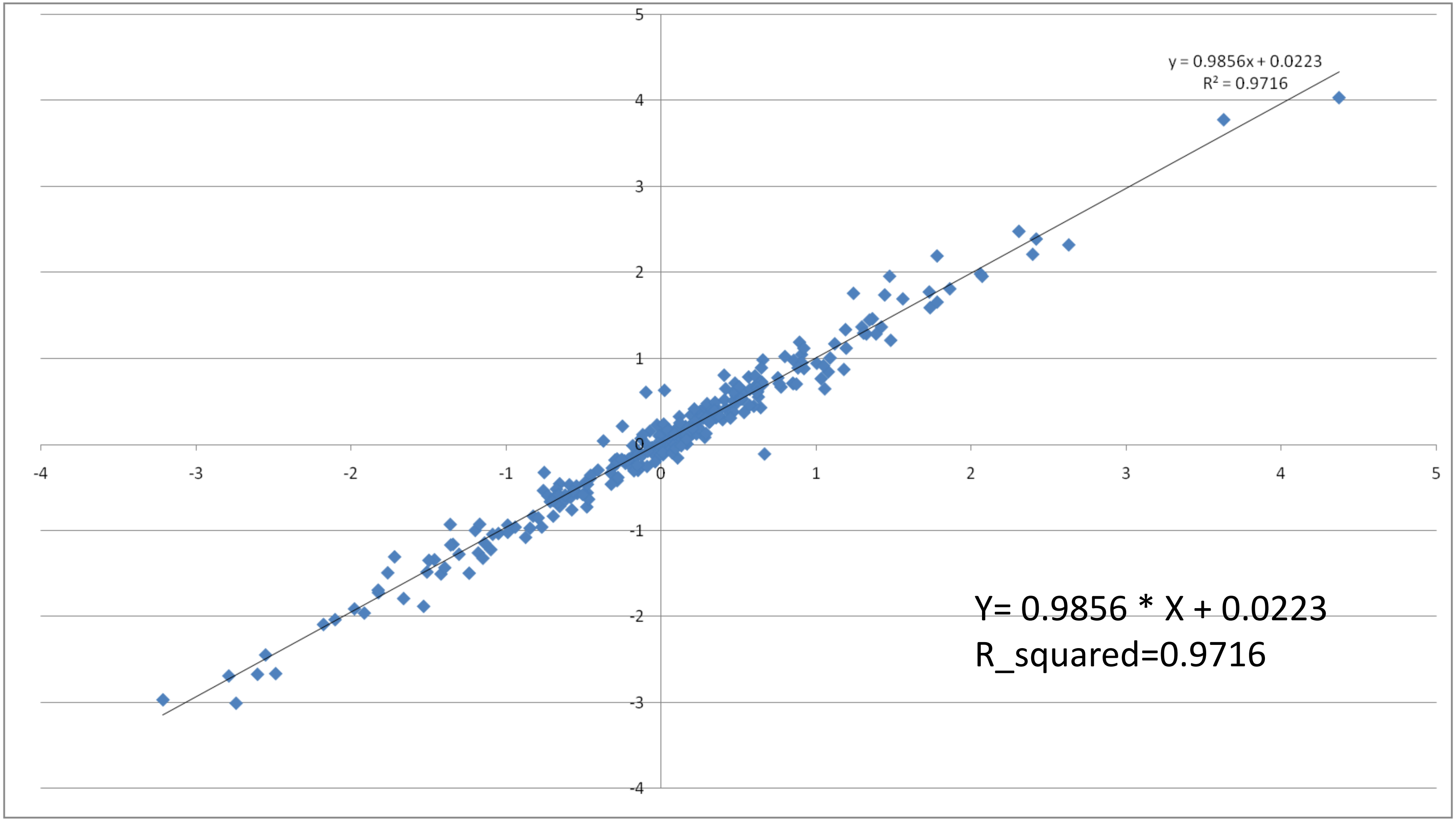
- Perform PCA on the empirical correlation matrix with 1 year moving window

- Consider the correlation matrix of the residuals after removing 1, 2 ,3… eigenportfolios

- Compare the DOS of the correlation of the residuals with the spectrum of the correlation matrix of pure noise (Marcenko-Pastur)

-The number of significant factors corresponds to the **first m for which the matrix of residuals is close to MP** (e.g. in the sense of hypothesis testing)

First eigenportfolio returns compared with S&P 500 returns (1/5/2009 to 1/29/2010)



Capital Asset Pricing Model (Lintner, Sharpe, 1965)

The capital asset-pricing model is a one-factor model to explain stock returns and stock prices.

$$\boxed{\tilde{R}_i = \beta_i \tilde{F}_1 + \varepsilon_i, \qquad \langle \varepsilon_i \rangle = 0,\ \langle \varepsilon_i \varepsilon_j \rangle = 0,\ \text{if } i \neq j}$$

$$\tilde{F}_1 = \text{ return of the market portfolio (cap - weighted market portfolio)}$$

$$E\left(\tilde{R}_i\right) = \beta_i E\left(\tilde{F}_1\right), \qquad \beta_i = \frac{Cov\left(R_i, F_1\right)}{Var\left(F_M\right)} = \frac{\sigma_i \rho_i}{\sigma_M}$$

Arbitrage Pricing Theory (Ross, 1971)

Assumptions:

•There are N stocks
•m <u>tradable assets</u> as factors (e.g. ETFs or baskets, or portfolios)
• Linear regression of stock returns on factors has <u>uncorrelated residuals</u> (uncorrelated with the factors and among each other)

$$\tilde{R}_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} \tilde{F}_k + \varepsilon_i$$

$$\tilde{R}_i = R_i - r\Delta t = \text{ return on stock } i, \text{ financed}$$

$$\tilde{F}_k = F_k - r\Delta t = \text{return on factor } k, \text{financed}$$

$$\varepsilon_i = \text{residual of the linear regression of stock ret. on factor returns}$$

$$\alpha_i = \text{excess returns}$$

Testing APT with data from Jan 5, 2009 to Jan 29 2010

Using daily returns from Jan 5 2009 to Jan 29, 2010 for the components of the S&P 500 index, we explored the ``partition'' of the eigenvectors/eigenvalues into significant and noise components.

We did this by testing for $\alpha = 0$ and for the correlations of residuals.

One way to analyze the correlations of residuals is by doing a PCA again and analyzing the corresponding eigenvalues and DOS.

The assumption that residuals are uncorrelated allows comparison with theoretical eigenvalue distributions such as Marcenko-Pastur.

One Factor Model (CAPM)

1. Compute correlation matrix of S&P 500

2. Compute the first eigenportfolio

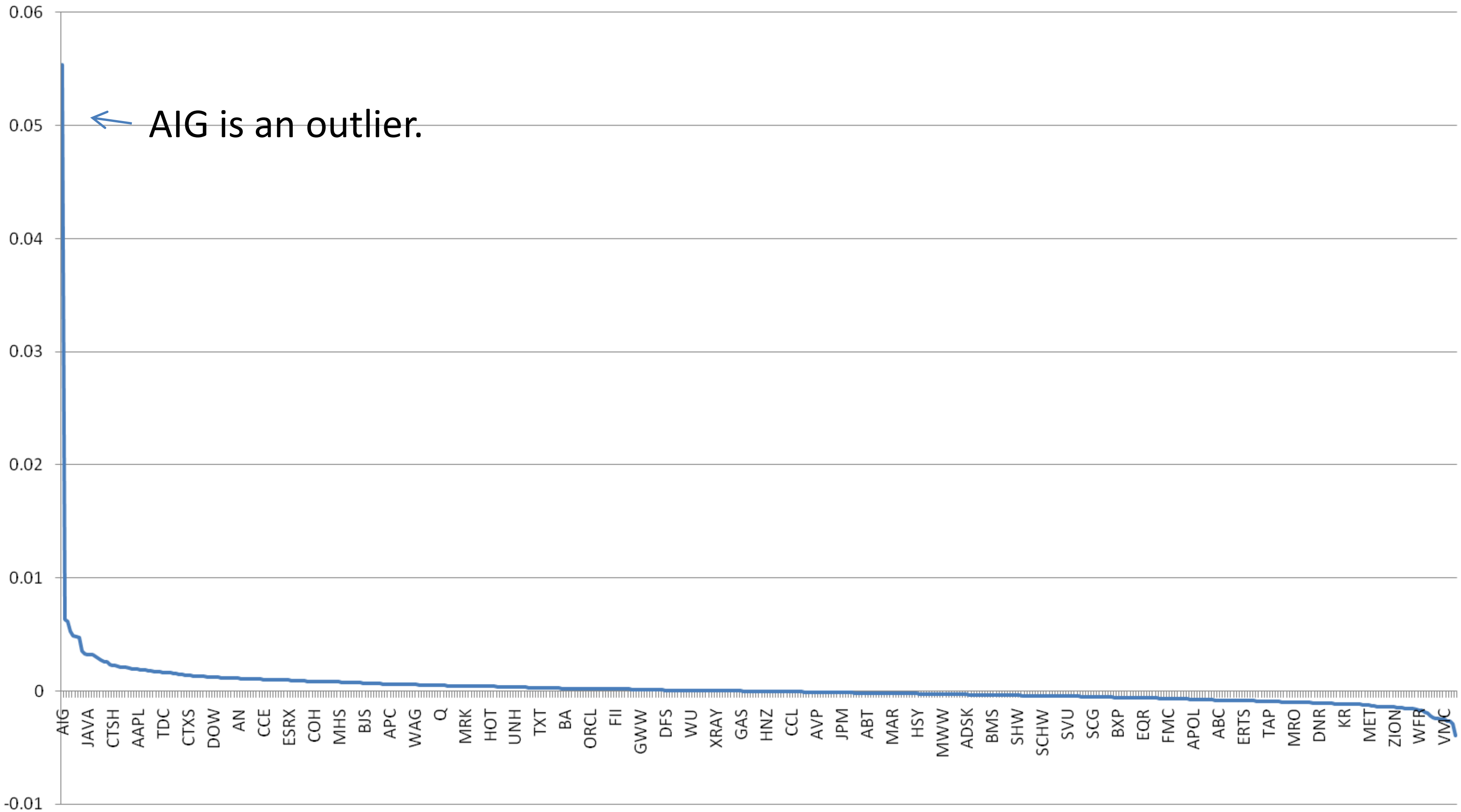
3. Compute residuals for all 500 stocks by regression

