Principal Component Analysis of High-Frequency Data

Yacine Aït-Sahalia
Department of Economics
Princeton University and NBER

Dacheng Xiu
Booth School of Business
University of Chicago

NBER NSF Conference, Vienna
Sep, 2015

1. introduction

1.1. Classical PCA and Its Applications

- Principal component analysis (PCA) is one of the most popular and oldest techniques for multivariate analysis.

  1. Pearson (1901, Philosophical Magazine)

  2. Hotelling (1933, J. Educ. Psych.)

- PCA is a dimension reduction technique that seeks a parsimonious representation of the multivariate structure of the data.

- PCA is widely used in macroeconomics and finance.

  1. Litterman and Scheinkman (1991): three-factor (level, slope, and curvature) structure of the term structure of yields

  2. Egloff, Leippold, and Wu (2010): two-factor (long-run and short-run) volatility term structure

  3. Stock and Watson (1999): Chicago Fed National Activity Index

  4. Baker and Wurgler (2006): sentiment measure

  5. Baker, Bloom, and Davis (2013): policy uncertainty index

1.2. Statistical Inference on Classical PCA

- Estimating eigenvalues of the sample covariance matrix is the key step towards PCA.

- Anderson (1963, AOS) studies the statistical inference problem of the eigenvalues, and find that

$$\sqrt{n}(\widehat{\lambda} - \lambda) \xrightarrow{d} N\left(0, 2\mathrm{Diag}\left(\lambda_1^2, \lambda_2^2, \ldots, \lambda_d^2\right)\right).$$

  where $\widehat{\lambda}$ and λ are the vectors of eigenvalues of the sample and population covariance matrices. λ is simple.

- When eigenvalues are repeated, the asymptotic theory is rather complicated to use.

1.3. Drawbacks and Limitations of the Classical PCA

- i.i.d. and multivariate normality
  - Extensions to non-normality or time-series data is possible, but typically more assumptions are needed, e.g. a parametric factor model, stationarity.

- the curse of dimensionality
  - It is well known that when $n/d \to C \geq 1$,

$$\widehat{\lambda}_1 \xrightarrow{a.s.} (1 + C^{-1/2})^2$$

where the true eigenvalue is 1.

- linear restriction
  - The identified “factors” are linear combinations of the data.

- static representation
  - The factor loadings are constants determined by the entire sample.

These drawbacks hinder the application of the PCA with financial data.

- Stock returns exhibit time-varying volatility and heavy tails, which deviate from i.i.d. normality to a great extent.

- The covariance matrix of 30 stocks has 465 parameters, if no additional structure is imposed. Years of daily data are required, raising the issue of survivorship bias, potential non-stationarity, and parameter constancy.

- Option returns are nonlinear functions of the underlying factors, stock prices, volatility, etc.

- Moreover, the factor loadings are time-varying.

1.4. Main Contribution of the Paper

- We define the concept of **(realized) PCA** for data sampled from a continuous-time stochastic process within a fixed time window.

- We propose asymptotic theory for **spectral functions**, **eigenvalues**, **eigenvectors**, and **principal components**, under general nonparametric models, using intraday data.

- Empirically, we use this new technique to analyze constituents of S&P 100 Index nonparametrically.

1.5. Literature

PCA and Factor Models

- Applications in Finance and Macro: Ross (1976, JET), Stock and Watson (2002, JBES), Litterman and Scheinkman (1991, J Fixed Income)

- Classic PCA and Factor Analysis: Hotelling (1933, J. Educ. Psych.), Thomson (1934, J. Educ. Psych.), Anderson and Amemiya (1988, AOS).

- Large d Setting: Chamberlain and Rothschild (1983, ECMA), Connor and Korajczyk (1998, JFE), Stock and Watson (2002, JASA), Bai and NG (2002, ECMA), Bai (2003, ECMA).

Our model consists of Itô semimartingales, and is fully nonparametric without any assumptions on the existence of a factor structure

Other Related HF Papers

- Fixed T, Fixed d, Small Δ: Eigenvalue Related Problems
  - Test of Rank: Jacod, Lejay, and Talay (2008, Bernoulli), Jacod and Podolskij (2013, AOS).
- Fixed T, Large d, Small Δ:
  - Sparse Covariance Matrix Estimation: Wang and Zou (2010, JASA), Tao, Wang, and Zhou (2013, AOS), Tao, Wang, and Chen (2013, ET), and Tao, Wang, Yao, and Zou (2011, JASA).
  - Continuous-Time Factor Model for High-Frequency Panel: Fan, Furger, and Xiu (2015, JBES) and Aït-Sahalia and Xiu (2015).

2. Main Theory

2.1. Review of Classical PCA

Suppose R is a d-dimensional vector-valued random variable. The first component is a linear combination of R, $\gamma_1^\intercal R$, which maximize its variation. The weight γ_1 satisfies the following optimization problem:

$$\max_{\gamma_1} \gamma_1^\intercal c \gamma_1, \quad \text{subject to} \quad \gamma_1^\intercal \gamma_1 = 1$$

where $c = \text{cov}(R)$. Using the Lagrange multiplier, the problem is to maximize

$$\gamma_1^\intercal c \gamma_1 - \lambda_1(\gamma_1^\intercal \gamma_1 - 1)$$

which yields $c\gamma_1 = \lambda_1 \gamma_1$, and $\text{var}(\gamma_1^\intercal R) = \gamma_1^\intercal c \gamma_1 = \lambda_1$.

- Therefore, λ_1 is the largest eigenvalue of the population covariance matrix c, and γ_1 is the corresponding eigenvector.

- The second principal component solves the following optimization problem:

  $$\max_{\gamma_2} \gamma_2^\intercal c \gamma_2, \quad \text{subject to} \quad \gamma_2^\intercal \gamma_2 = 1, \text{ and } \mathrm{cov}(\gamma_1^\intercal R, \gamma_2^\intercal R) = 0.$$

  It turns out that the solution γ_2 corresponds to the second eigenvalue λ_2.

- One can keep doing this, regardless of whether the eigenvalues are simple or repeated.

2.2. Continuous-Time Factor Model

We consider a d-dimensional Itô semimartingale, defined on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ with the following representation:

$$X_t = X_0 + \int_0^t b_s \, ds + \int_0^t \sigma_s dW_s + J_t,$$

and $c_t = (\sigma\sigma^\intercal)_t$ is another Itô semimartingale. J_t is of infinite activity and finite variation.

2.3. Principal Component Analysis

How do we introduce PCA in this setting?

