Lecture 2:
PCA and Risk Factors

Marco Avellaneda

G63.2936.001

Spring Semester 2009

Principal components analysis for financial 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

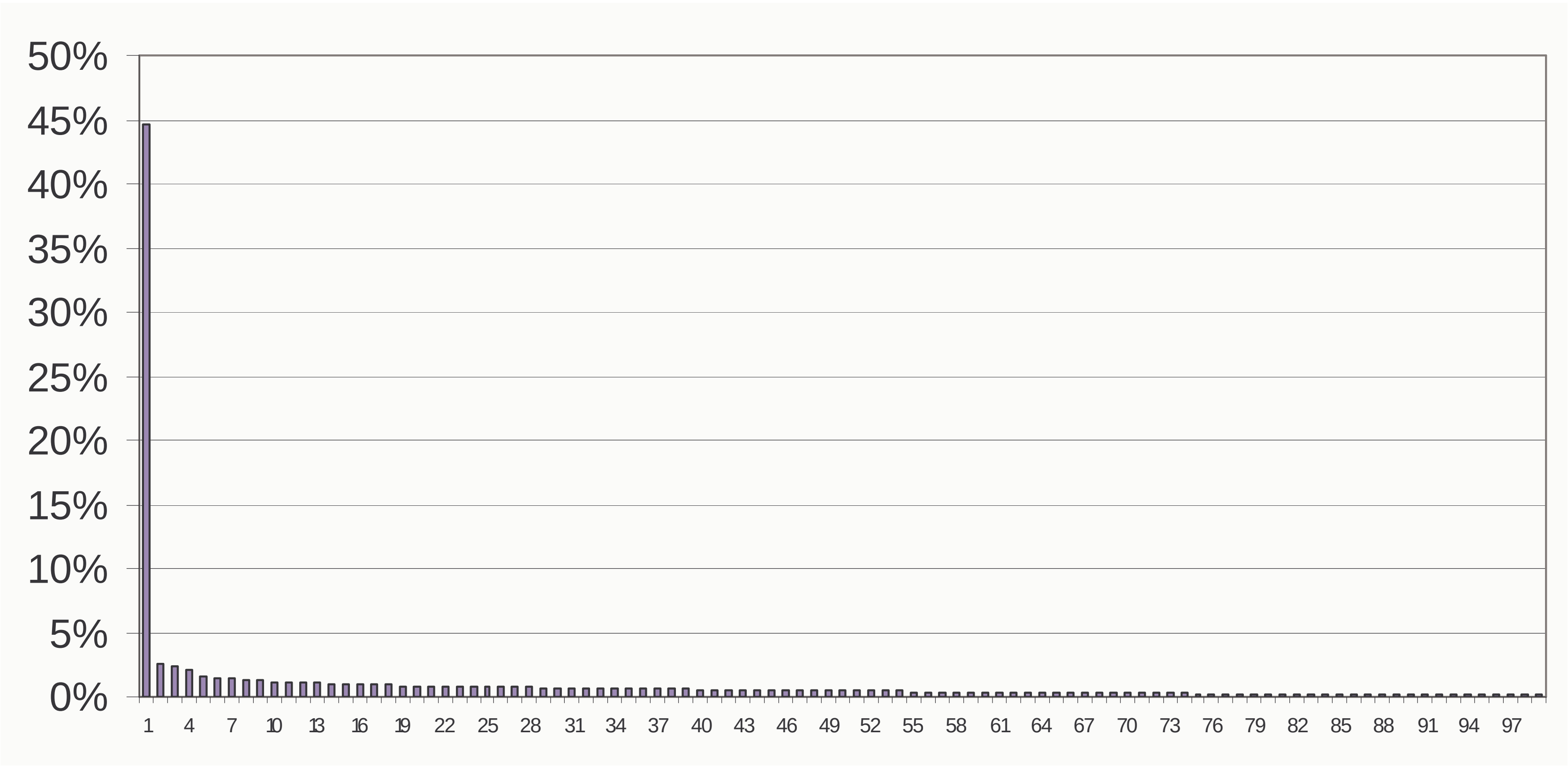
-- Estimate the number of significant components