$$\tilde{R}_i = \alpha_i + \beta_i \tilde{F}_1 + \varepsilon_i$$

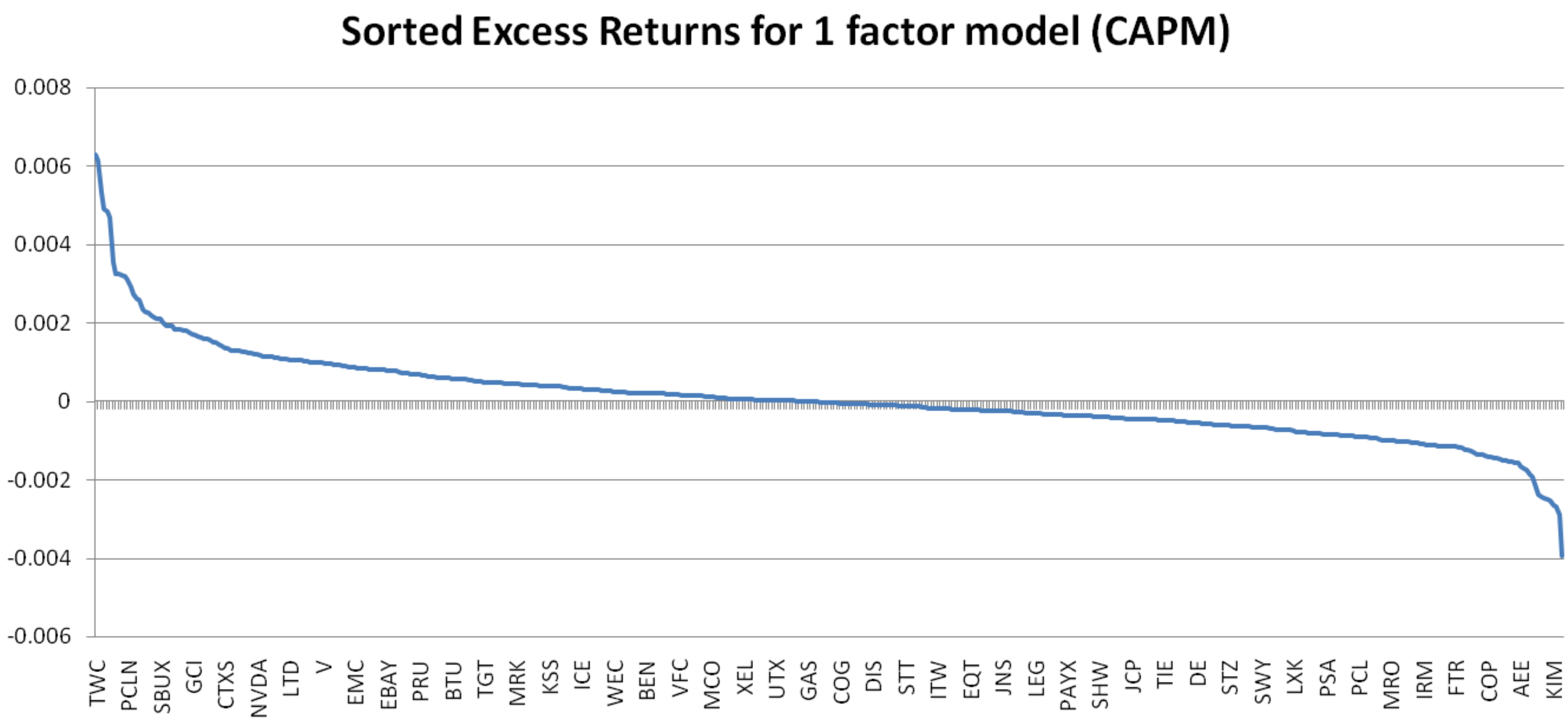
4. Analyze the vector of alphas

5. Analyze the correlation matrix of the residuals

Sorted Excess Returns, 1-factor

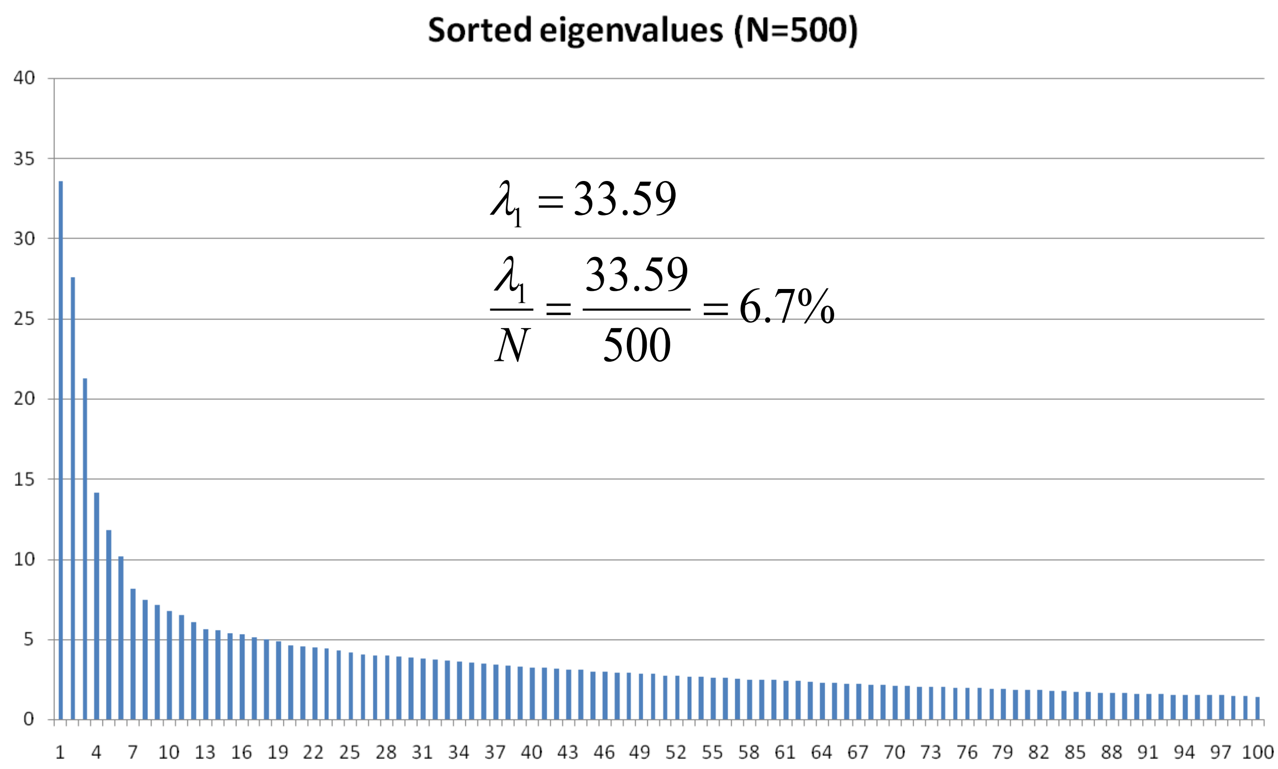


Sorted Excess Returns 1-factor, excluding AIG



Max=60 pbs, Min=-40 bps, average=1.9 bps, stdev=11bps

Eigenvalues of the correlation matrix of residuals (m=1)



$$\lambda_1 = 33.59$$

$$\frac{\lambda_1}{N} = \frac{33.59}{500} = 6.7\%$$

Recall that lambda_1 for the original correlation matrix was ~ 220, so the residuals matrix has ``smaller'' correlations.

Ratio lambda/N is a proxy for the average correlation.

First eigenvalue & average correlation

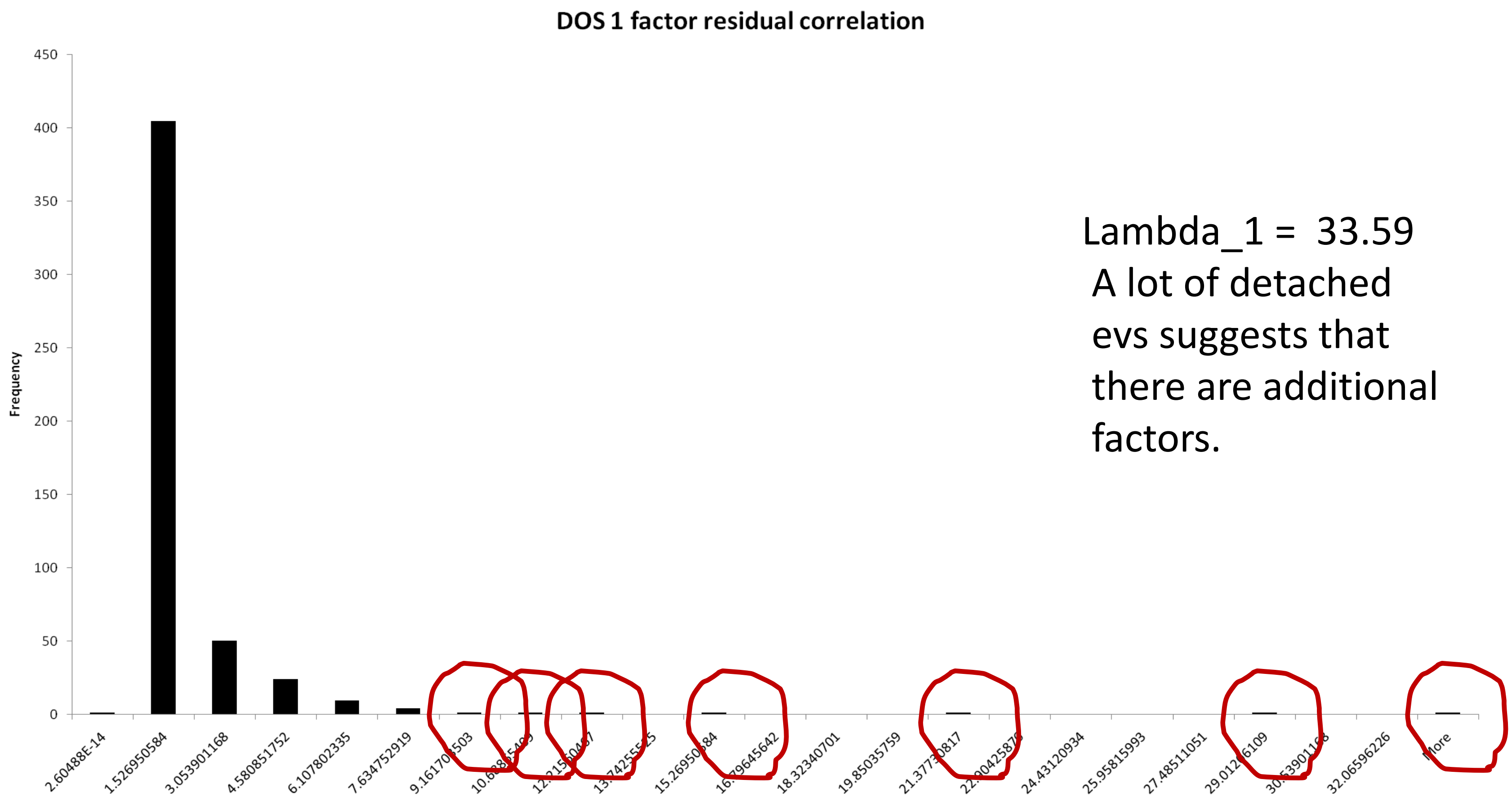
$$\begin{aligned}
\lambda_1 &= V^{(1)T} C V^{(1)} \\
&= \sum_{i=1}^{N} \left(V_i^{(1)}\right)^2 + \sum_{i \neq j} V_i^{(1)} V_j^{(1)} \rho_{ij} \\
&= 1 + \sum_{i \neq j} V_i^{(1)} V_j^{(1)} \rho_{ij} \\
&= 1 + \left(\sum_{i \neq j} V_i^{(1)} V_j^{(1)} \right) \cdot \frac{\sum_{i \neq j} V_i^{(1)} V_j^{(1)} \rho_{ij}}{\sum_{i \neq j} V_i^{(1)} V_j^{(1)}}
\end{aligned}$$

$$\frac{\lambda_1 - 1}{\sum_{i \neq j} V_i^{(1)} V_j^{(1)}} = \frac{\sum_{i \neq j} V_i^{(1)} V_j^{(1)} \rho_{ij}}{\sum_{i \neq j} V_i^{(1)} V_j^{(1)}} \qquad \therefore \quad V_i^{(1)} \approx \frac{1}{\sqrt{N}}, \quad \sum_{i \neq j} V_i^{(1)} V_j^{(1)} \approx \frac{N(N-1)}{N} = N - 1$$

$$\frac{\lambda_1 - 1}{N - 1} \approx \langle \rho \rangle$$

$$\boxed{\langle \rho \rangle \approx \frac{\lambda_1}{N}}$$

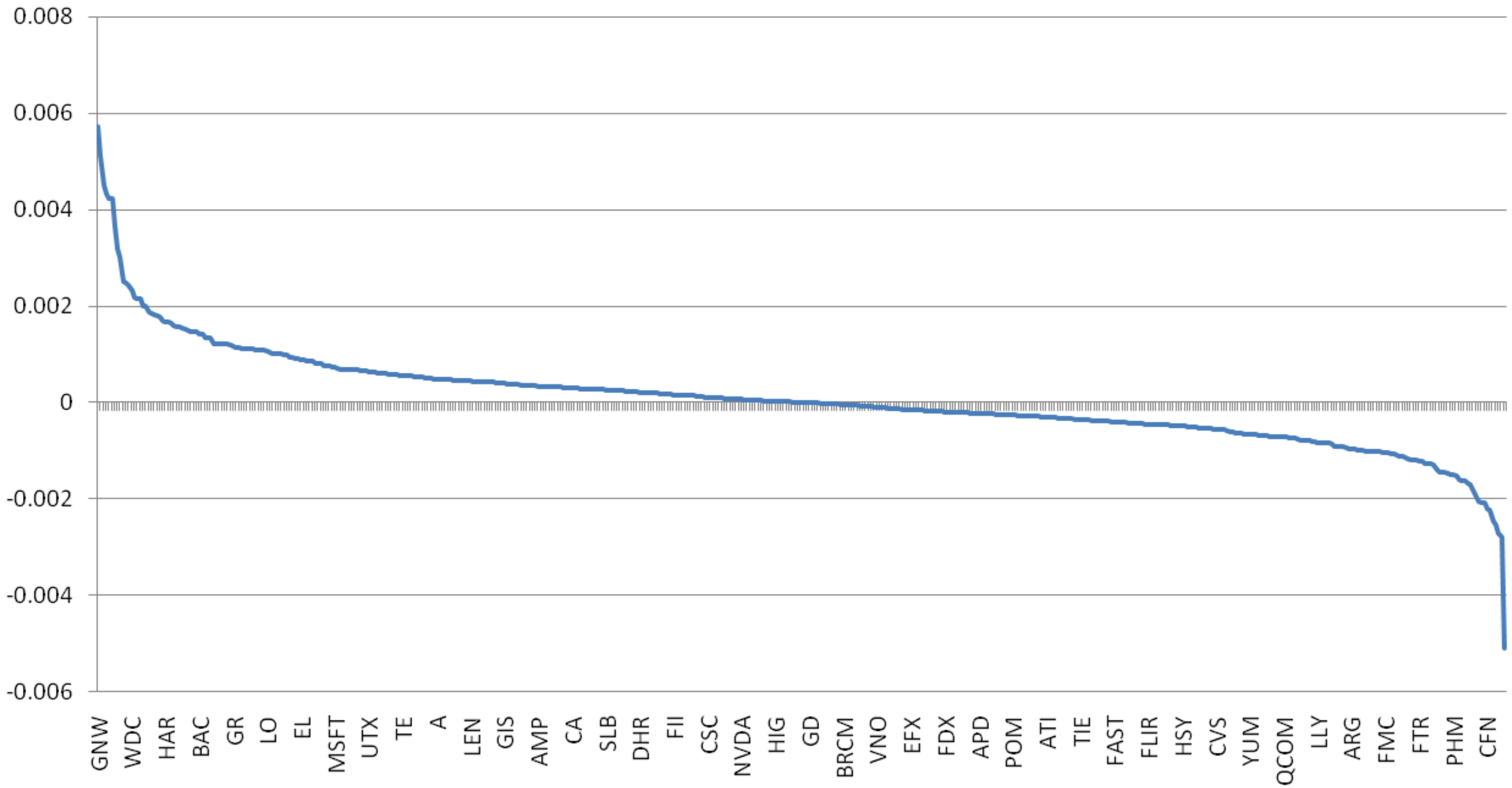
Density of States, or Histogram, of Eigenvalues