- Instead of maximizing the **variance**, we maximize the continuous component of the **quadratic variation.**

- **Theorem:** There exists a sequence of $\{\lambda_{g,s}, \gamma_{g,s}\}$, $1 \leq g \leq d$, $0 \leq s \leq t$, such that

$$c_s \gamma_{g,s} = \lambda_{g,s} \gamma_{g,s}, \quad \gamma_{g,s}^{\intercal} \gamma_{g,s} = 1, \quad \text{and} \quad \gamma_{h,s}^{\intercal} c_s \gamma_{g,s} = 0,$$

where $\lambda_{1,s} \geq \lambda_{2,s} \geq \ldots \geq \lambda_{d,s} \geq 0$. Moreover, for any càdlàg and vector-valued adapted process γ_s, such that $\gamma_s^{\intercal} \gamma_s = 1$, and for $1 \leq h \leq g-1$,

$$\left[\int_0^u \gamma_{s-}^{\intercal} dX_s, \int_0^u \gamma_{h,s-}^{\intercal} dX_s\right]^c = 0, \text{ and } \int_0^u \lambda_{g,s} ds \geq \left[\int_0^u \gamma_{s-}^{\intercal} dX_s, \int_0^u \gamma_{s-}^{\intercal} dX_s\right]^c.$$

2.4. Estimation Strategy

- The idea is simple.

  1. Decompose the interval $[0, t]$ into many subintervals

  2. Estimate c_s within each subinterval using sample covariance matrix.

  3. Aggregate the eigenvalues of $\widehat{c}_s$, $\lambda(\widehat{c}_s)$.

- Apparently, we need some idea about the derivatives of $\lambda(\cdot)$ with respect to a matrix, as the estimation error depends on the smoothness of $\lambda(\cdot)$.

- **Lemma:** The function $\lambda : \mathcal{M}_d^+ \to \bar{\mathbb{R}}_d^+$ is **Lipchitz**.
  - $\bar{\mathbb{R}}_d^+$ is the subset of ordered nonnegative numbers of $\mathbb{R}_d$.
  - $\mathcal{M}_d^+$ is the space of non-negative matrices.

- Moreover, λ_g, if **simple**, is a C^∞-function. So is its corresponding eigenvector γ_g, which is unique up to a sign.

2.5. Spectral Functions

- It turns out we should consider estimating an even more general quantity $\int_0^t F(c_s)ds$.

- A spectral function is a function of non-negative matrices, which satisfies, $F(c) = F(O^{\intercal} c O)$, for any orthogonal matrix O.

- In other words, a spectral function depends on the matrix only through its eigenvalues.

- Therefore, we can write $F(c) = (f \circ \lambda)(c)$.

- F is spectral $\Longleftrightarrow$ f is symmetric, i.e. $f(Px) = f(x)$, for any vector $x \in \bar{\mathbb{R}}_d^+$, where P is a permutation matrix.

- **Lemma:** The symmetric function f is kth continuously differentiable at a point $\lambda(c) \in \bar{\mathbb{R}}_d^+$ if and only if the spectral function $F = f \circ$

λ is kth continuously differentiable at the point $c \in \mathcal{M}_d^+$, for $k = 0, 1, 2, \ldots, \infty$. The gradient and the Hessian matrix are given below:

$$\partial_{jk}(f \circ \lambda)(A) = \sum_{p=1}^{d} O_{pj} \partial_p f(\lambda(A)) O_{pk},$$

$$\partial^2_{jk,lm}(f \circ \lambda)(A) = \sum_{p,q=1}^{d} \partial^2_{pq} f(\lambda(A)) O_{pl} O_{pm} O_{qj} O_{qk} + \sum_{p,q=1}^{d} \mathcal{A}^f_{pq}(\lambda(A)) O_{pl} O_{pj} O_{qk} O_{qm},$$

where O is any orthogonal matrix that satisfies $A = O^{\mathsf{T}} \mathrm{Diag}\left(\lambda(A)\right) O$. Moreover, f is convex if and only if F is.

Examples

- $f(x) = \sum_{j=1}^{d} x_j \implies F(c) = \mathrm{Tr}(c)$.

- $f(x) = \prod_{j=1}^{d} x_j \implies F(c) = \det(c)$

- $f(x) =$ the k-th largest entry in $x =: \bar{x}_k$.
  - $F(c) =$ the k-th eigenvalue of c and it is simple.
  - This function is only differentiable when $x_{k-1} > x_k > x_{k+1}$.

- $f(x) = \frac{1}{g_l - g_{l-1}} \sum_{j=g_{l-1}+1}^{g_l} \bar{x}_j$.
  - $F(c) =$ the k-th eigenvalue of c, and it is repeated.
  - This function is differentiable when $x_{g_{l-1}} > x_{g_{l-1}+1} \geq \ldots \geq x_{g_l} > x_{g_l+1}$.

- Many spectral functions are only differentiable at certain points of the matrix space, because of the potential repeated eigenvalues.

- We need a more careful analysis of the topology of the set of differentiable matrices.

Topology of Sets of Differentiable Matrices

- **Lemma** For any $1 \leq g_1 < g_2 < \ldots < g_r \leq d$, the set
$$\begin{aligned}&\mathcal{M}(g_1, g_2, \ldots, g_r)\\ =&\{A \in \mathcal{M}_d^{++} \mid \lambda_{g_l}(A) > \lambda_{g_l+1}(A),\ \text{for any}\ \ l = 1, 2, \ldots, r-1\}\end{aligned}$$
is dense and open in $\mathcal{M}_d^{++}$. In particular, the set of positive-definite matrices with distinct eigenvalues, i.e., $\mathcal{M}(1, 2, \ldots, d)$, is dense and open in $\mathcal{M}_d^{++}$.

2.6. Asymptotic Theory

2.6.1. Estimator

1. Denote the distance between adjacent observations by Δ_n. We form blocks of length k_n. At each $ik_n\Delta_n$, we can estimate $c_{ik_n\Delta_n}$ by

$$\widehat{c}_{ik_n\Delta_n} = \frac{1}{k_n\Delta_n}\sum_{j=1}^{k_n}\left(\Delta^n_{ik_n+j}X\right)^{\intercal}\left(\Delta^n_{ik_n+j}X\right)1_{\left\{\left\|\Delta^n_{ik_n+j}X\right\|\leq u_n\right\}}.$$

where $u_n = \alpha\Delta_n^{\varpi}$, and $\Delta_l^n X = X_{l\Delta_n} - X_{(l-1)\Delta_n}$.

2. Our estimator of $\int_0^t F(c_s)ds$ is

$$V(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} f\left(\widehat{\lambda}_{ik_n\Delta_n}\right),$$

where $\widehat{\lambda}_{ik_n\Delta_n} := \lambda(\widehat{c}_{ik_n\Delta_n})$.