-- Analyze the corresponding eigenvectors and eigenportfolios

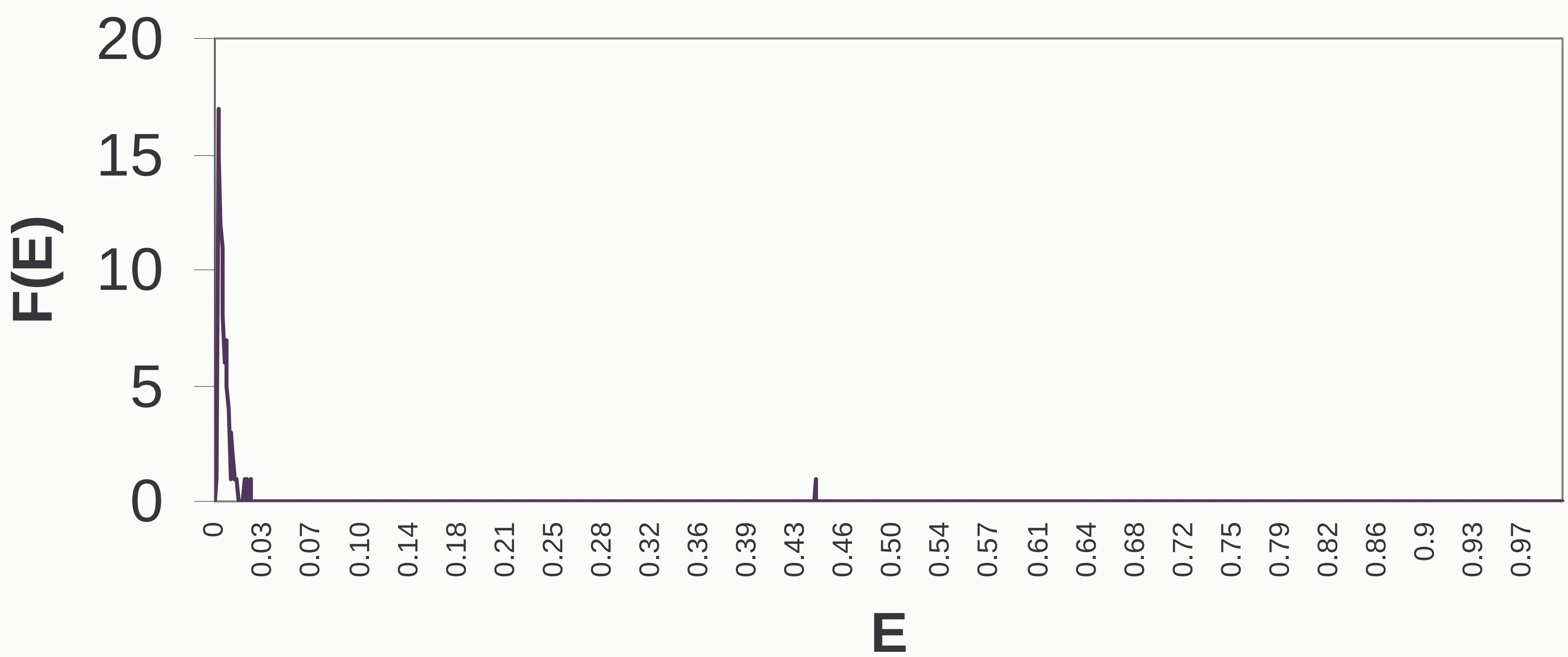
-- Dynamics: evolution of the spectrum in time.