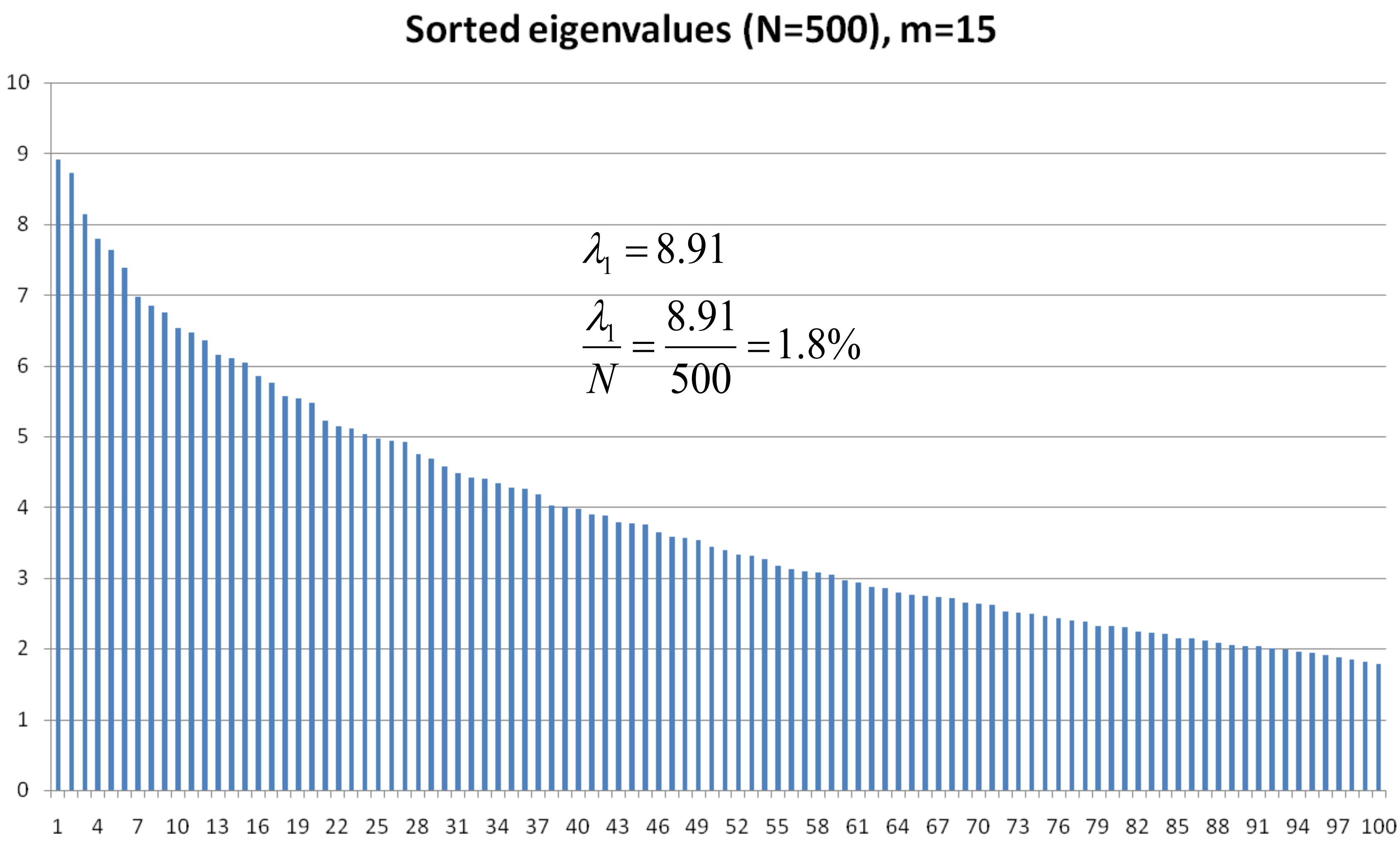
Lambda_1 = 33.59

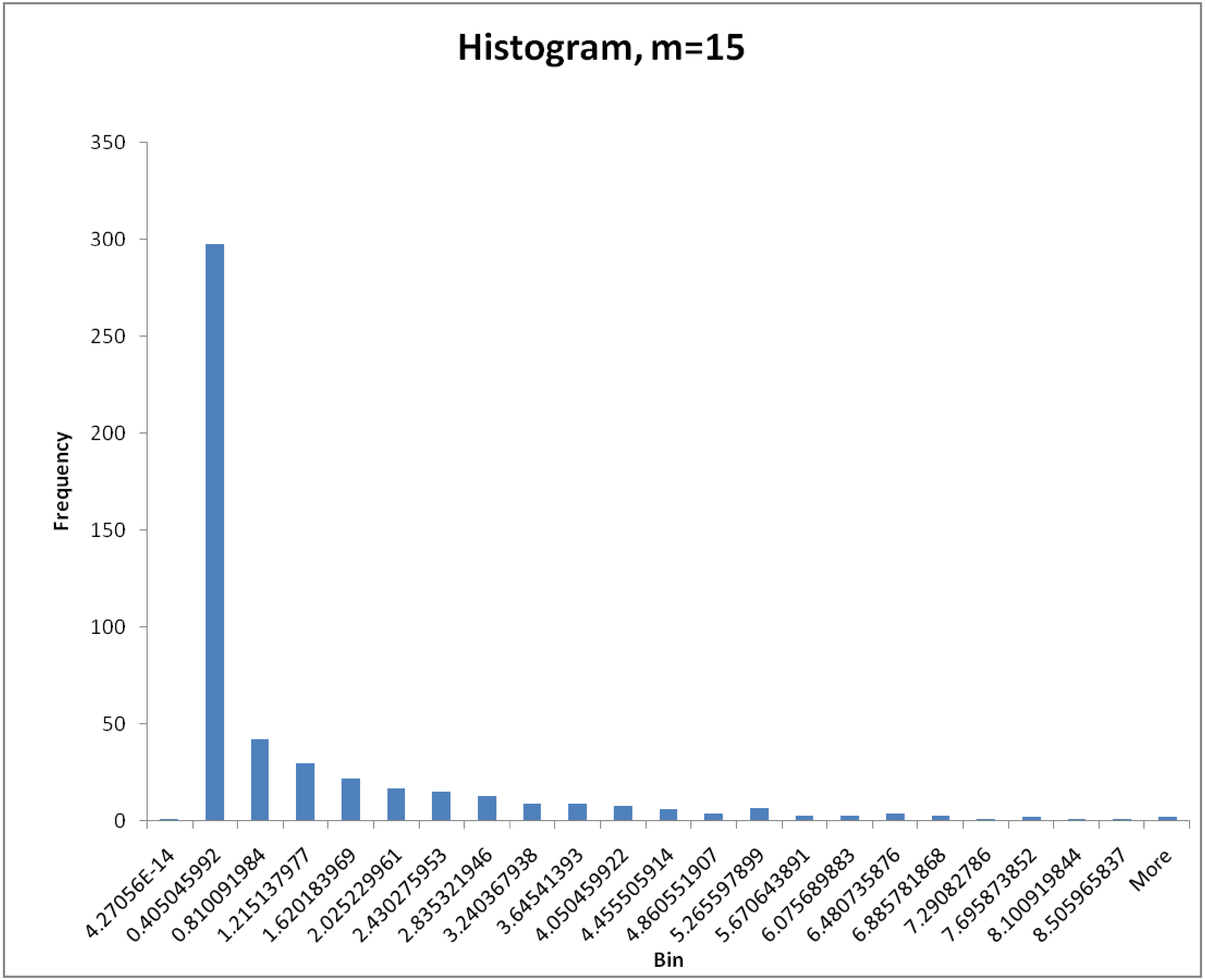
A lot of detached evs suggests that there are additional factors.

Sorted excess returns, m=15 (without AIG)

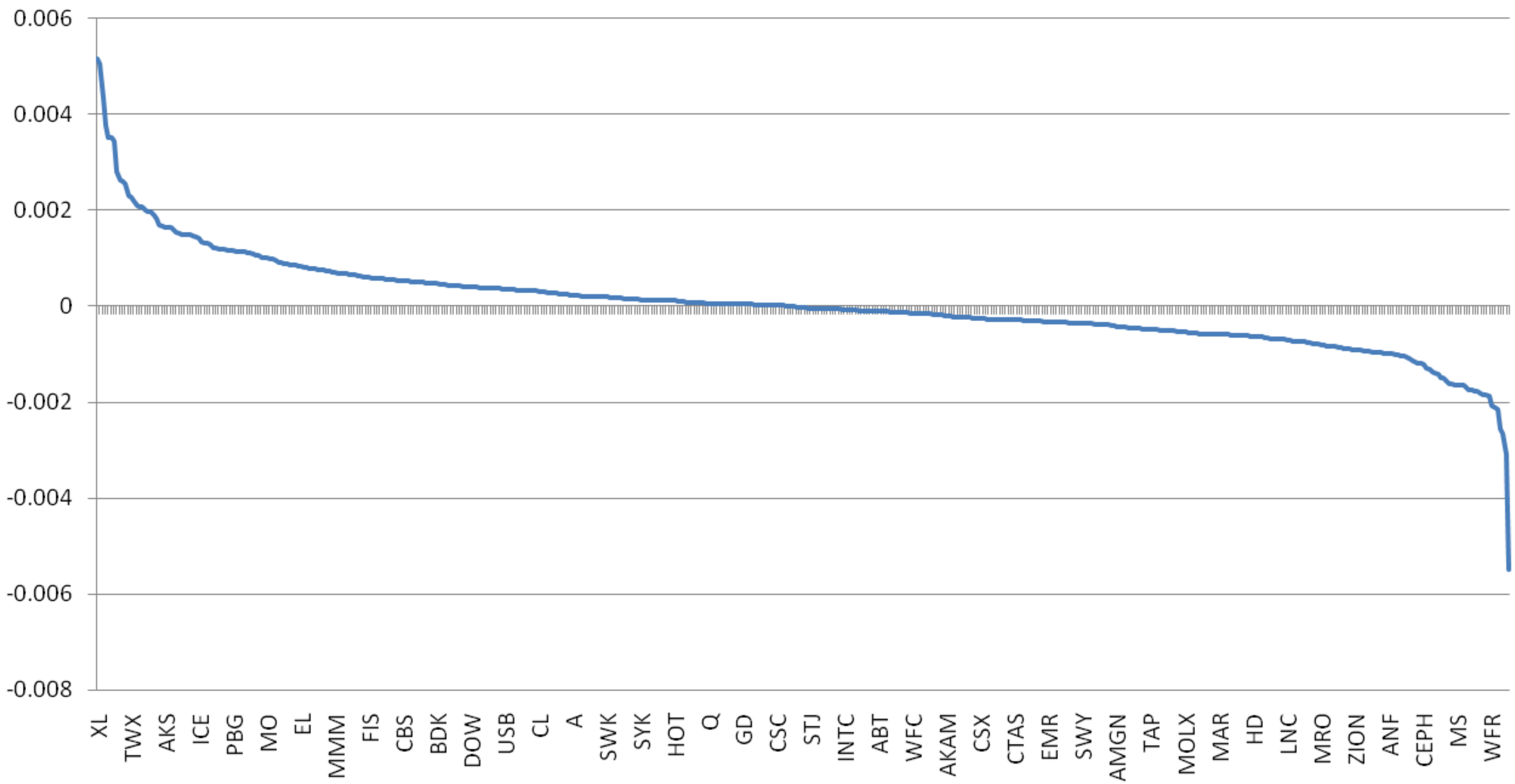


Top 100 eigenvalues of the correlation matrix of residuals (m=15)

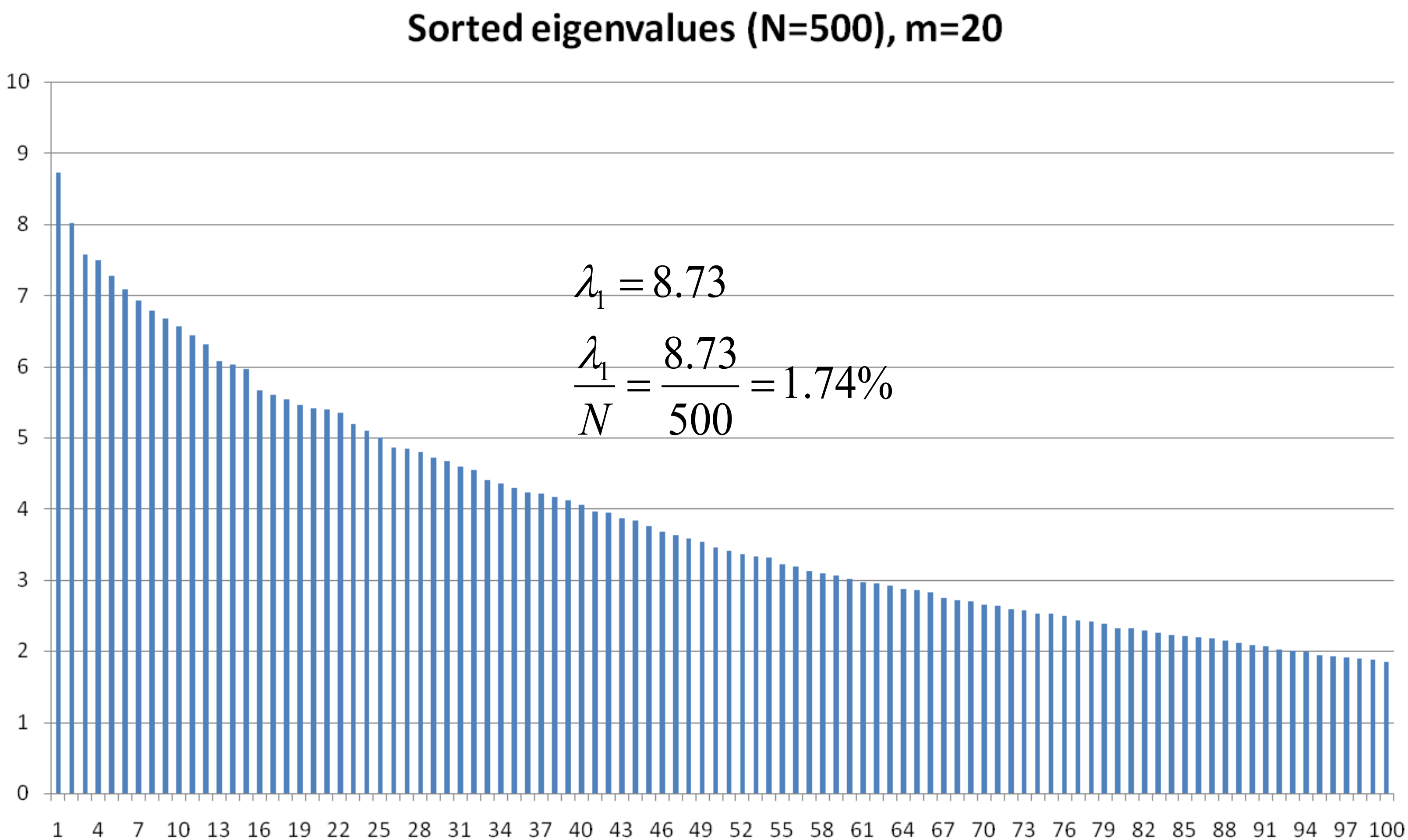



Histogram, m=15
Frequency
350
300
250
200
150
100
50
0
4.27056E-14
0.405045992
0.810091984
1.215137977
1.620183969
2.025229961
2.430275953
2.835321946
3.240367938
3.64541393
4.050459922
4.455505914
4.860551907
5.265597899
5.670643891
6.075689883
6.480735876
6.885781868
7.29082786
7.695873852
8.100919844
8.505965837
More
Bin