2.6.2. Assumptions

- Suppose F is a vector-valued spectral function, and f is the corresponding vector-valued symmetric function such that $F = f \circ \lambda$. f is a continuous function, and satisfies $\|f(x)\| \leq K(1 + \|x\|^{\zeta})$, for some $\zeta > 0$.

- There exists some open and convex set $\mathcal{C}$, such that $\bar{\mathcal{C}} \subset \mathcal{M}(g_1, g_2, \ldots, g_r)$, where $1 \leq g_1 < g_2 < \ldots < g_r \leq d$, and that for any $0 \leq s \leq t$,

$c_s \in \mathcal{C} \cap \mathcal{M}^*(g_1, g_2, \ldots, g_r)$. Moreover, f is C^3 on $\mathcal{D}(g_1, g_2, \ldots, g_r)$. where $\mathcal{D}(g_1, g_2, \ldots, g_r) = \lambda(\mathcal{M}(g_1, g_2, \ldots, g_r))$, and

$$\begin{aligned} &\mathcal{M}^*(g_1, g_2, \ldots, g_r) \\ =&\Big\{ A \in \mathcal{M}_d^{++} \mid \lambda_1(A) = \ldots = \lambda_{g_1}(A) > \lambda_{g_1+1}(A) = \ldots \\ &= \lambda_{g_2}(A) > \ldots \lambda_{g_{r-1}}(A) > \lambda_{g_{r-1}+1}(A) = \ldots = \lambda_{g_r}(A) \Big\}. \end{aligned}$$

- If one is only interested in spectral functions that depends on a simple eigenvalue, then the assumption can be weakened:

  There exists some open and convex set $\mathcal{C}$, such that $\bar{\mathcal{C}} \subset \mathcal{M}(g)$, for some $g = 1, 2, \ldots, d$, and that for any $0 \leq s \leq t$, $c_s \in \mathcal{C}$. Moreover, f is C^3 on $\mathcal{D}(g)$.

2.6.3. Consistency and Failure of CLT

- **Theorem:** Suppose either $\zeta \leq 1$ or $\zeta > 1$ and $\varpi \in [\frac{\zeta-1)}{2\zeta-\gamma}, \frac{1}{2})$. Then our estimator is consistent: as $k_n \to \infty$ and $k_n\Delta_n \to 0$,

$$V(\Delta_n, X; F) \xrightarrow{\mathsf{p}} \int_0^t F(c_s)ds.$$

- **Theorem:** Suppose $k_n \asymp \Delta_n^{-\varsigma}$ and $u_n \asymp \Delta_n^{\varpi}$ for some $\varsigma \in (\frac{\gamma}{2}, \frac{1}{2})$ and $\varpi \in [\frac{1-\varsigma}{2-\gamma}, \frac{1}{2})$. As $\Delta_n \to 0$,

$$k_n\left(V(\Delta_n, X; F) - \int_0^t F(c_s)ds\right)$$
$$\xrightarrow{\mathsf{p}} \frac{1}{2}\sum_{j,k,l,m=1}^{d} \int_0^t \partial^2_{jk,lm}F(c_s)\left(c_{jl,s}c_{km,s} + c_{jm,s}c_{kl,s}\right)ds.$$

2.6.4. The Bias-Corrected Estimator and its CLT

- The Bias-Corrected Estimator:

$$\widetilde{V}(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \Big\{ F(\widehat{c}_{ik_n\Delta_n}) - \frac{1}{2k_n} \times$$
$$\sum_{j,k,l,m=1}^{d} \partial^2_{jk,lm} F(\widehat{c}_{ik_n\Delta_n}) \left(\widehat{c}_{jl,ik_n\Delta_n}\widehat{c}_{km,ik_n\Delta_n} + \widehat{c}_{jm,ik_n\Delta_n}\widehat{c}_{kl,ik_n\Delta_n} \right) \Big\}.$$

- **Theorem (CLT):** $k_n \asymp \Delta_n^{-\varsigma}$ and $u_n \asymp \Delta_n^{\varpi}$ for some $\varsigma \in (\frac{\gamma}{2}, \frac{1}{2})$ and $\varpi \in [\frac{1-\varsigma}{2-r}, \frac{1}{2})$. As $\Delta_n \to 0$, we have,

$$\frac{1}{\sqrt{\Delta_n}} \left(\widetilde{V}(\Delta_n, X; F) - \int_0^t F(c_s)ds \right) \xrightarrow{\mathcal{L}-\mathsf{s}} \mathcal{W}_t.$$

2.6.5. Return to the Integrated Eigenvalues

- Previous assumption implies that eigenvalues maintain the following structure:

$$\begin{aligned}\lambda_1(c_s) = \ldots = \lambda_{g_1}(c_s) > \lambda_{g_1+1}(c_s) = \ldots = \lambda_{g_2}(c_s) > \ldots \lambda_{g_{r-1}}(c_s)\\ > \lambda_{g_{r-1}+1}(c_s) = \ldots = \lambda_{g_r}(c_s) > 0.\end{aligned}$$

  Moreover, their structure, i.e. $\{g_1, g_2, \ldots, g_r\}$, does not vary over $[0, t]$. **Not needed for consistency**. Moreover, if we only care about one eigenvalue block, say, the largest, then this assumption can be weakened to $\lambda_{g_1}(c_s) > \lambda_{g_1+1}(c_s)$.

- Choose the spectral function F accordingly:

$$F^{\lambda}(\cdot) = \left(\frac{1}{g_1} \sum_{j=1}^{g_1} \lambda_j(\cdot), \frac{1}{g_2 - g_1} \sum_{j=g_1+1}^{g_2} \lambda_j(\cdot), \ldots, \frac{1}{g_r - g_{r-1}} \sum_{j=g_{r-1}+1}^{g_r} \lambda_j(\cdot) \right)^{\intercal}.$$

2.6.6. The CLT of Eigenvalues

- **Corollary:** The bias-corrected estimator takes on the following form:

$$\widetilde{V}(\Delta_n, X; F_p^{\lambda}) = \frac{\Delta_n}{g_p - g_{p-1}} \sum_{i=0}^{[t/(k_n\Delta_n)]} \sum_{h=g_{p-1}+1}^{g_p} \left\{ \widehat{\lambda}_{h,ik_n\Delta_n} - \frac{1}{k_n} \mathrm{Tr}\left((\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n})^{+} \widehat{c}_{ik_n\Delta_n} \right) \widehat{\lambda}_{h,ik_n\Delta_n} \right\}.$$