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

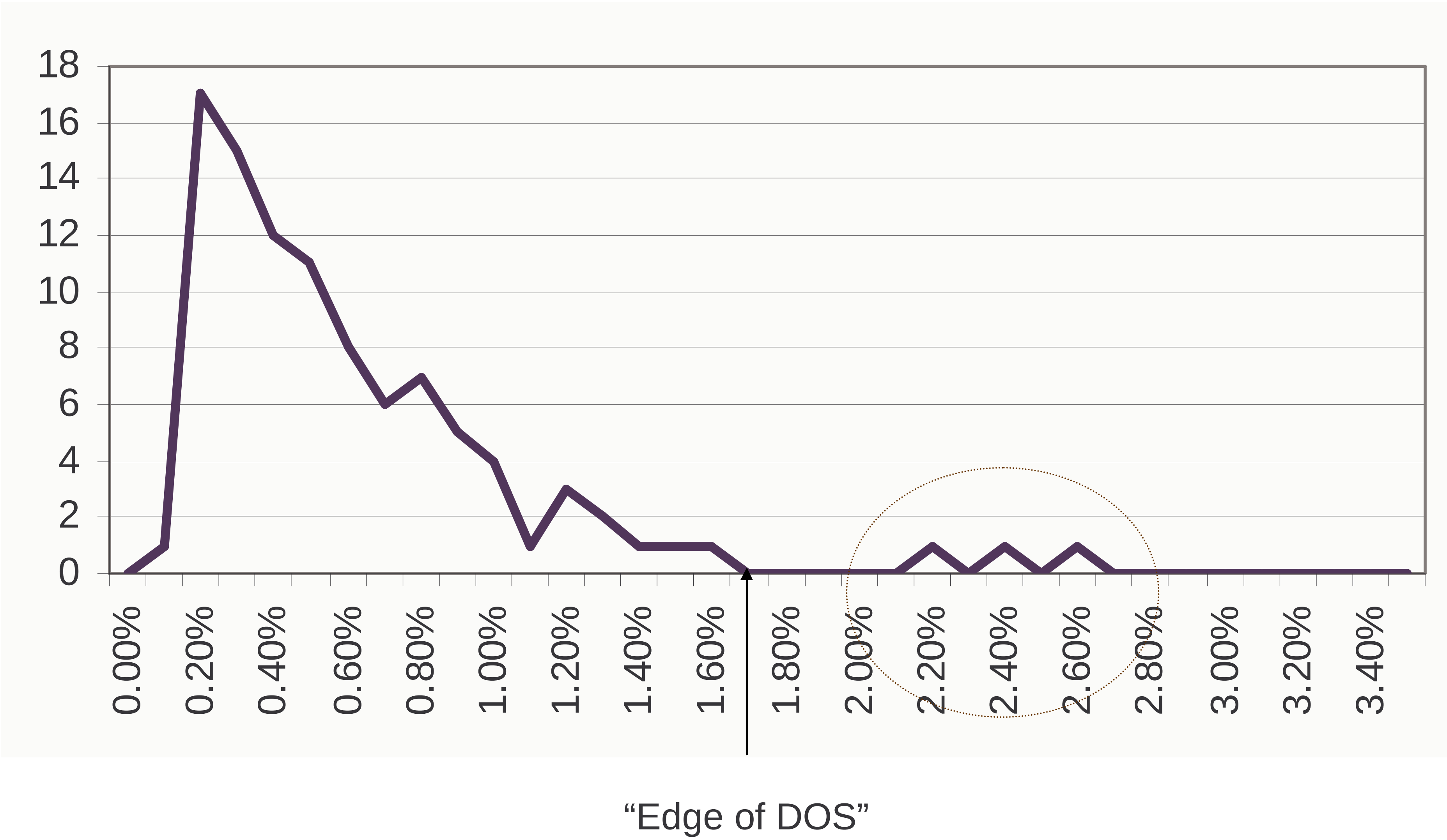


Density of States (from previous data)