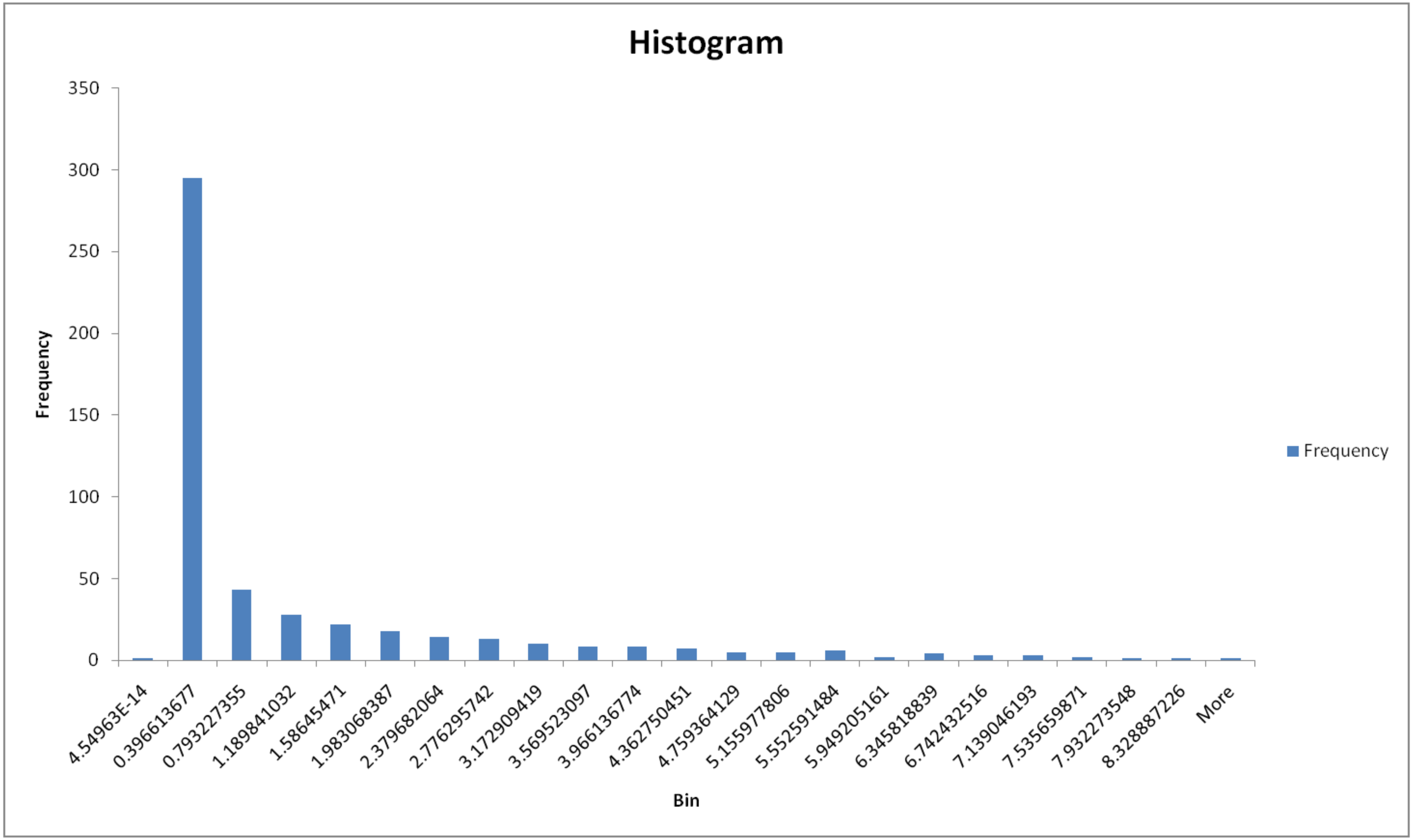
Sorted Excess Returns, m=20



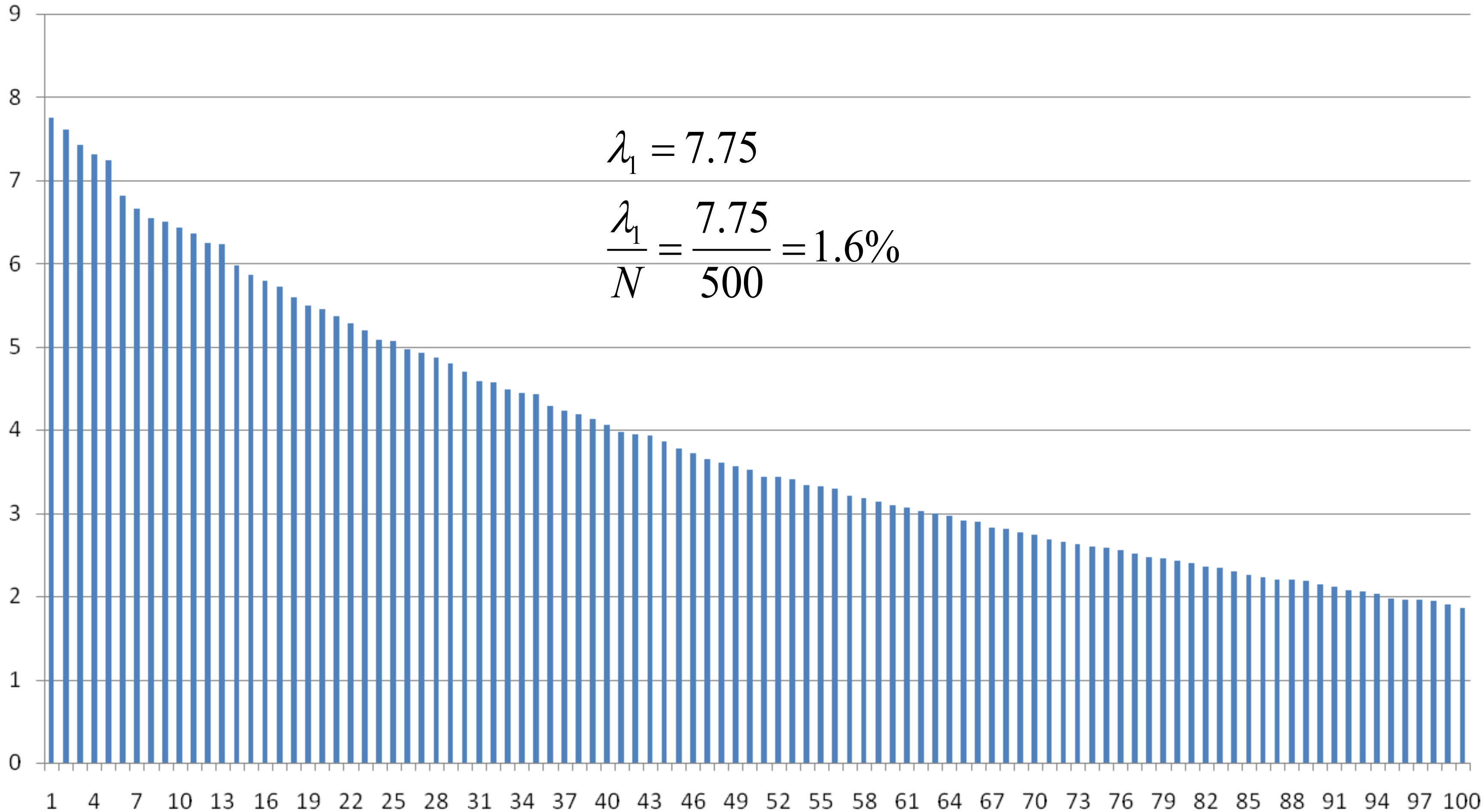
Sorted Eigenvalues, m=20



Density of States m=20



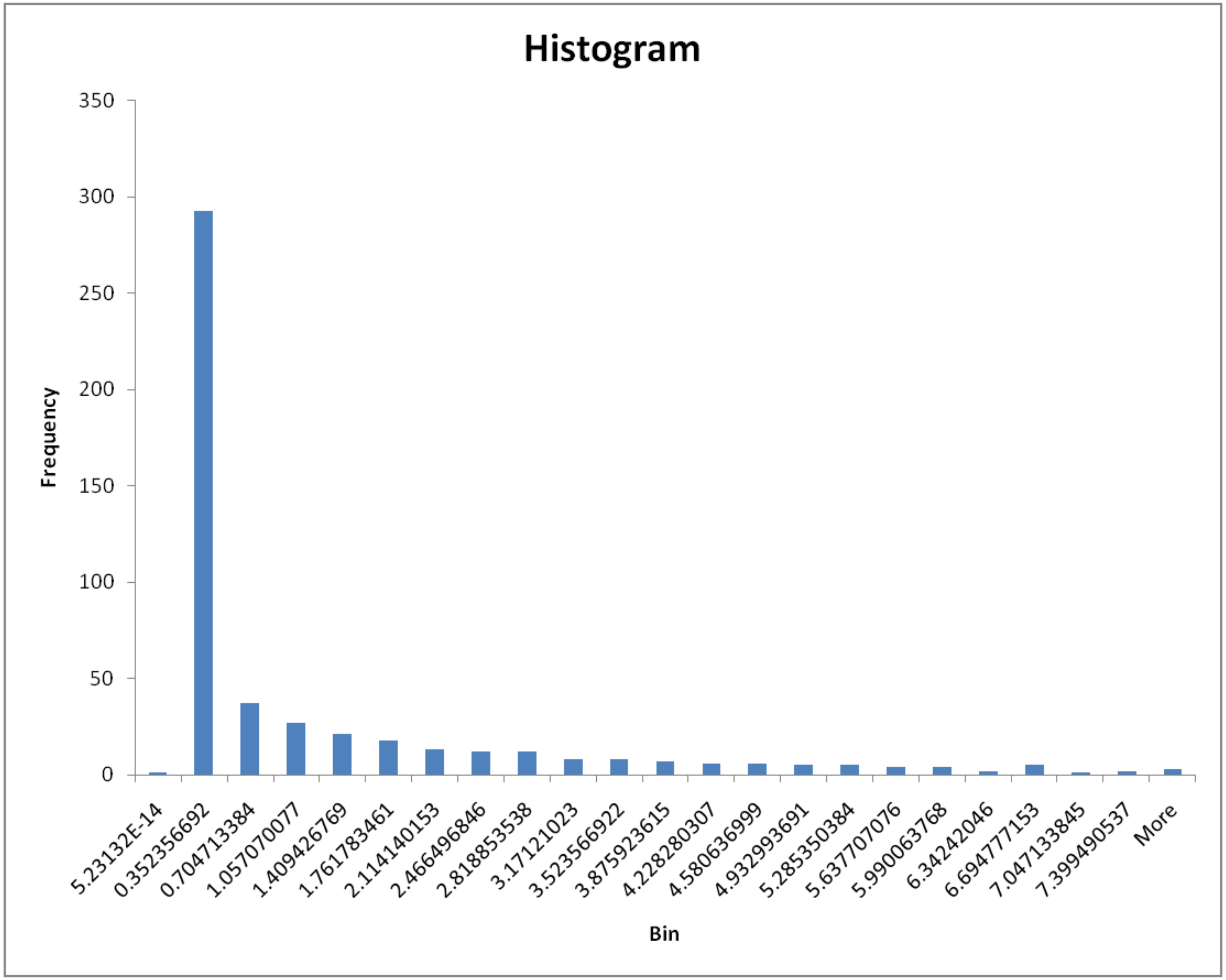
Sorted eigenvalues (N=500), m=30



$$\lambda_1 = 7.75$$

$$\frac{\lambda_1}{N} = \frac{7.75}{500} = 1.6\%$$

Density of States m=30



Excess returns (alpha) as a function of the number of eigenportfolios (m)

m	max	min	average	average abs	stdev
1	0.6283%	-0.3941%	0.0196%	0.0776%	0.1129%
15	0.5095%	-0.5096%	0.0065%	0.0687%	0.1004%
20	0.5095%	-0.5485%	0.0968%	0.0667%	0.0968%
30	0.5095%	-0.3957%	0.0049%	0.0664%	0.0960%

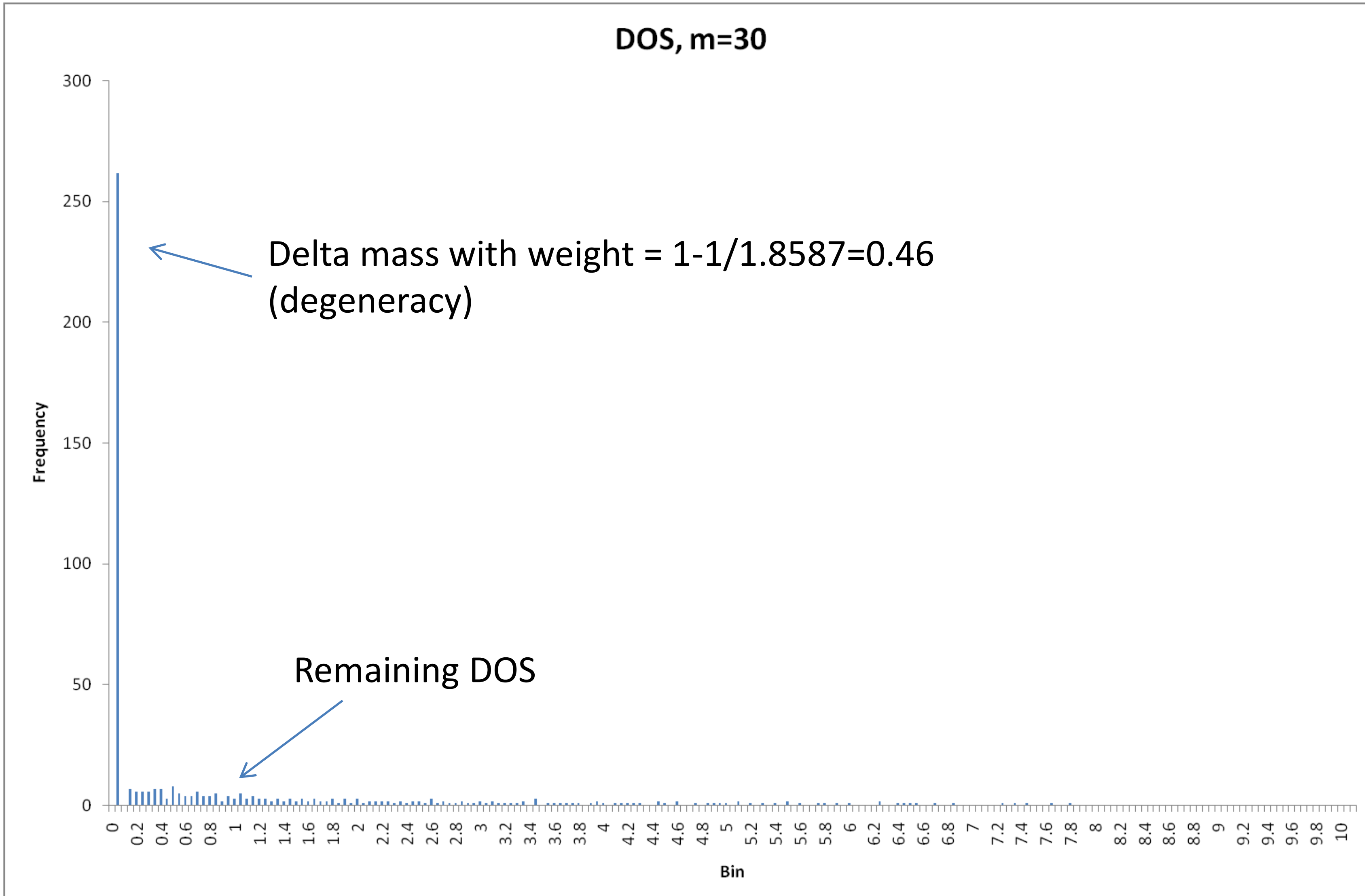
Marcenko-Pastur Distribution for the DOS of a Random Correlation Matrix

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.