The asymptotic covariance matrix is given by

$$\mathrm{E}(\mathcal{W}_t^{\lambda}(\mathcal{W}_t^{\lambda})^{\intercal}|\mathcal{F}) = \begin{pmatrix} \frac{2}{g_1}\int_0^t \lambda_{g_1,s}^2 ds & & & \\ & \frac{2}{g_2-g_1}\int_0^t \lambda_{g_2,s}^2 ds & & \\ & & \ddots & \\ & & & \frac{2}{g_r-g_{r-1}}\int_0^t \lambda_{g_r,s}^2 ds \end{pmatrix}.$$

2.6.7. Eigenvectors

Suppose $\gamma_{g,s}$ is a vector-valued function that corresponds to the eigenvector of c_s with respect to a **simple** $\lambda_{g,s}$, for each $s \in [0,t]$. We have

$$\frac{1}{\sqrt{\Delta_n}}\left(\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]}\left(\widehat{\gamma}_{g,ik_n\Delta_n} + \frac{1}{2k_n}\sum_{p\neq g}\frac{\widehat{\lambda}_{g,ik_n\Delta_n}\widehat{\lambda}_{p,ik_n\Delta_n}}{(\widehat{\lambda}_{g,ik_n\Delta_n} - \widehat{\lambda}_{p,ik_n\Delta_n})^2}\widehat{\gamma}_{g,ik_n\Delta_n}\right) - \int_0^t \gamma_{g,s} ds\right)$$

$$\xrightarrow{\mathcal{L}-\mathsf{s}} W_t^{\gamma},$$

where the covariance matrix is given by

$$\mathsf{E}(\mathcal{W}_t^{\gamma}(\mathcal{W}_t^{\gamma})^{\mathsf{T}}|\mathcal{F}) = \int_0^t \lambda_{g,s}\left((\lambda_{g,s}\mathbb{I} - c_s)^{+} c_s (\lambda_{g,s}\mathbb{I} - c_s)^{+}\right) ds.$$

2.6.8. Principal Components

- Suppose $\gamma_{g,s}$ is a vector-valued function that corresponds to the eigenvector of c_s with respect to a **simple** root $\lambda_{g,s}$, for each $s \in [0, t]$. We have

$$\sum_{i=1}^{[t/(k_n\Delta_n)]-1} \widehat{\gamma}^{\mathsf{T}}_{g,(i-1)k_n\Delta_n}(X_{(i+1)k_n\Delta_n} - X_{ik_n\Delta_n}) \xrightarrow{\mathsf{p}} \int_0^t \gamma^{\mathsf{T}}_{g,s-} dX_s.$$

- So far, we have estimated $\int_0^t \lambda(c_s)ds$, $\int_0^t \gamma^{\mathsf{T}}_{g,s-}dX_s$, and $\int_0^t \gamma_{g,s}ds$.

2.7. PCA on Integrated Covariance?

Why not apply the usual PCA technique to the integrated covariance matrix $\int_0^t c_s ds$?

- The "eigenvalues" and "principal components" do not have the usual interpretations, i.e. the first eigenvalue $\lambda_1(\int_0^t c_s ds)$ is not the "variance(in any sense)" of the first principal component $\gamma_1^\intercal(X_t - X_0)$. By contrast:

$$\int_0^t \lambda_{1,s} ds = \left[\int_0^t \gamma_{1,s}^\intercal dX_s, \int_0^t \gamma_{1,s}^\intercal dX_s\right]^c.$$

3. Simulations and Empirical Work

3.1. Simulation Results

The cross-section of log stock prices X in continuous-time follows a factor model:

$$dX_t = \beta_t dF_t + dZ_t,$$

where F is unknown, and Z is a idiosyncratic component, orthogonal to Y. Y follows a three-factor model with stochastic volatility and jumps, whereas Z is a Brownian motion with jumps.

- We simulate intraday returns of up to 100 stocks at 5-second frequency spanning 1 week.

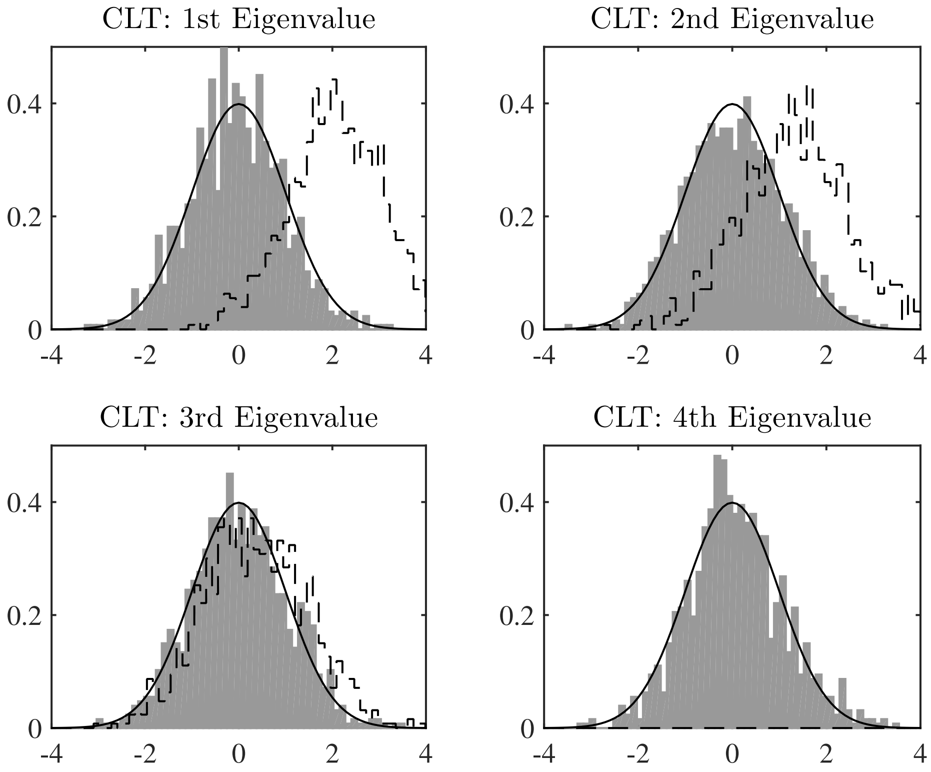
- There are 3 distinct eigenvalues which reflect the local factor structure of the simulated data.

- The remaining 97 population eigenvalues are identical, due to idiosyncratic variations.

$$\frac{d[X,X]_t^c}{dt} = \underbrace{\beta_t \frac{d[Y,Y]_t^c}{dt} \beta_t^{\intercal}}_{\text{Rank=3}} + \underbrace{\frac{d[Z,Z]_t^c}{dt}}_{\text{Diagonal Full Rank}}$$

- We fix $k_n = \theta \Delta_n^{-1/2} \sqrt{\log(d)}$, with $\theta = 0.05 - 0.25$ and d is the dimension of X.