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

Zoom of the DOS for low eigenvalues



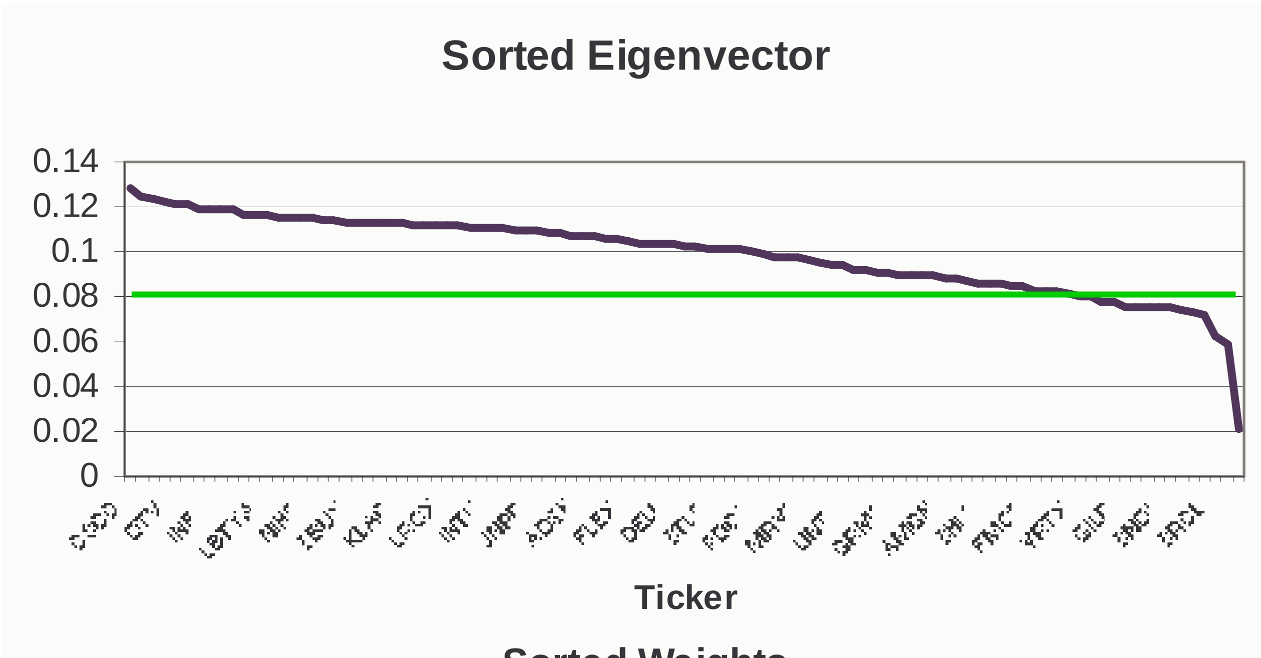
Results of PCA with DOS analysis

-- 4 significant eigenvectors/eigenvalues

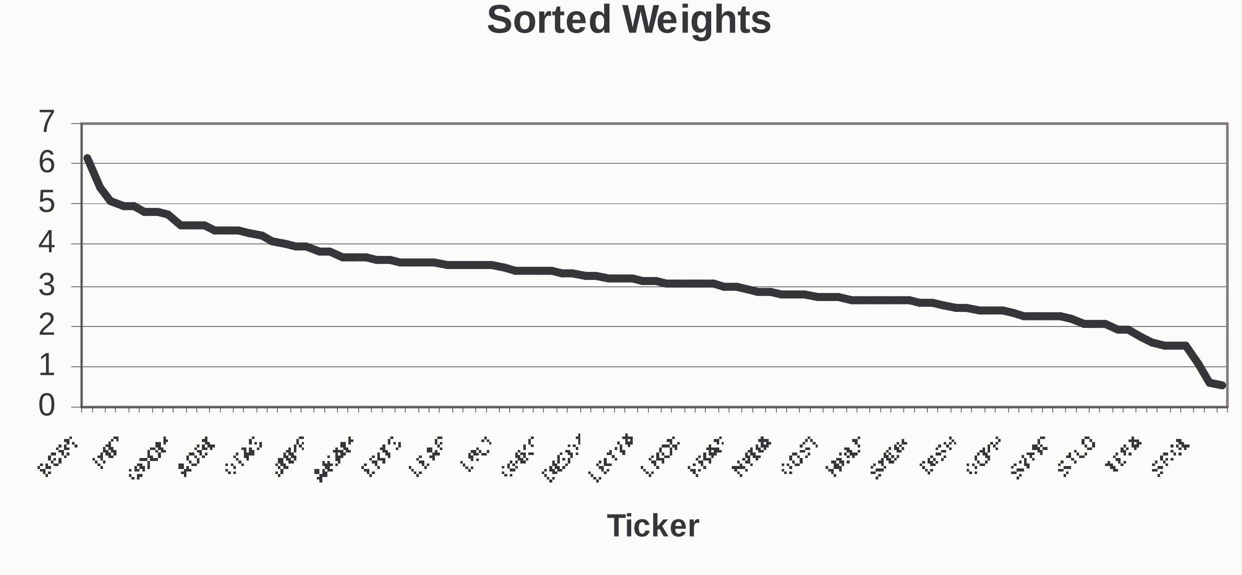
-- first Eigen-state explains about 44% of the correlation