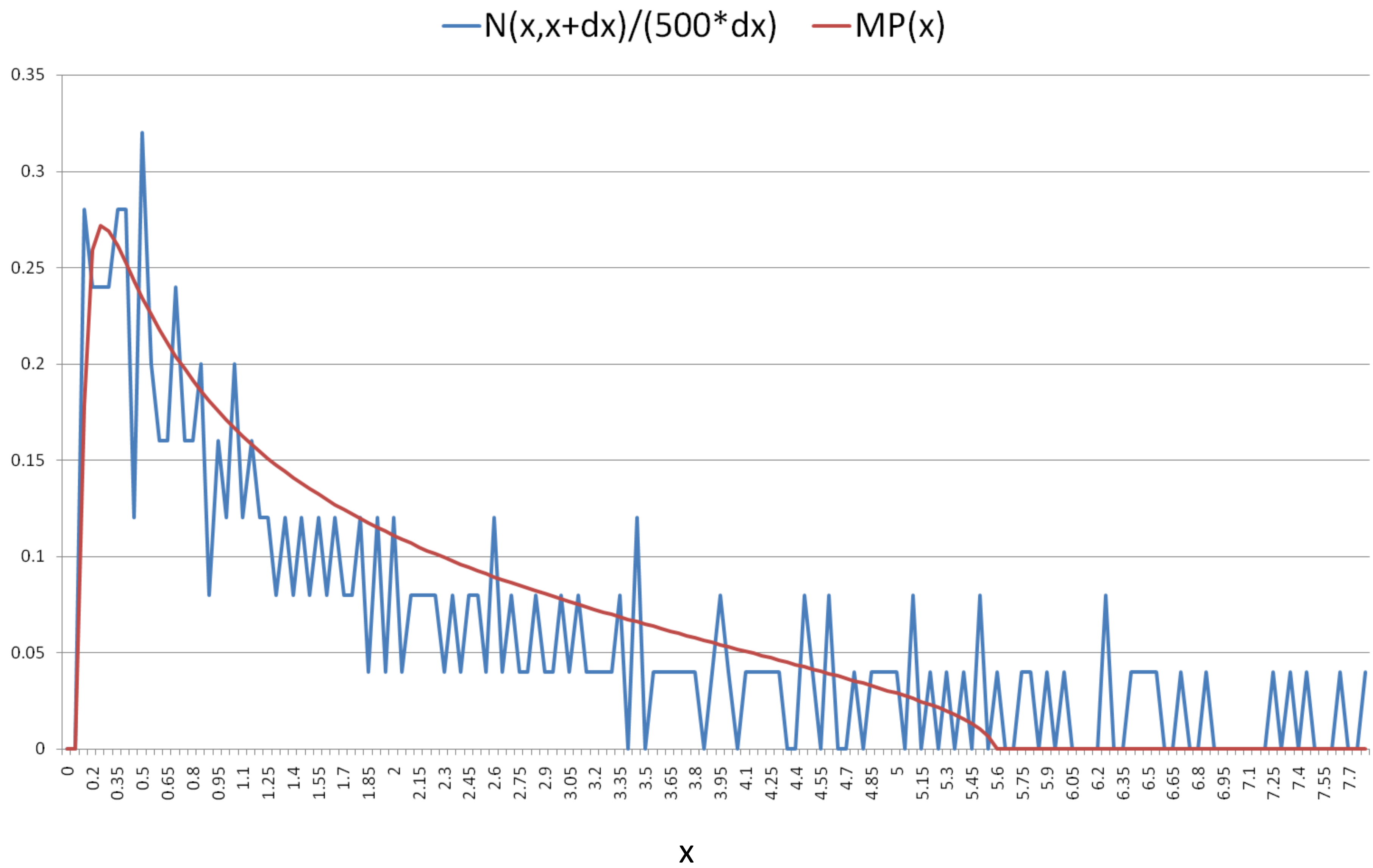
$$\gamma = \frac{N}{T} \qquad \lambda_+ = \left(1+\sqrt{\gamma}\right)^2 \qquad \lambda_- = \left(1-\sqrt{\gamma}\right)^2$$

$$MP(\lambda) = \left(1-\frac{1}{\gamma}\right)^+ \delta(\lambda) + \frac{1}{2\pi\gamma}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}$$

Interpreting the DOS for the residuals in terms of Marcenko Pastur



Marcenko Pastur compared to data with m=30



Evaluating the use of ETFs as factors in APT

We found out how many eigenportfolios are needed approximately
to explain the systematic portion of stock returns using panel data for stock returns.

We obtained a matrix of random residuals if we choose m=15 or higher.

Since eigenportfolios are not tradable (except perhaps for the first one), this leaves us with the identification problem.

We perform an analysis of APT using sector ETFs as factors.

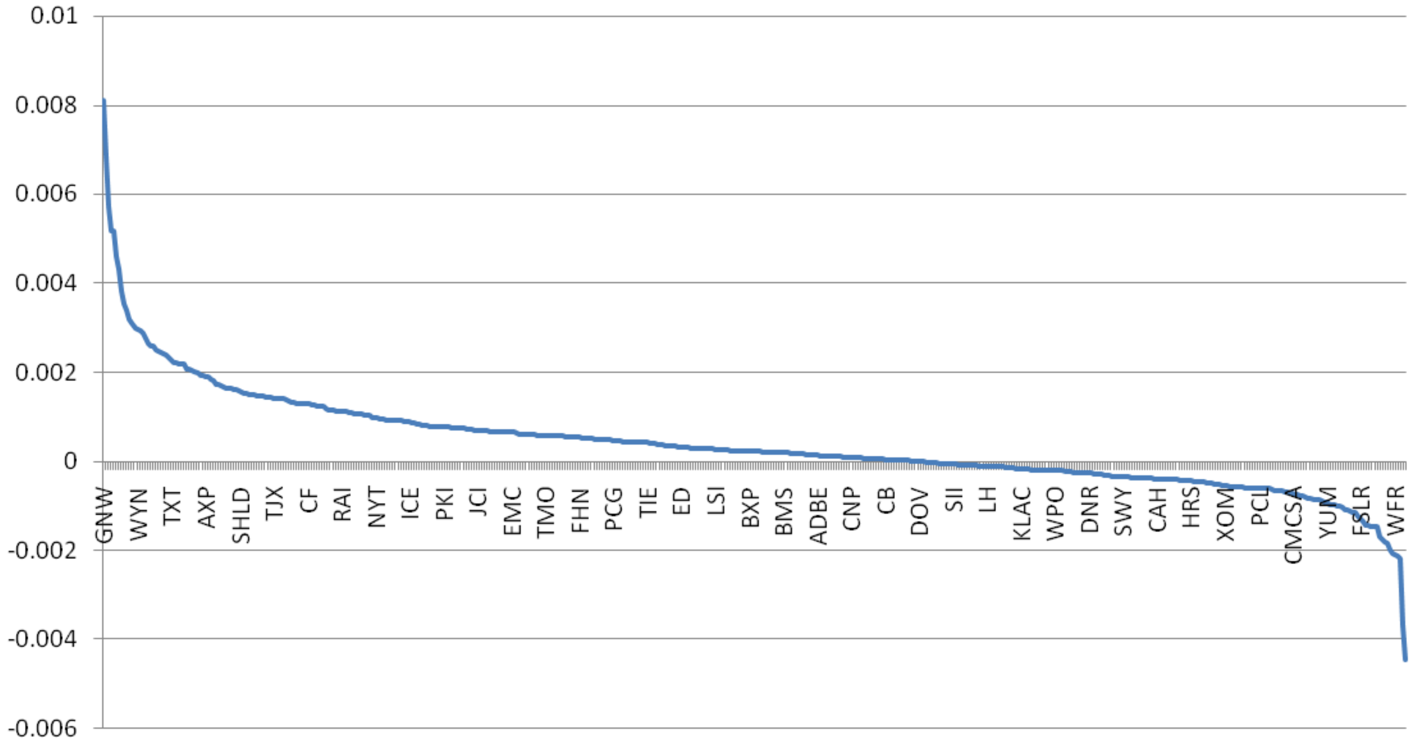
Three experiments:

* Multiple regression on 19 ETFs

* Matching pursuit on 19 ETFs

* Association of a single ETF to each stock

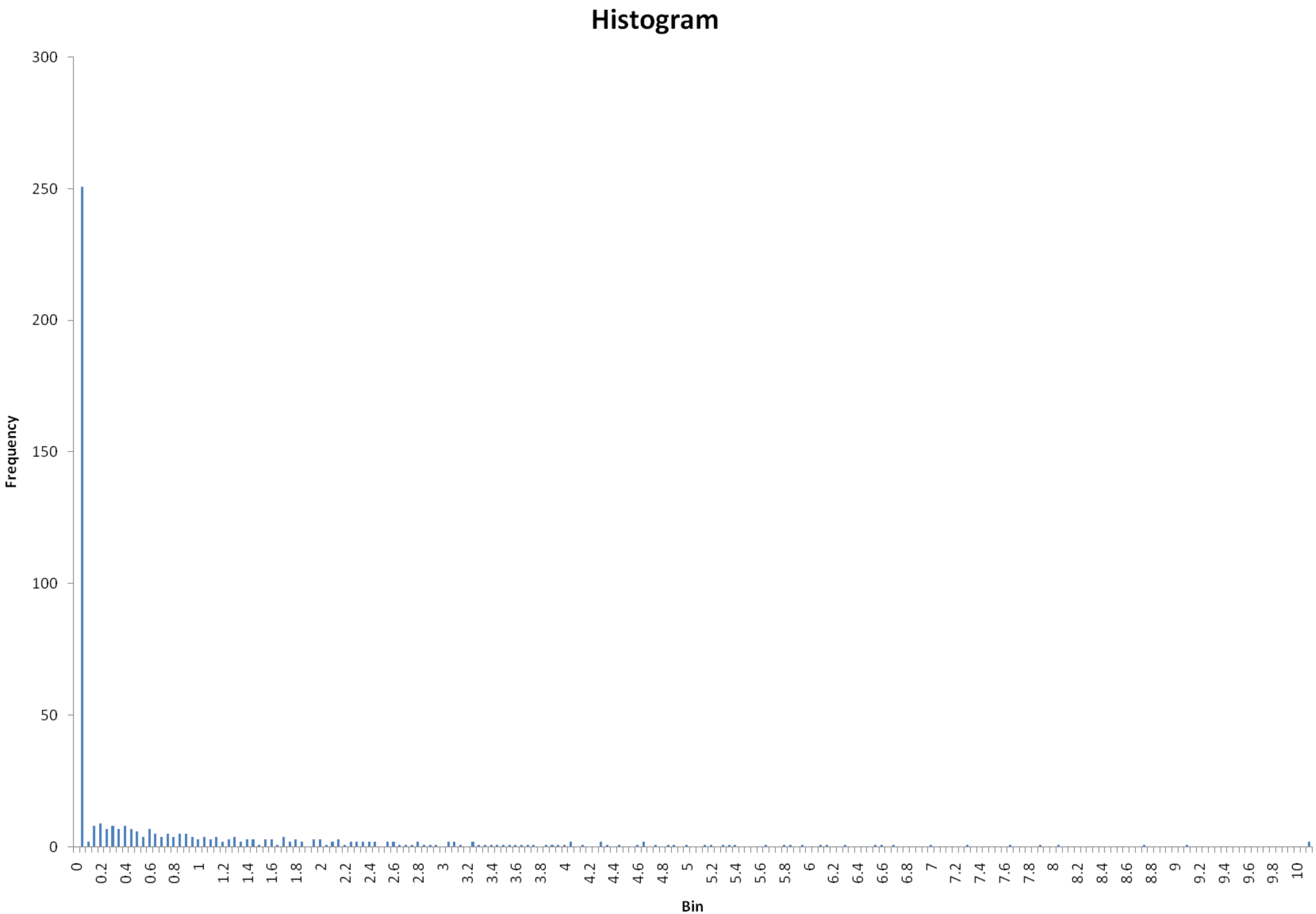
Sorted Excess Returns, Factors=19 ETFs Multiple Regression