We estimate 3 simple integrated eigenvalues as well as the average of the remaining identical eigenvalues.



	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.4686	-0.0001	0.0075	0.4703	0.0007	0.0152
10	0.6489	0.0001	0.0110	0.6552	-0.0014	0.0223
15	0.8927	-0.0004	0.0149	0.8975	-0.0011	0.0296
20	1.3044	0.0003	0.0225	1.3148	-0.0018	0.0424
30	2.1003	-0.0002	0.0356	2.1134	-0.0033	0.0688
50	2.9863	0.0002	0.0514	3.0104	-0.0054	0.1002
100	6.6270	-0.0004	0.1141	6.6732	-0.0127	0.2179
	1 Week, 5 Minutes			1 Month, 5 Minutes		
5	0.4879	0.0107	0.0452	0.5642	0.0006	0.0247
10	0.6839	0.0084	0.0574	0.7143	-0.0024	0.0258
15	0.9397	0.0128	0.0724	1.0167	-0.0024	0.0382
20	1.3765	0.0130	0.1076	1.3882	-0.0041	0.0503
30	2.2157	0.0210	0.1670	2.2383	-0.0073	0.0806
50	3.1554	0.0267	0.2410	3.1518	-0.0125	0.1155
100	7.0000	0.0552	0.5314	6.9632	-0.0270	0.2451

Table 1: 1st Eigenvalue Estimation

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.3145	0.00004	0.0051	0.3173	0.0003	0.0110
10	0.4268	-0.0001	0.0071	0.4289	0.0006	0.0146
15	0.5531	-0.0003	0.0086	0.5572	0.0004	0.0188
20	0.6517	-0.0008	0.0103	0.6556	-0.0006	0.0215
30	0.9186	-0.0015	0.0144	0.9251	-0.0017	0.0305
50	1.2993	-0.0017	0.0205	1.3080	-0.0026	0.0458
100	2.3273	-0.0041	0.0361	2.3441	-0.0065	0.0766
	1 Week, 5 Minutes			1 Month, 5 Minutes		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.3255	0.0020	0.0269	0.3639	-0.0003	0.0189
10	0.4384	0.0028	0.0379	0.4469	-0.0003	0.0186
15	0.5751	0.0014	0.0427	0.6524	-0.0017	0.0257
20	0.6758	0.0046	0.0535	0.7779	-0.0021	0.0297
30	0.9586	0.0017	0.0725	1.1365	-0.0036	0.0438
50	1.3508	-0.0022	0.1046	1.5630	-0.0064	0.0683
100	2.4312	-0.0084	0.1787	2.9148	-0.0136	0.1113

Table 2: 2nd Eigenvalue Estimation

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.1338	0.0001	0.0022	0.1345	0.0003	0.0044
10	0.2132	0.0001	0.0035	0.2149	0.0002	0.0070
15	0.2941	0.0002	0.0046	0.2954	0.0002	0.0093
20	0.3212	0.0001	0.0052	0.3242	0.0000	0.0105
30	0.4825	0.0005	0.0078	0.4853	0.0001	0.0153
50	0.6943	0.0001	0.0113	0.7016	-0.0005	0.0226
100	1.3808	0.0004	0.0221	1.3935	-0.0013	0.0438
	1 Week, 5 Minutes			1 Month, 5 Minutes		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.1364	0.0041	0.0124	0.1404	0.0006	0.0054
10	0.2199	0.0033	0.0183	0.2367	-0.0001	0.0091
15	0.3018	0.0056	0.0264	0.3396	0.0009	0.0131
20	0.3324	0.0037	0.0294	0.3546	0.0003	0.0132
30	0.4971	0.0055	0.0409	0.5686	0.0006	0.0211
50	0.7216	0.0028	0.0577	0.8051	-0.0010	0.0306
100	1.4302	-0.0016	0.1119	1.6278	-0.0028	0.0612

Table 3: 3rd Eigenvalue Estimation

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.0596	0.0002	0.0007	0.0597	0.0005	0.0015
10	0.0596	0.0001	0.0004	0.0597	0.0003	0.0008
15	0.0596	0.0001	0.0003	0.0597	0.0004	0.0006
20	0.0596	0.0001	0.0002	0.0597	0.0004	0.0005
30	0.0596	0.0001	0.0002	0.0597	0.0004	0.0004
50	0.0596	0.0001	0.0001	0.0597	0.0004	0.0003
100	0.0596	0.0001	0.0001	0.0597	0.0004	0.0002
	1 Week, 5 Minutes			1 Month, 5 Minutes		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.0595	0.0026	0.0041	0.0595	0.0006	0.0017
10	0.0595	0.0029	0.0029	0.0595	0.0006	0.0009
15	0.0595	0.0027	0.0024	0.0595	0.0006	0.0007
20	0.0595	0.0028	0.0021	0.0595	0.0006	0.0006
30	0.0595	0.0030	0.0020	0.0595	0.0007	0.0005
50	0.0595	0.0029	0.0018	0.0595	0.0006	0.0004
100	0.0595	0.0031	0.0016	0.0595	0.0006	0.0003