-- total explained variance= 51%

First Eigenvector: Market

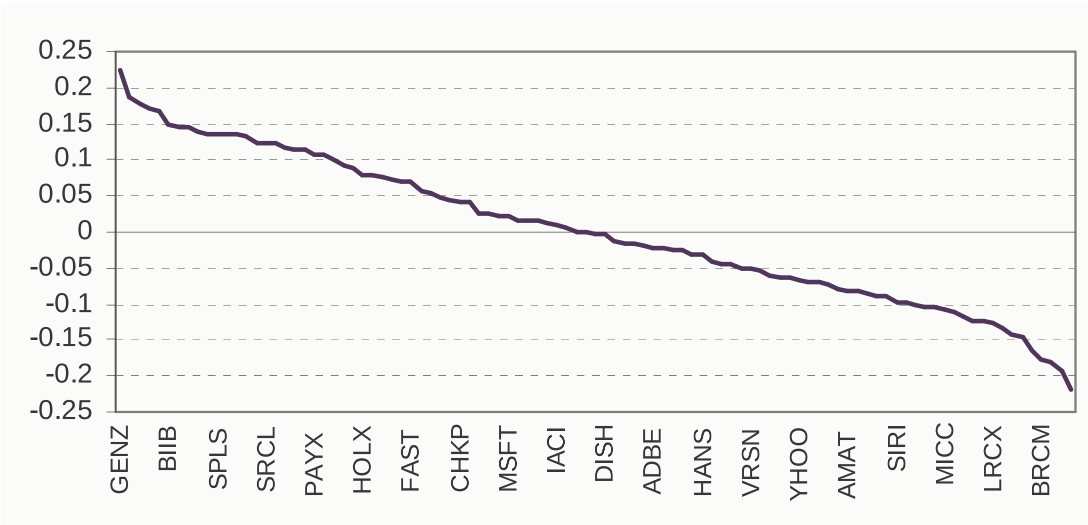


CSCO
INTC
ORCL
ADBE
DTV
SIAL
MSFT
SPLS
INFY
PAYX

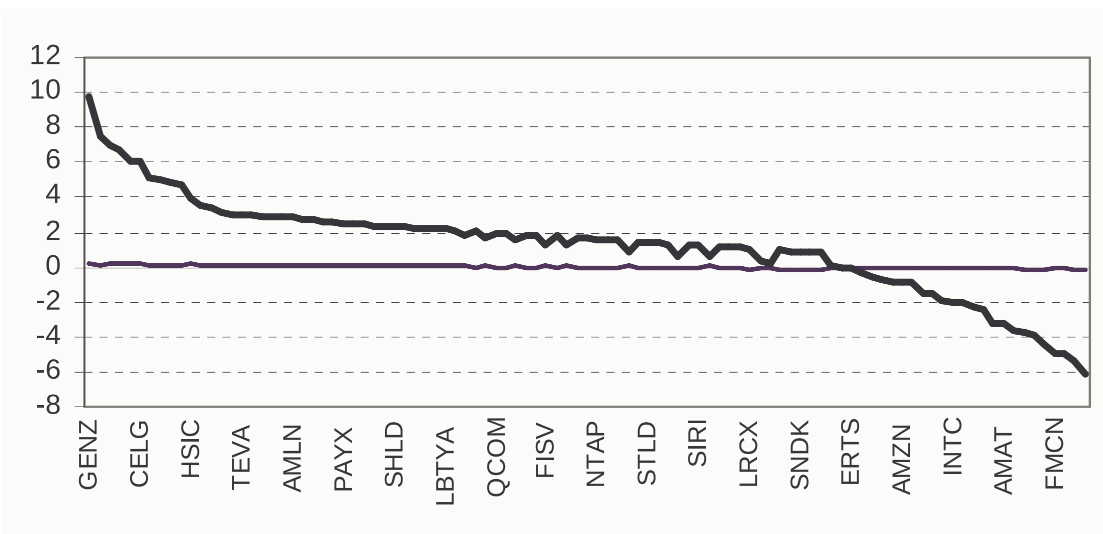


BBBY
LLTC
DELL
MRVL
INFY
ISRG
CMCSA
CTXS
QCOM
CSCO

Second Eigenvector: Biotech vs. Chips

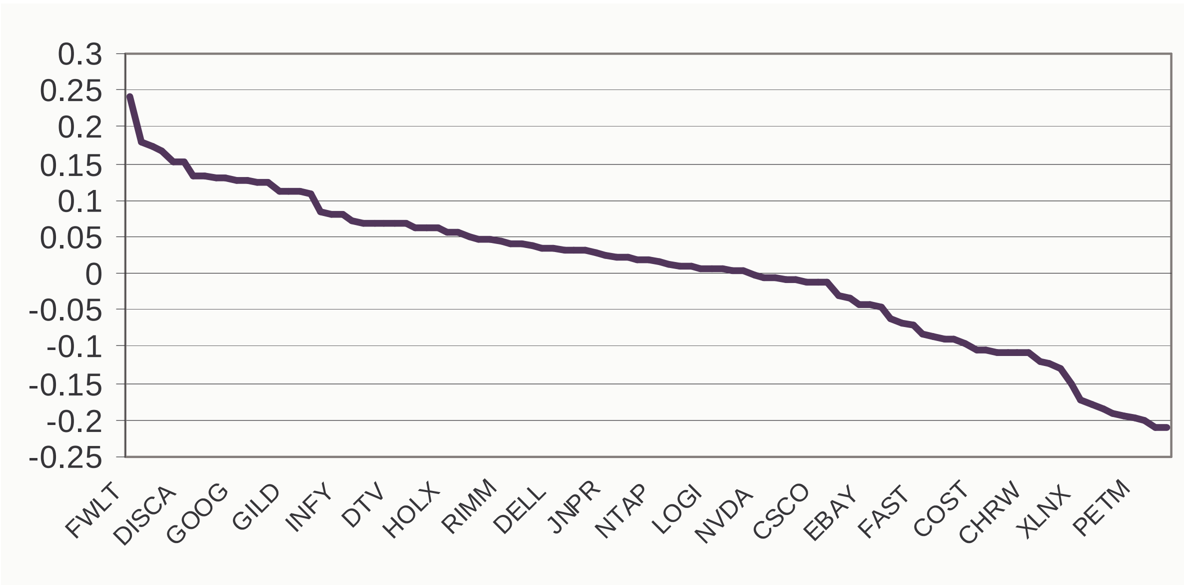


Top 10	Bottom 10
GENZ	MRVL
CEPH	NVDA
HSIC	FWLT
CELG	BRCM
GILD	SNDK
BIIB	JOYG
XRAY	RIMM
AMLN	BIDU
CTAS	LRCX