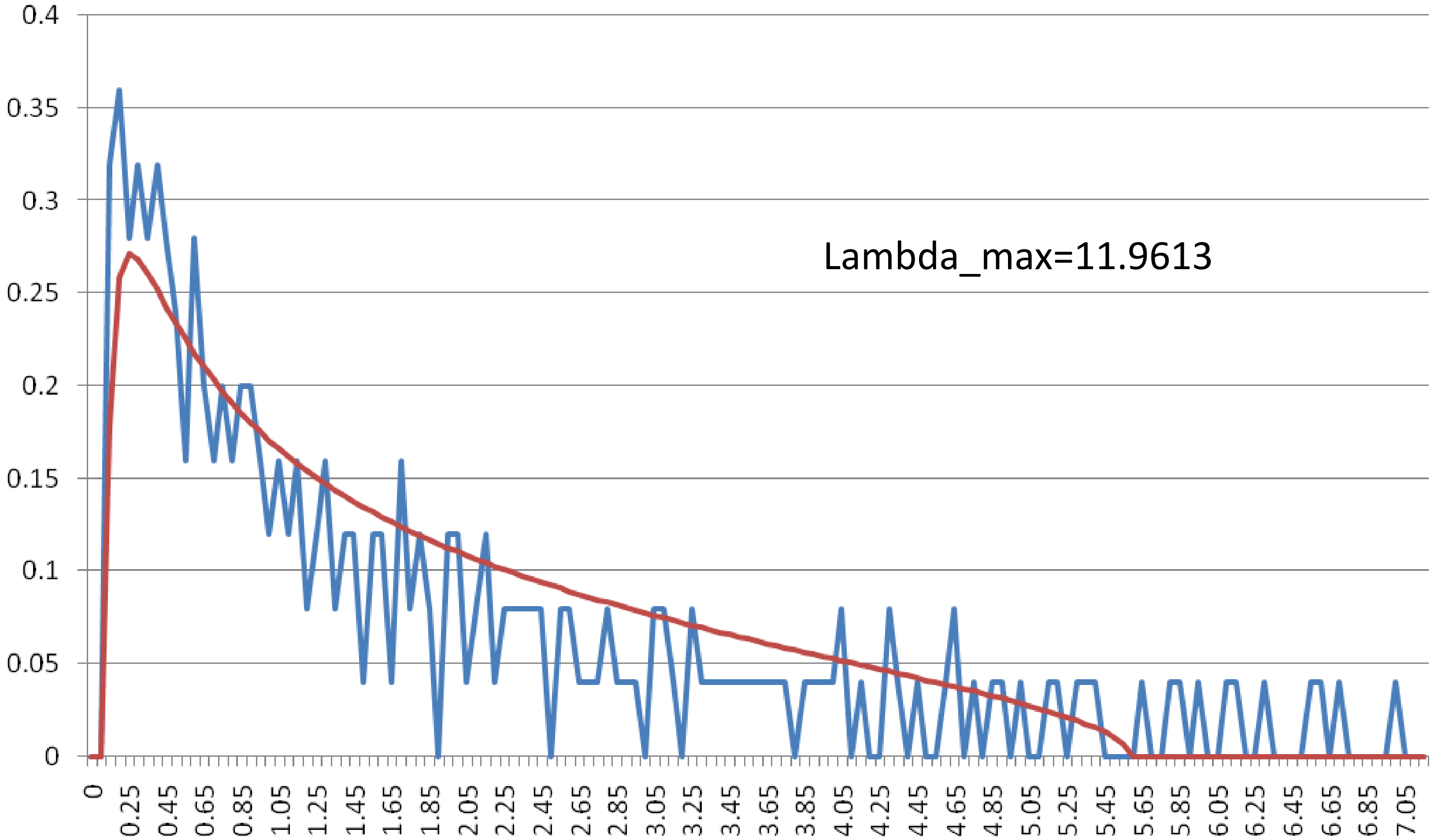
average	average abs	stdev
0.0390%	0.2755%	0.1167%

(AIG not included)

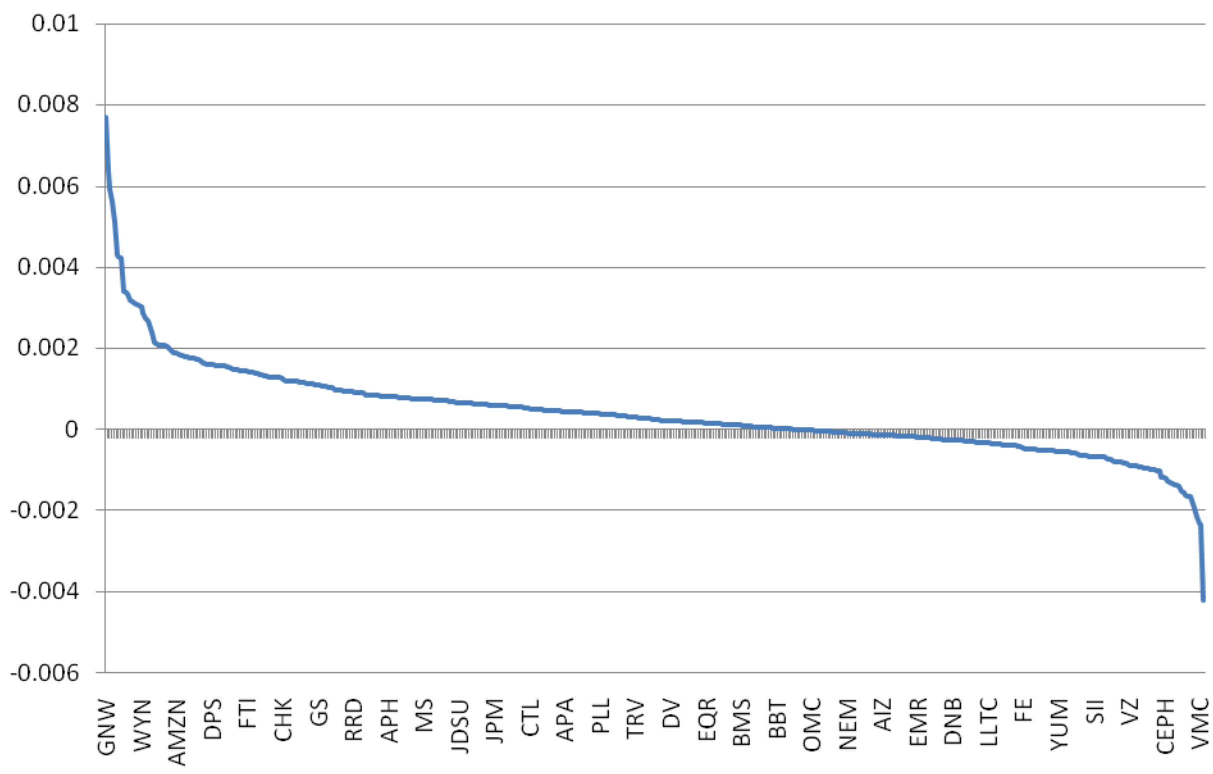
Density of States, residuals with 19 ETFs



After removing mass at zero (19 etfs, MR)
(Red line=Marcenko-Pastur)

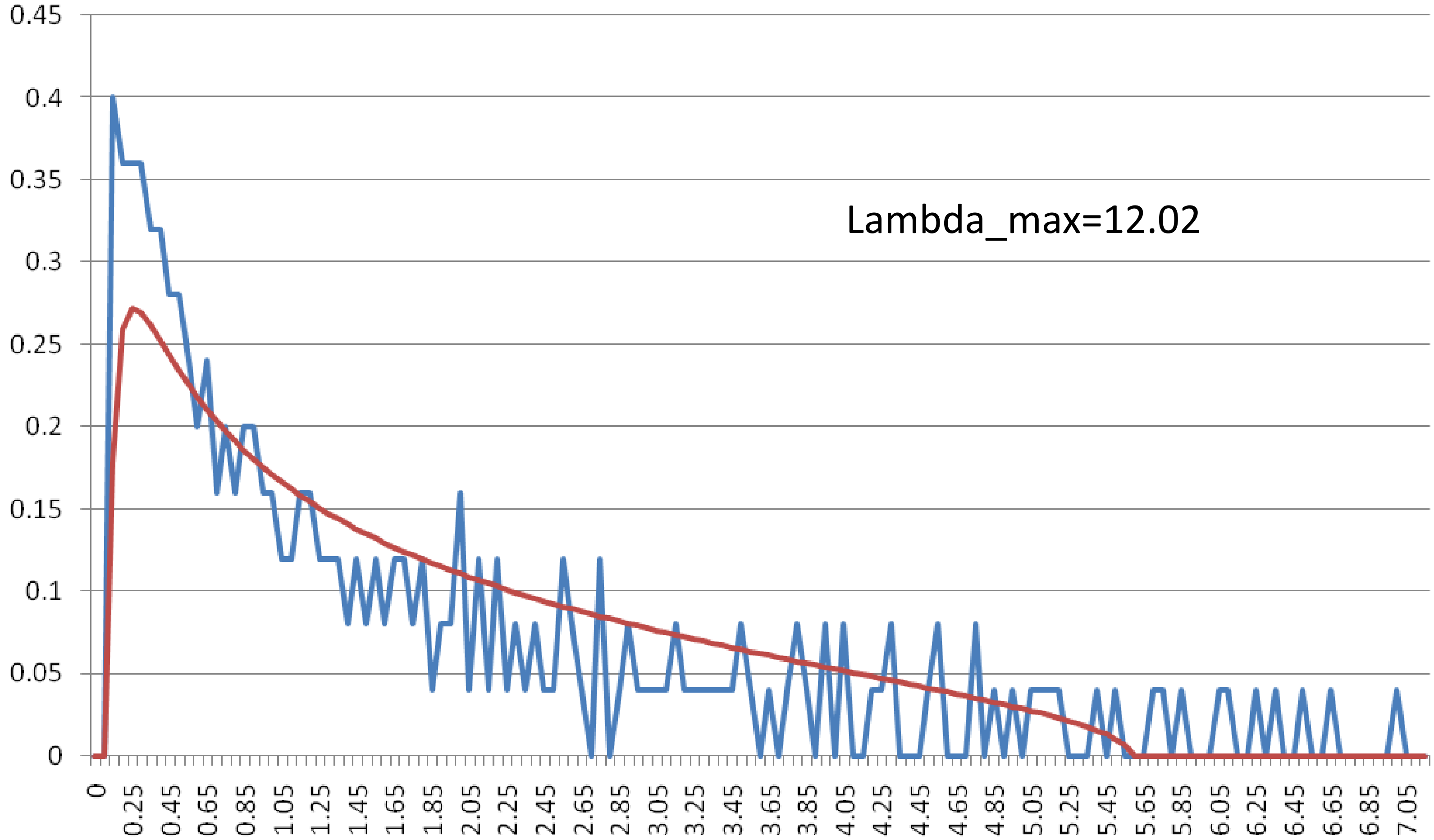


Excess Returns:
Matching Pursuit, 19 ETFs (w/o AIG)