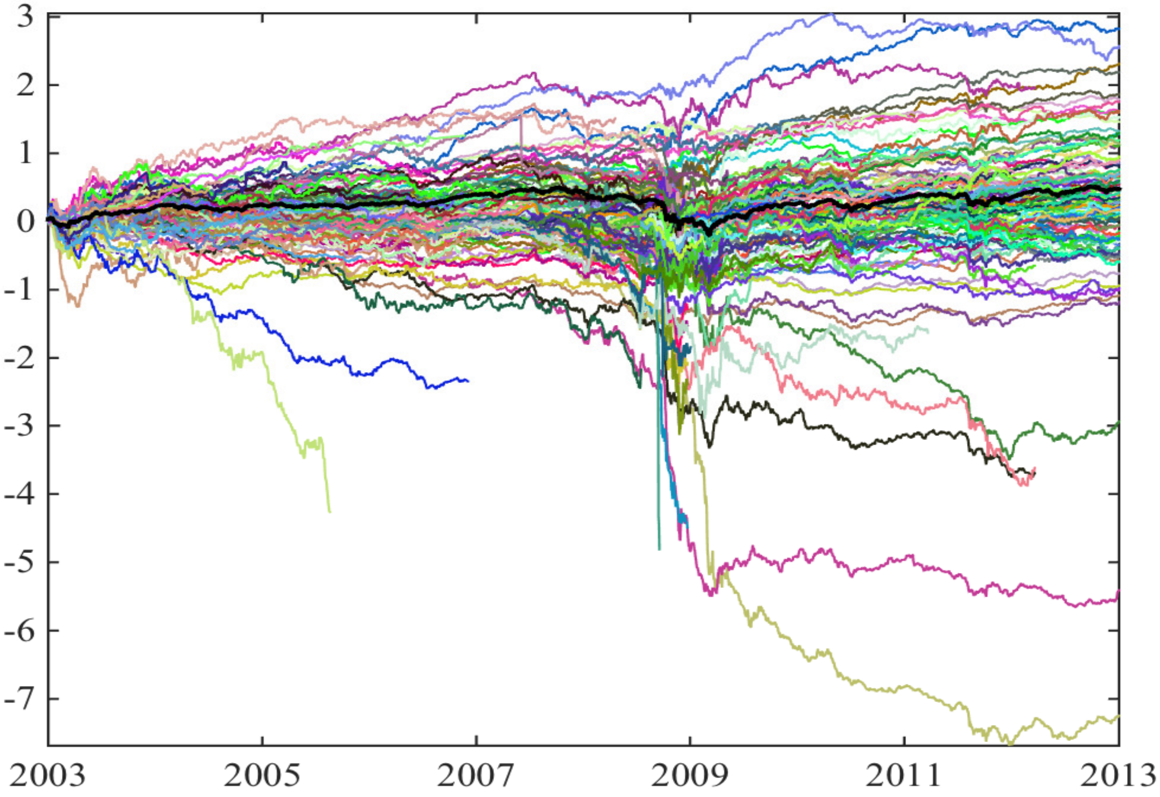
Table 4: Repeated Eigenvalue Estimation: 4th and Beyond

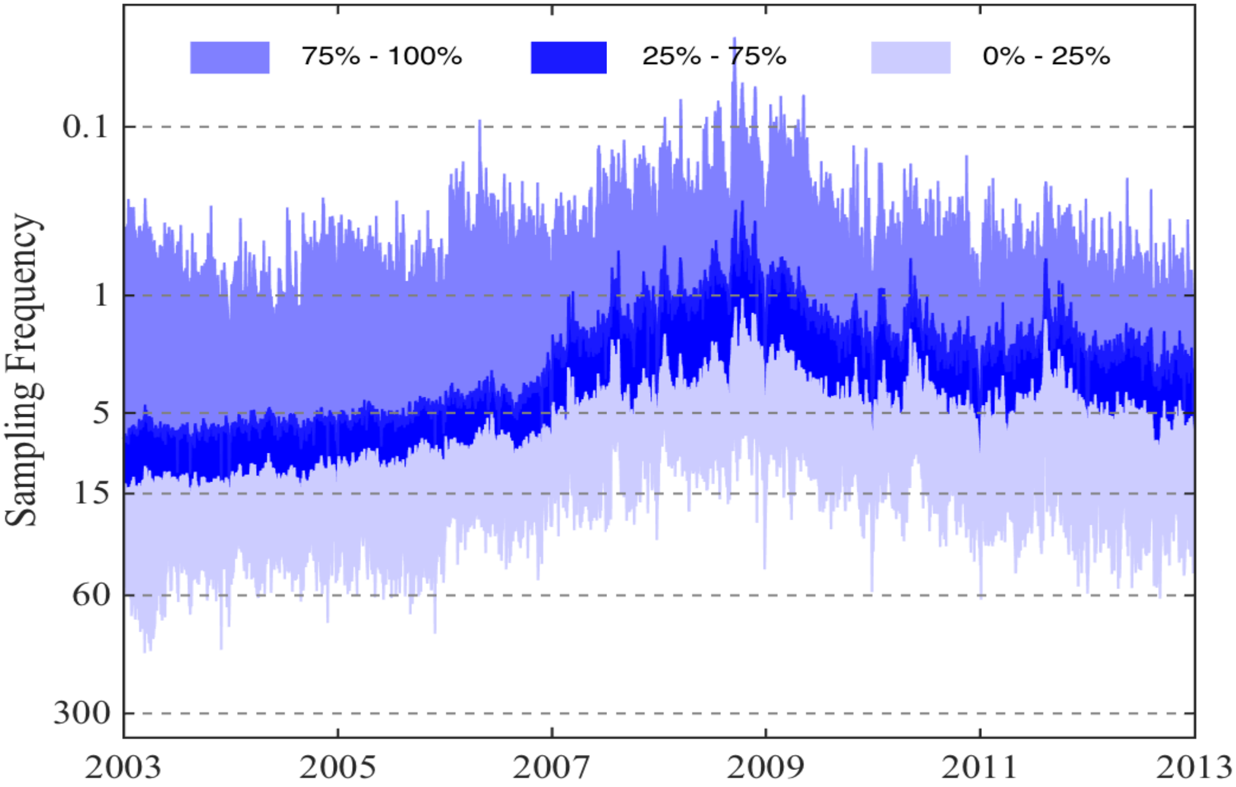
	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.2457	-0.0002	0.0173	0.2475	0.0074	0.1559
	0.8987	0.0028	0.0261	0.8963	0.0174	0.2135
	0.0566	0.0013	0.0235	0.0536	-0.0018	0.1281
	0.0633	0.0001	0.0119	0.0639	0.0037	0.0786
	0.3110	0.0017	0.0202	0.3086	0.0042	0.0645
10	0.0506	0.0004	0.0049	0.0495	0.0029	0.0266
	0.3444	0.0006	0.0087	0.3452	0.0080	0.0691
	0.4632	0.0011	0.0129	0.4619	0.0121	0.0966
	0.1422	0.0008	0.0137	0.1422	0.0094	0.0990
	0.4166	0.0004	0.0220	0.4164	0.0045	0.1562
	0.0864	0.0008	0.0158	0.0863	0.0089	0.1117
	0.3460	0.0008	0.0088	0.3440	0.0101	0.0600
	0.1268	0.0005	0.0076	0.1259	0.0066	0.0451
	0.3409	0.0005	0.0174	0.3415	0.0046	0.1268
	0.4262	0.0007	0.0112	0.4271	0.0099	0.0942