ESRX	ALTR

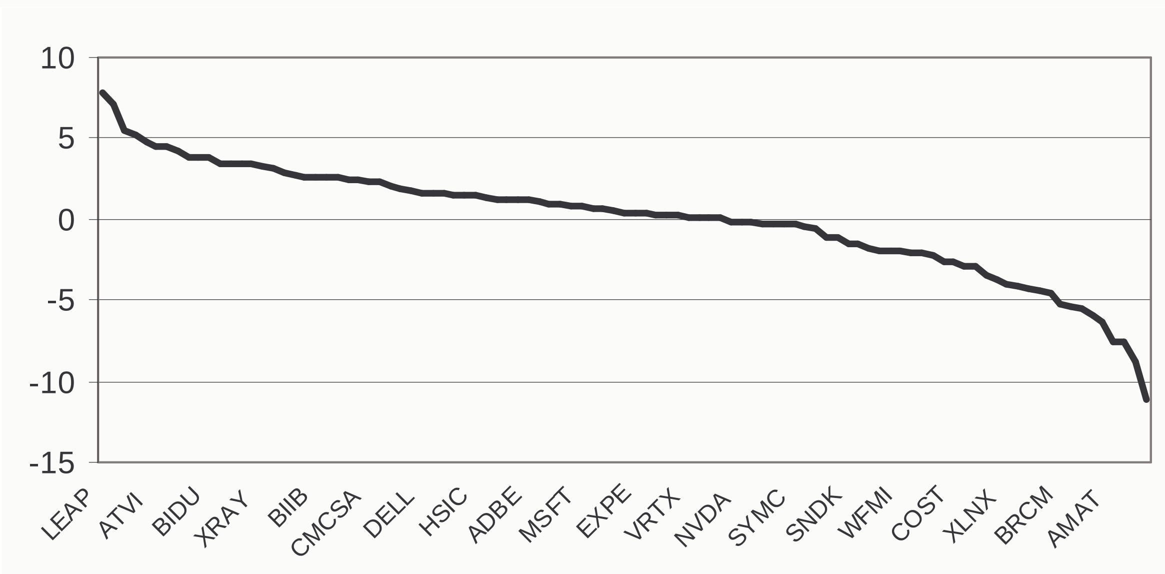


Top 10	Bottom 10
GENZ	BRCM
BBBY	FWLT
BIIB	DELL
GILD	FMCN
CEPH	AKAM
CELG	BIDU
ESRX	ALTR
CTAS	FLEX
AMGN	AMAT

Third Eigenvector: Manufacturing vs. Chips



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



Top 10	Bottom 10
LEAP	BBBY
FWLT	LLTC
DISCA	SHLD
FMCN	AMAT
NIHD	ALTR
ATVI	PETM
GILD	MRVL
CEPH	LRCX
GOOG	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 of sector