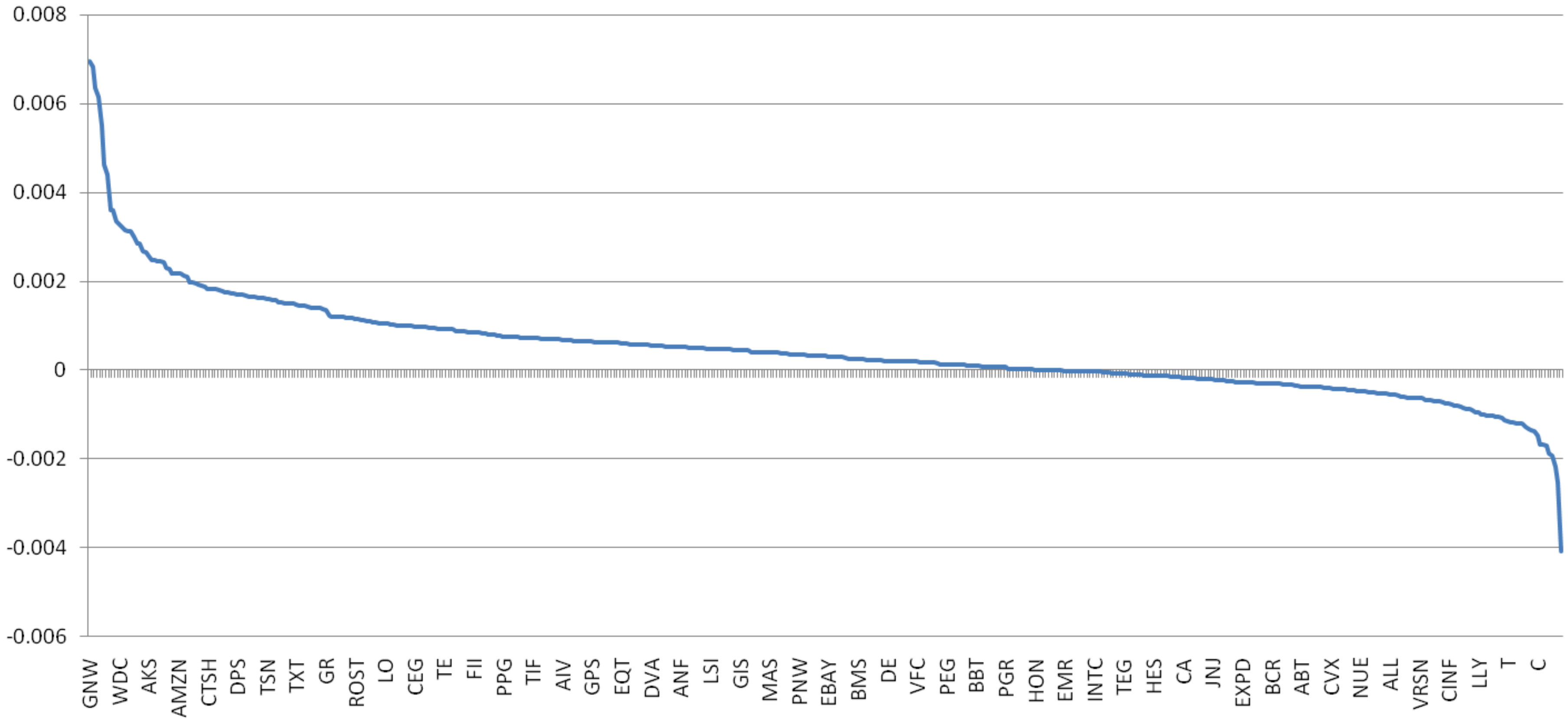


average	average abs	stdev
0.0405%	0.0799%	0.0904%

Noise Spectrum for Matching Pursuit Residuals

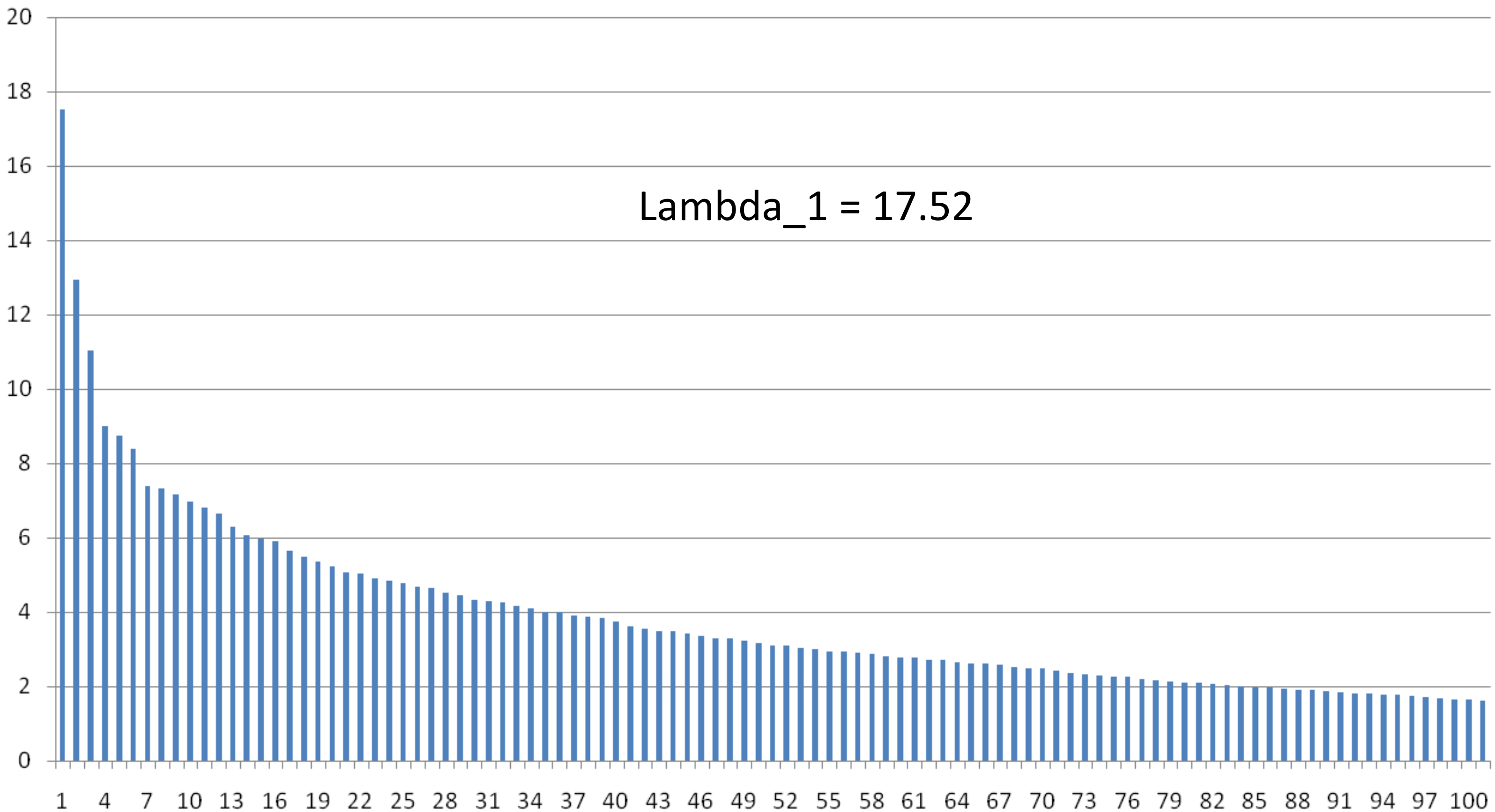


Excess Returns after projecting on the corresponding industry ETF for each stock

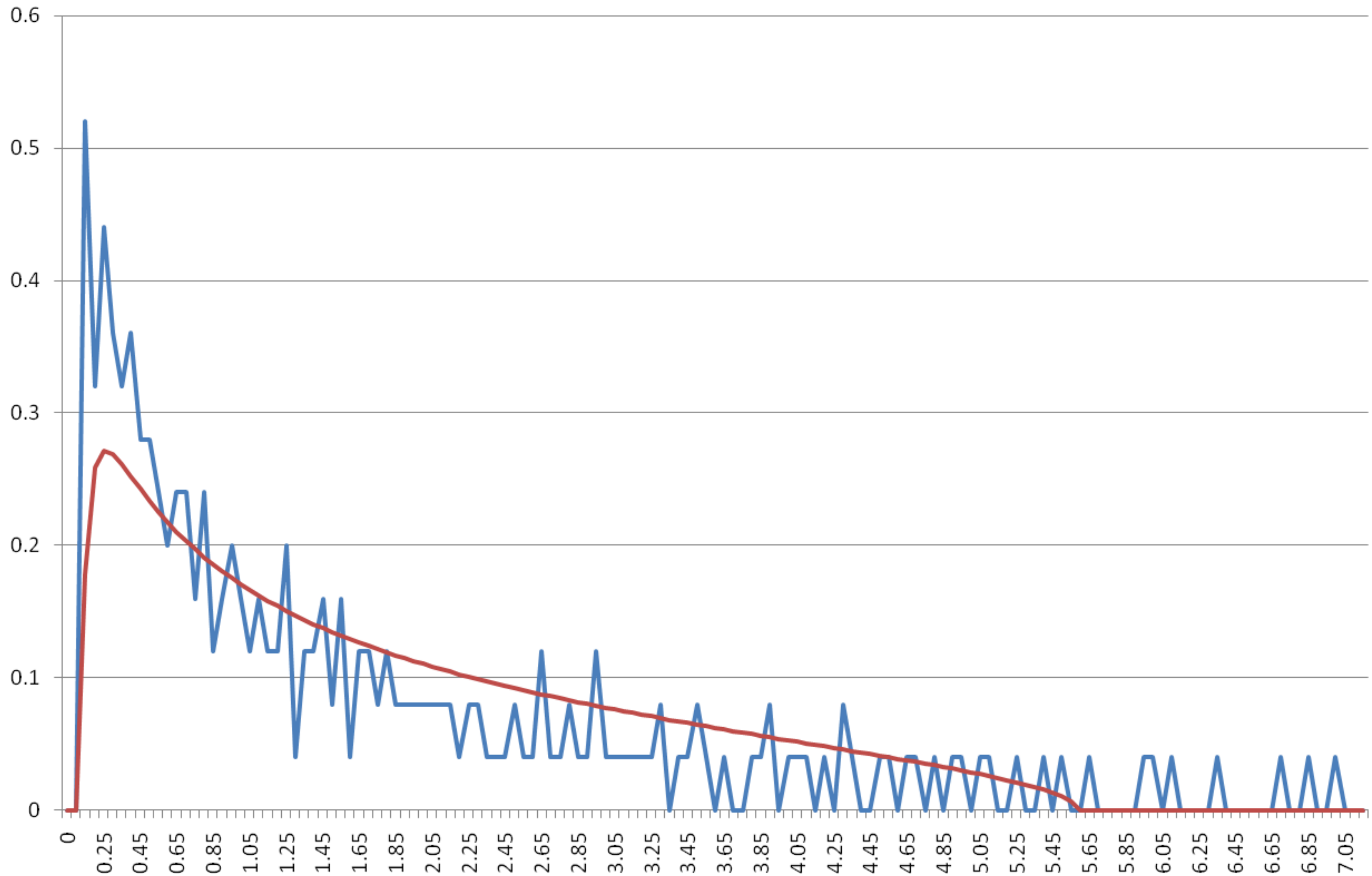


average	average abs	stdev
0.0461%	0.0823%	0.1153%

100 top eigenvalues for residuals after removing industry ETF



Density of States for Correlation matrix of residuals after removing the sector ETF for each stock



Building a Risk-Management System for Stock portfolios based on PCA

- A stock portfolio often comprises 100's of positions, both long and short
- Risk-management systems are necessary to evaluate the market risk
- This evaluation is critical to determine the capital requirements for the portfolio
- Estimate a PDF for portfolio variations over a specified time-period Dt
- Calculate the losses at different confidence levels and set a risk-management policy based on covering tail losses

PCA Eigenportfolios as Risk Factors

$$\frac{\Delta S}{S} = \sum_{j=1}^{m} \beta_{sj} F_j + \varepsilon_s$$

$$F_j = \text{ return of eigenportfolio } \# \, j$$

$$F_j = \sum_{i=1}^{N} \left(\frac{V_i^{(j)}}{\sigma_i} \right) R_i = \sum_{i=1}^{N} w_i^{(j)} R_i$$

-- Factors arise directly from data analysis
-- Uncorrelated factors

-- Identification problem (does “coherence” hold?)
-- Noise

Algorithm

- Perform PCA on large universe of stocks (e.g. s&p 500) and extract m=15 eigenportfolios

- Store the eigenportfolio returns in a matrix (15*T)

- For each stock in the portfolio, calculate the 15 regression coefficient as in

$$R_t = \alpha + \sum_{j=1}^{m} \beta_j F_t^j + \varepsilon_t$$

- Replace factors by standardized Student-t variables and consider the model

$$R_{it} = \alpha + \sum_{j=1}^{m} \beta_{ij} \varphi_t^{\ j} + \left(\sigma_i^2 - \sum_{j=1}^{m} \beta_{ij}^2 \right)^{1/2} \varphi_{it}^*$$

$\varphi_t^j, \varphi_{it}^*$ independent standardized Student t (t = 4)

Monte Carlo Simulation

- Generate a large number T of samples of the m=15 ``factor'' student variates and the idiosyncratic risk
- Calculate the losses for any given portfolio under the generated scenarios and determine the risk limit in this way.

Dynamics

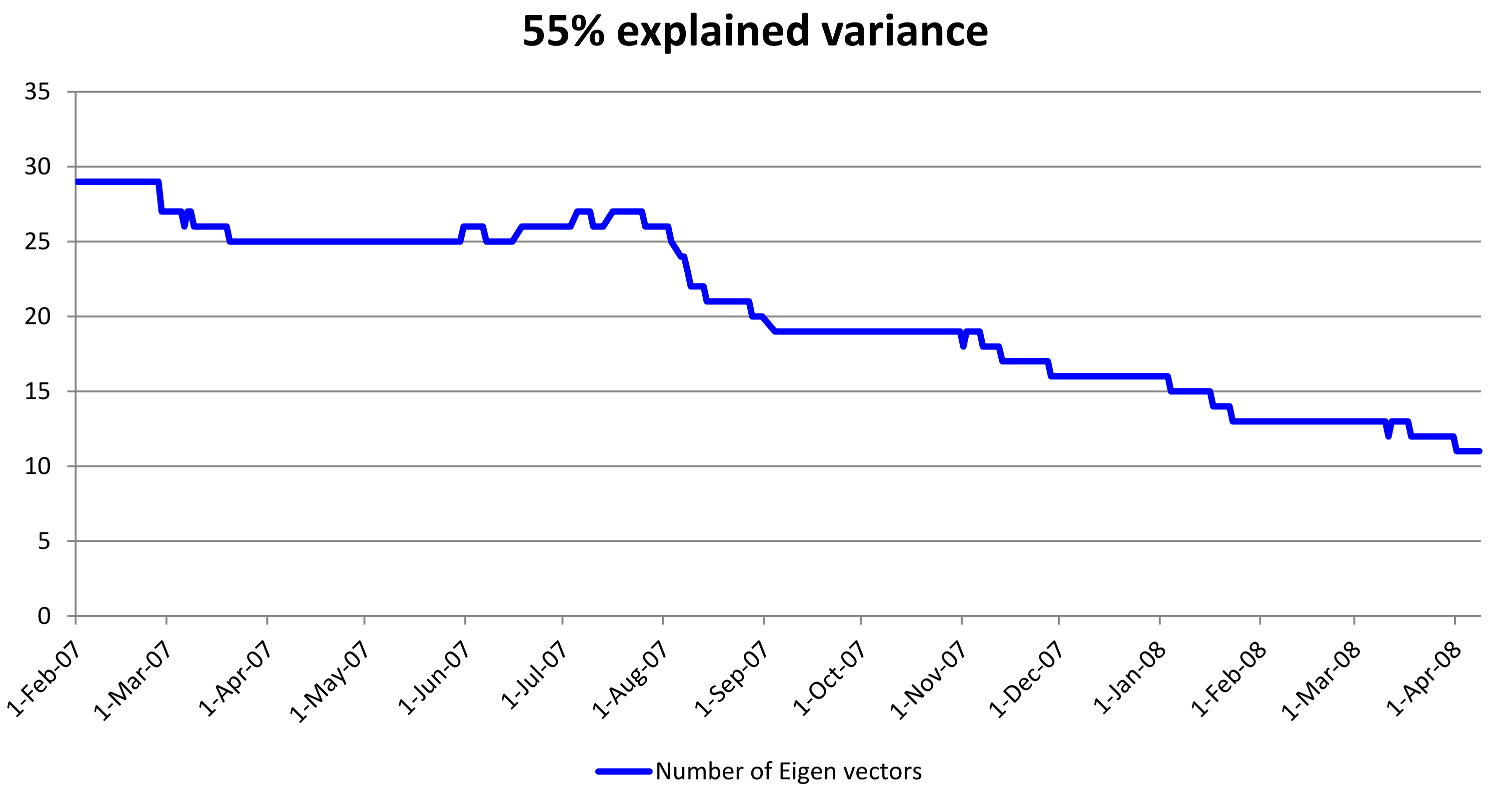
So far, the analysis that we made assumes a fixed window.

We should ask ourselves how these relationships change across time, i.e. if the factor count and the R-squared that we obtained are stable across time.