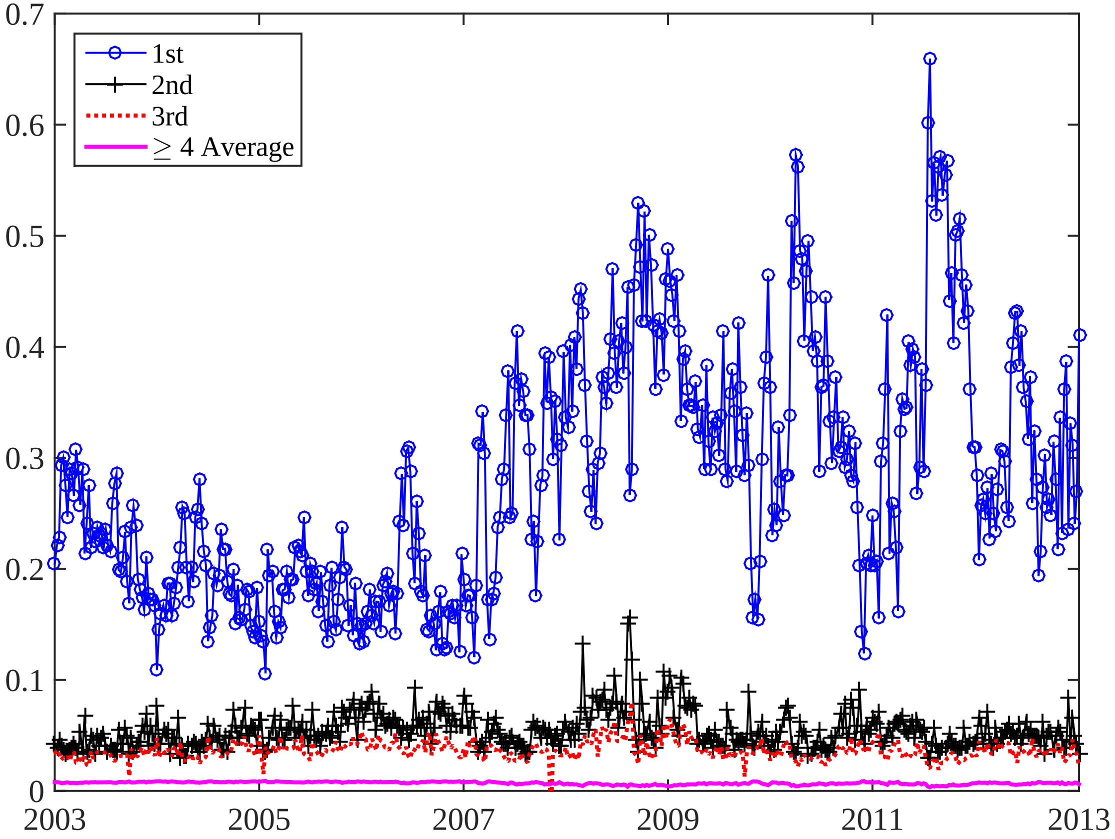
Table 5: 1st Eigenvector Estimation

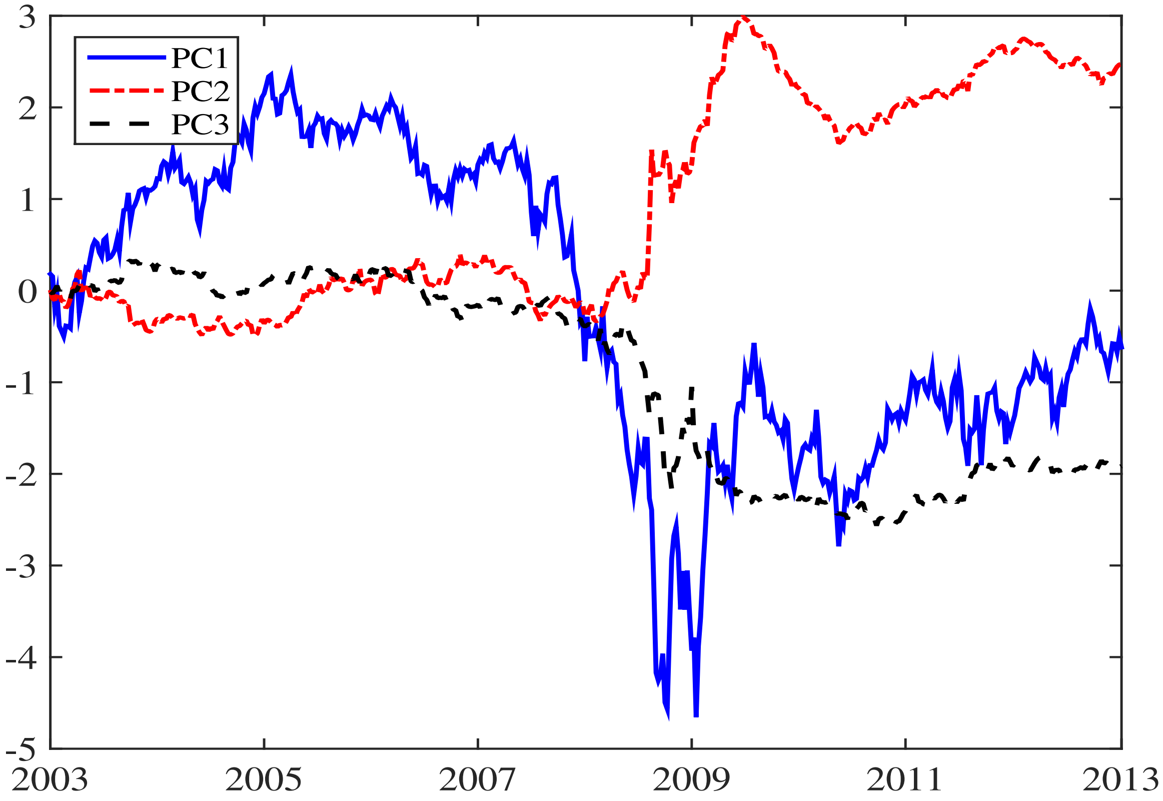
3.2. Empirical Findings

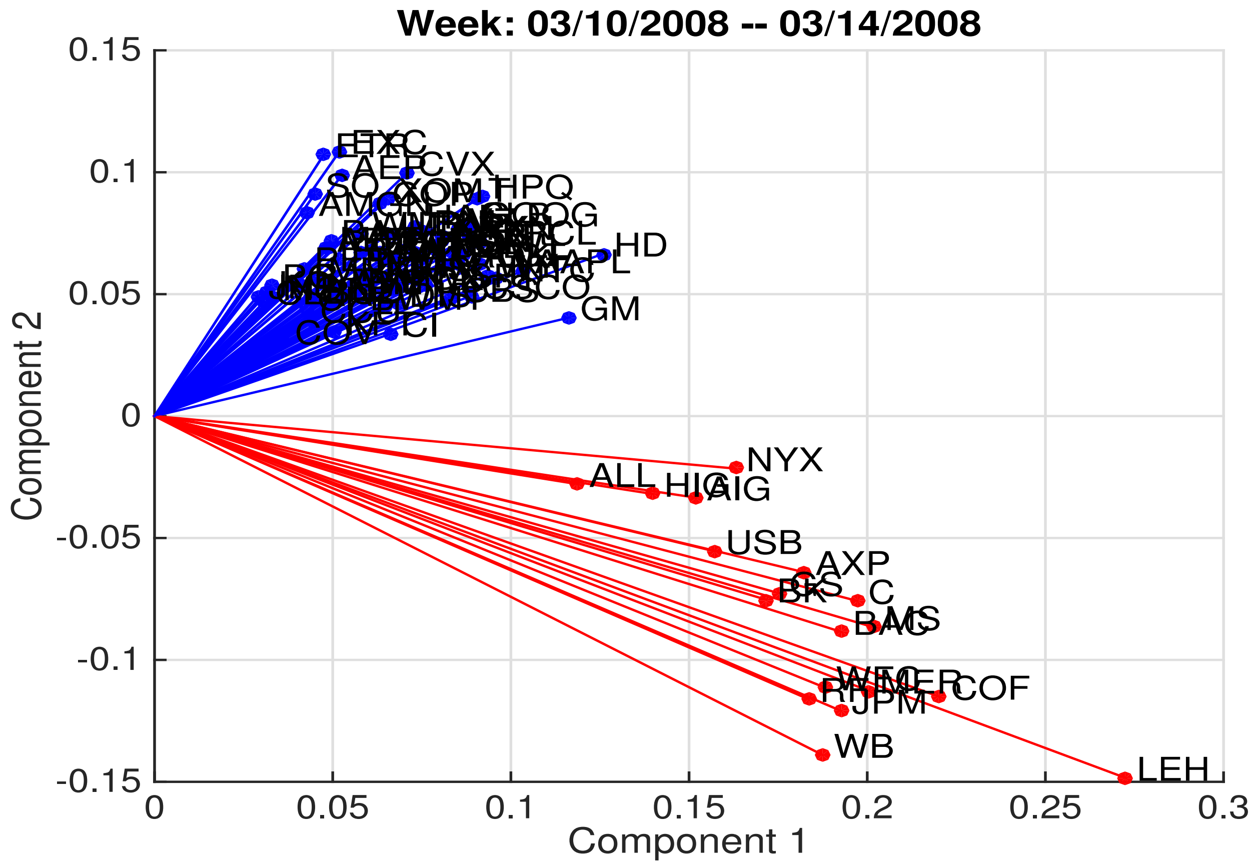
- We collect intraday returns of S&P 100 constituents over 2003 - 2012 periods from TAQ database.
- There are in total 158 different symbols as the constituents may have changed over time.
- We select the most liquid exchange for each ticker on each day.
- These stocks have superior liquidity, avoiding the issue of Microstructure noise and asynchronous trading.