Random Matrix Theory

$$X_{tn},\ t = 1,2,..,T, \quad n = 1,2,...,N$$

$$X \sim N(0,1)$$

$$W_{mn} = \sum_{t=1}^{T} X_{tm} X_{tn}, \qquad \mathbf{W} = \mathbf{X}^{\mathbf{t}}\mathbf{X}$$

$$\lambda_n,\ n = 1,2,...N \qquad \text{eigenvalues of } \mathbf{W}$$

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*?

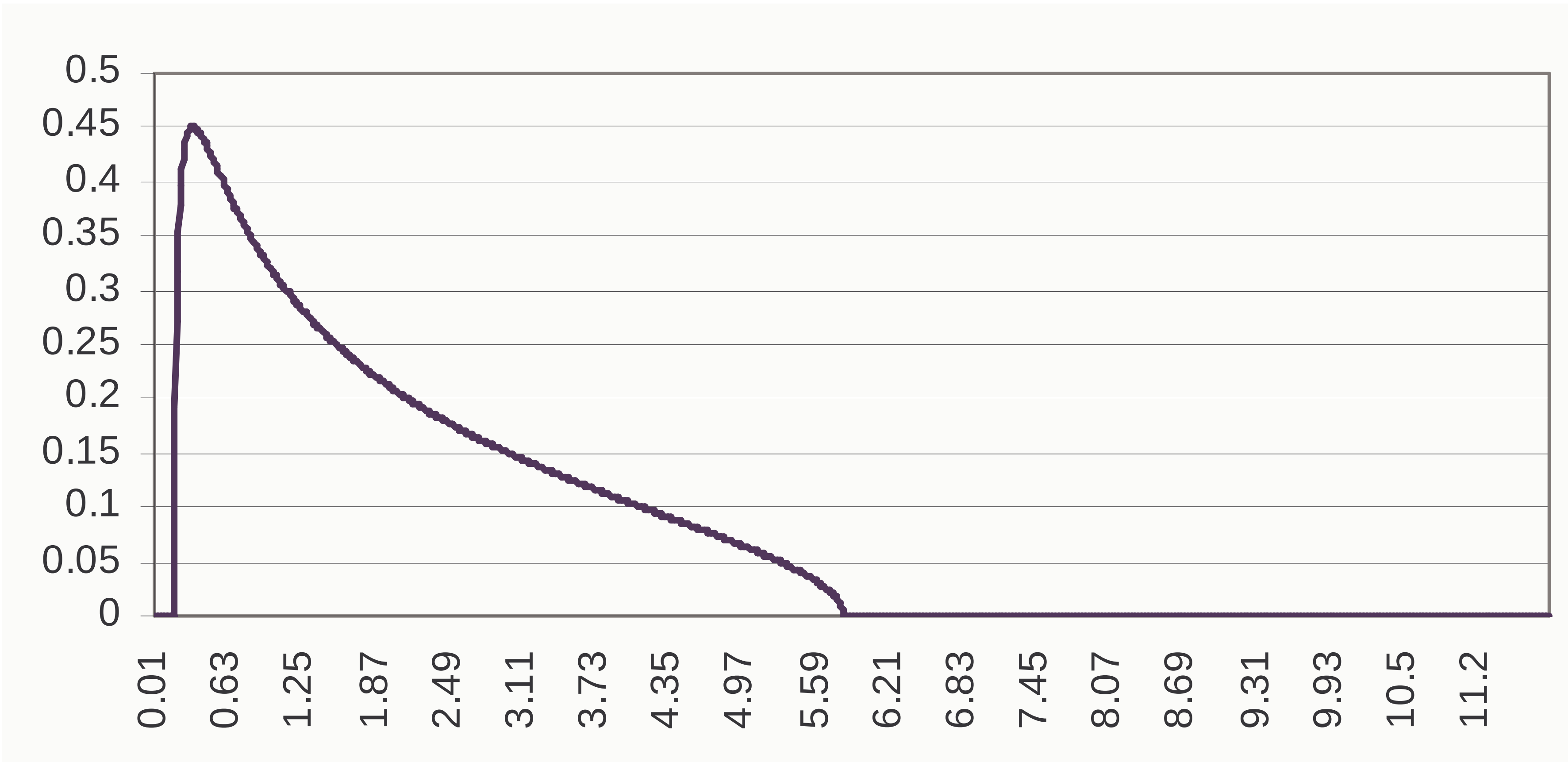
Marcenko-Pastur Theorem

$$F(E) = \lim_{\substack{N\to\infty \\ T\to\infty \\ N/T\to\gamma}} \frac{\#\{k : \lambda_k / N \le E\}}{N} \quad \text{integrated DOE}$$