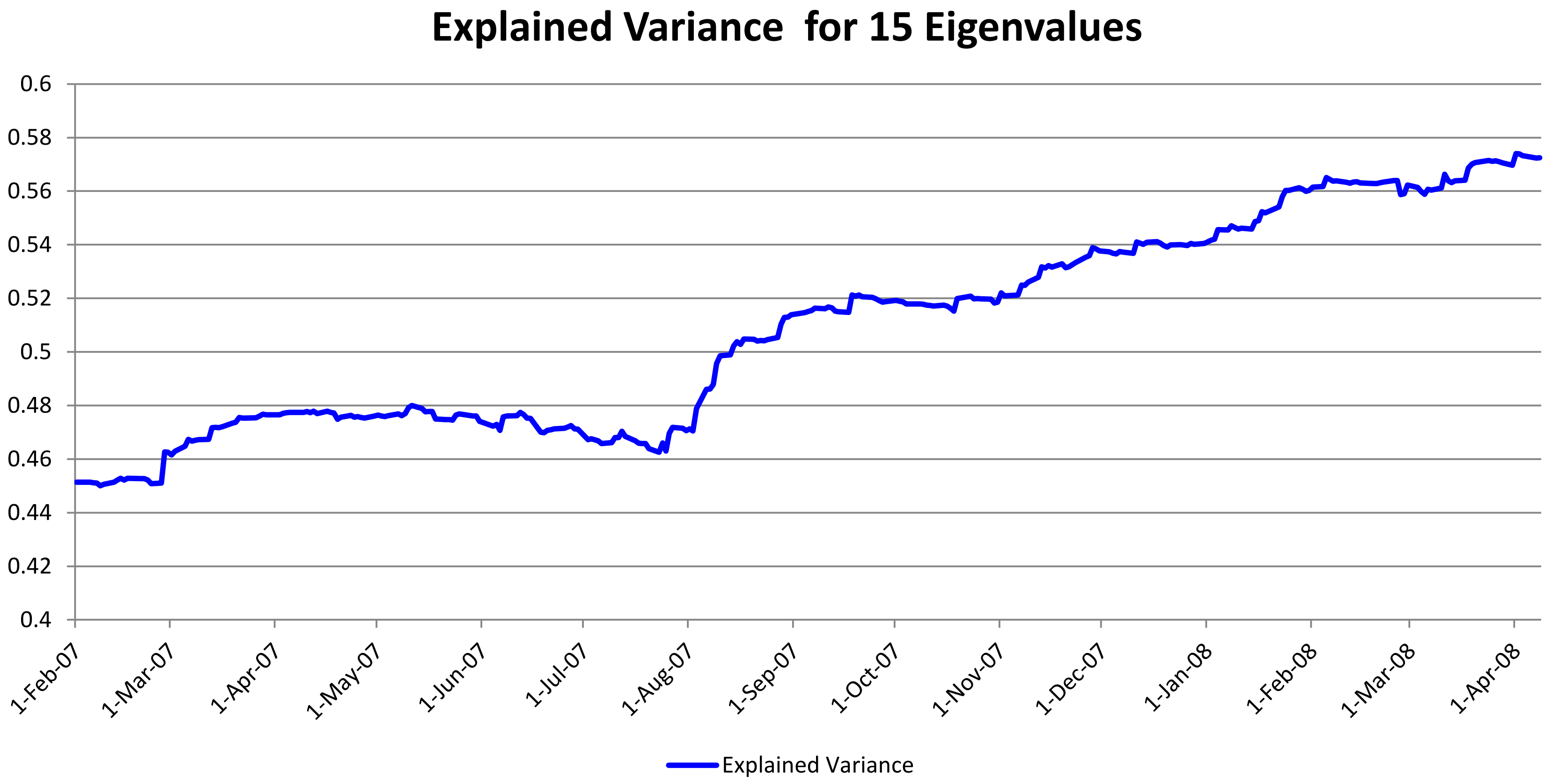
This is particularly important if the factor model is used for hedging or for relative valuation of stocks with respect to ETFs.

The following charts show some results of the PCA analysis viewed as time passes, using a moving window to calculate the eigenvalues and eigenvectors.

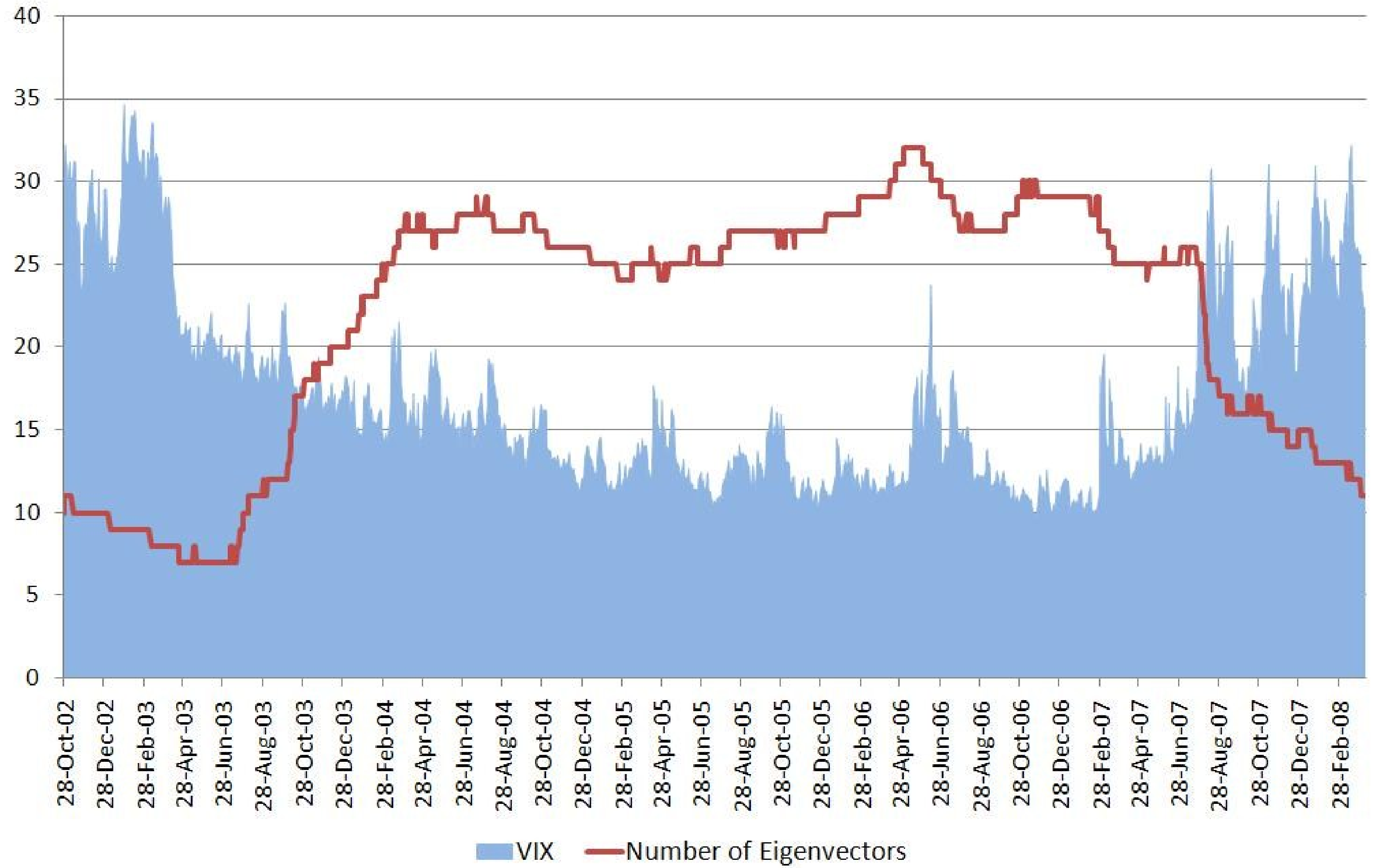
Number of significant eigenvectors at 55% level



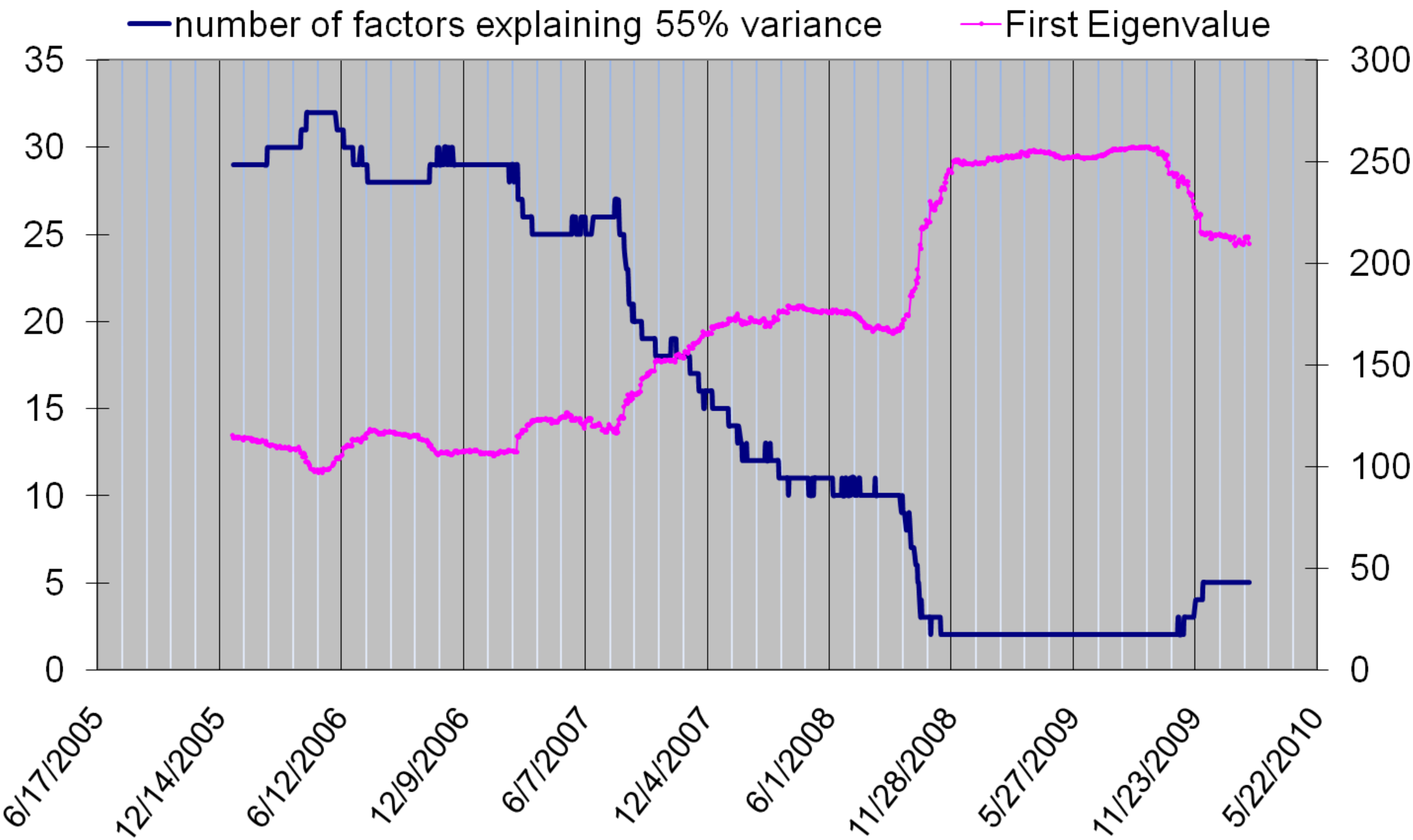
Fixed number of eigenvectors (factors)



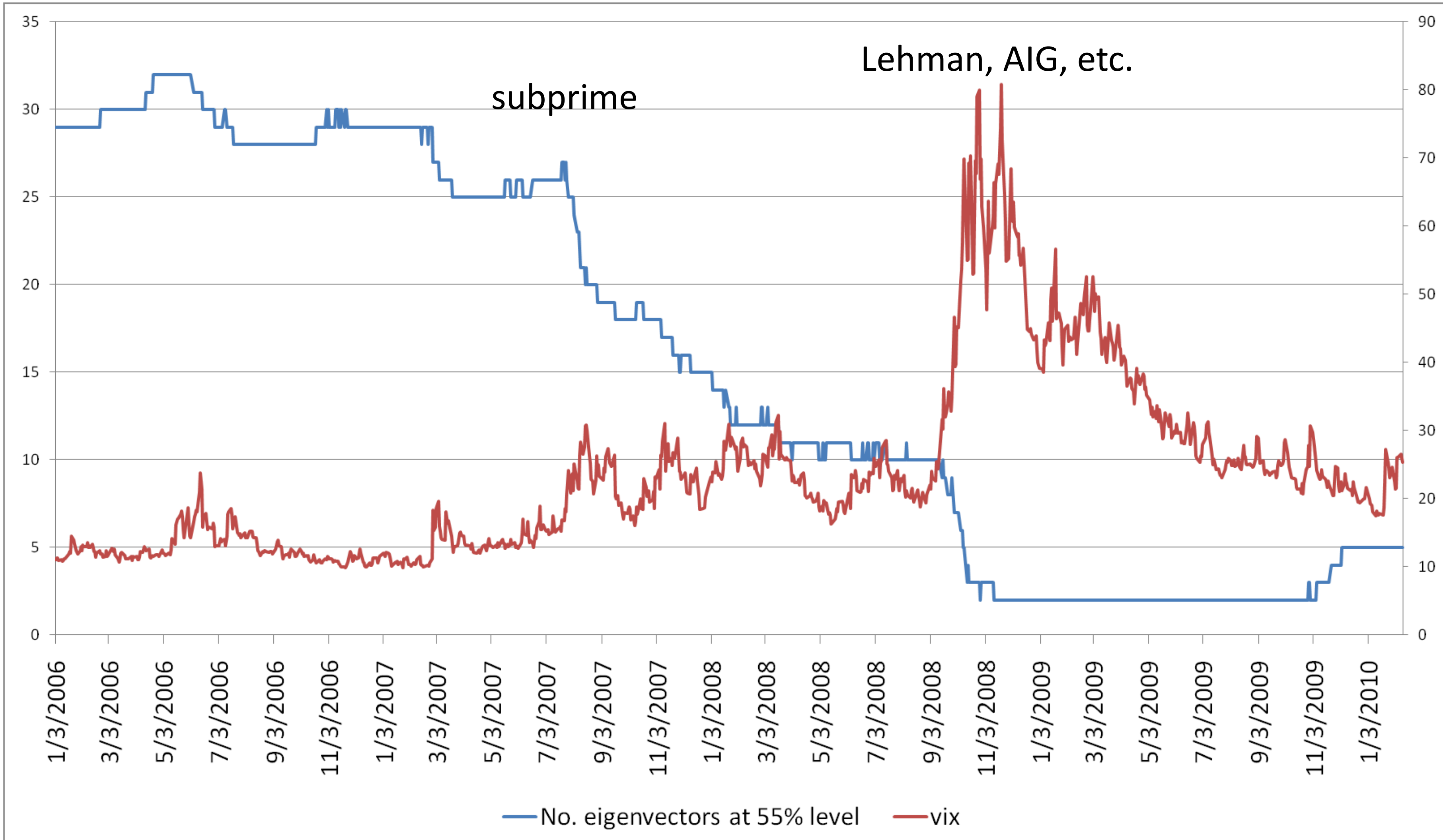
Number of factors explaining 55% of the variance versus VIX volatility index (2002-2008)



Number of explanatory factors vs. first eigenvalue of correlation matrix



Number of EVs versus VIX (1/2006-2/2010)



Dynamics are important

The previous slides show that the structure of the market is far from static.

This is obvious if we consider innovations in the market (new issues, new industries, the economic cycle, bubbles).

Equilibrium theories (e.g. APT, CAPM) are insufficient to explain prices, volatilities and correlations of financial assets.

Hence the need to model the evolution of financial variables using stochastic processes based on time-series analysis.

What can time-series analysis do for us?

-- Understand serial correlations in the data
-- Construct predictive models over suitable time-windows.
-- Discrete-time processes: important for data analysis.
-- Continuous-time processes: useful for theoretical purposes and to
   model high-dimensional data.