- Data are sampled at 1-min frequency, and grouped by weeks.
- We remove 10 least liquid stocks.


3
2
1
0
-1
-2
-3
-4
-5
-6
-7
2003
2005
2007
2009
2011
2013

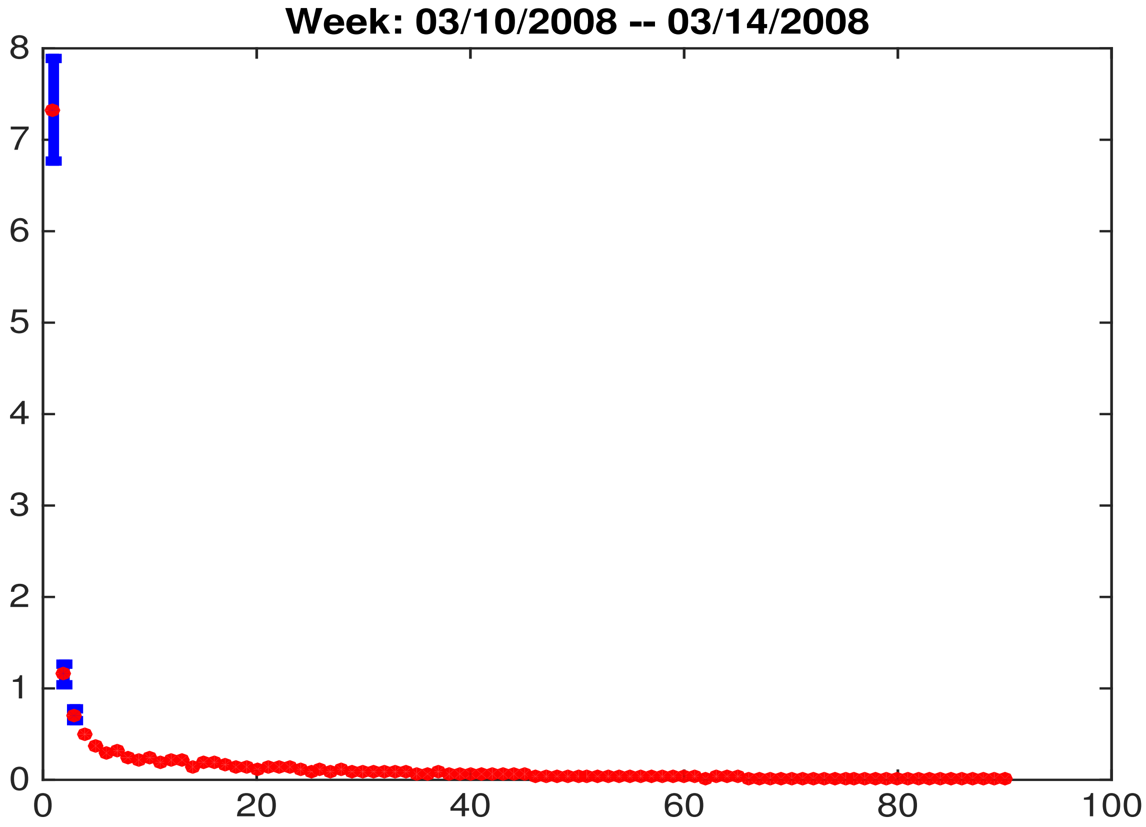

75% - 100%
25% - 75%
0% - 25%
Sampling Frequency
0.1
1
5
15
60
300
2003
2005
2007
2009
2011
2013


1st
2nd
3rd
≥ 4 Average
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2003
2005
2007
2009
2011
2013


PC1
PC2
PC3
3
2
1
0
-1
-2
-3
-4
-5
2003
2005
2007
2009
2011
2013



Week: 03/10/2008 -- 03/14/2008


Week: 03/10/2008 -- 03/14/2008
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100