$$f(E) = \frac{dF(E)}{dE}$$

$$\boxed{f(E) = \frac{1}{2\pi\gamma} \frac{\sqrt{(E_+ - E)(E - E_-)}}{E} \qquad E_\pm = \left(1 + \sqrt{\gamma}\right)^2}$$

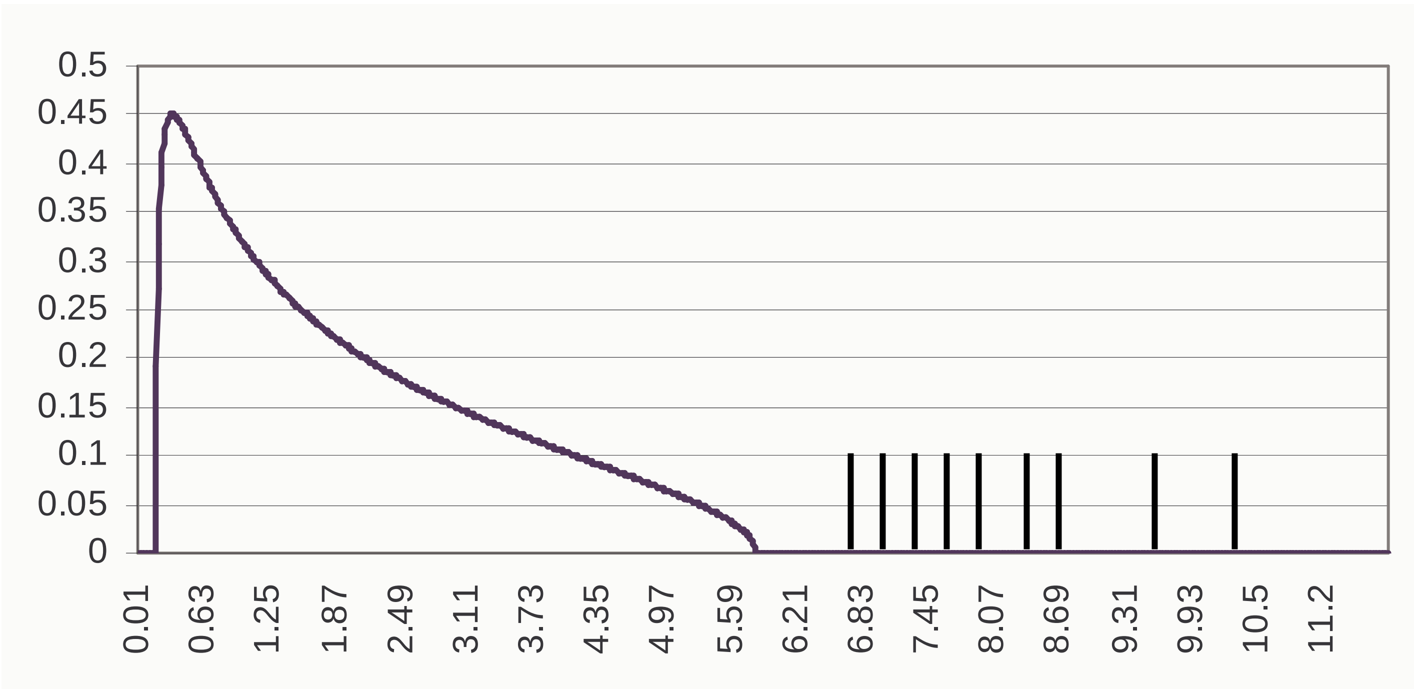
Marcenko Pastur (N/T=2)



DOS for a random Gaussian ensemble

Bouchaud, Cizeau, Laloux, Potters (PRL, 1999)

The bulk distribution is described approximately by Marcenko Pastur properly normalized



Bulk spectrum significant eigenvalues