The Five Trolls under the Bridge: Principal Component Analysis with Asynchronous and Noisy High Frequency Data

Dachuan Chen[1] Per A Mykland[2] Lan Zhang[3]

[1]University of Illinois at Chicago, and Nankai University

[2]University of Chicago

[3]University of Illinois at Chicago

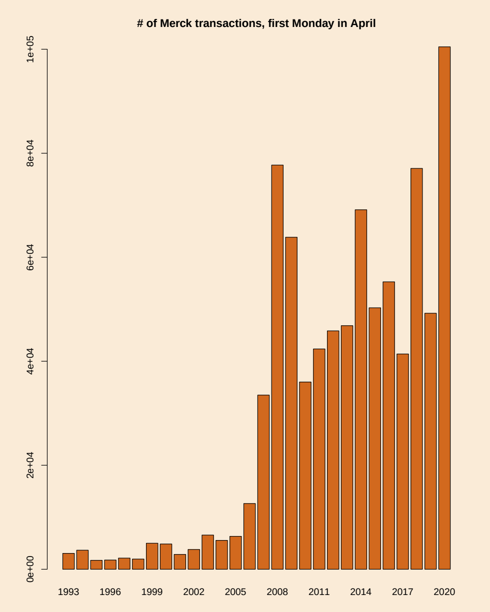
Oslo, September 2022

Outline

High Frequency Data

- Financial prices, volumes, number of trades, order time
- Intra-day:
  - transactions tick-by-tick, from TAQ, Refinitiv (ex Thompson-Reuters), CME
  - quotes - bid, ask - same sources
  - limit order books, harder to get but more information
  - stocks, bonds, futures, currencies, ...
- HF data can also be found in internet data, neuroscience, survival analysis, geoscience, climate recordings, wind measurements, turbulence, fish, ...

Evolution of Data Size per Day



Intraday Trading: almost time continuous

	Time	Size	Price
	9:00:05.897	100	601.740
	9:00:11.257	100	601.700
	9:00:11.340	100	601.730
	9:00:12.190	100	601.700
	9:00:12.393	500	601.700
Apple	9:00:12.807	200	601.700
April 2, 2012	9:00:13.060	100	601.700
	9:00:13.460	100	601.650
	9:00:14.240	100	601.700
Number of Trades	9:00:14.913	100	601.700
102,986	9:00:14.913	200	601.700
	9:00:15.310	100	601.700
	9:00:18.380	100	601.530
	⋮	⋮	⋮

Intraday Trading: almost time continuous

	Time	Size	Price
Apple April 2, 2020 Number of Trades 376,731	10:00:00.000985678	65	239.390
	10:00:00.010742509	2	239.390
	10:00:00.010744971	100	239.390
	10:00:00.010748759	6	239.400
	10:00:00.010752774	100	239.450
	10:00:00.010887597	1	239.390
	10:00:00.011109135	34	239.450
	10:00:00.019740536	2	239.423
	10:00:00.042692078	9	239.440
	10:00:00.044256462	3	239.390
	10:00:00.047250042	20	239.390
	10:00:00.064590362	100	239.390
	10:00:00.073841728	20	239.430
	⋮	⋮	⋮

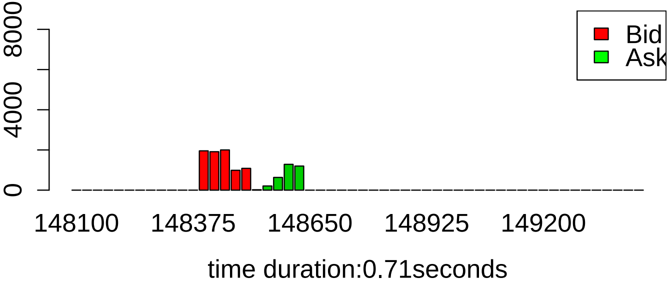
Observation times are : (1) down to nano-seconds per trade,

(2) non-equidistant, (3) could be endogenous.

High Dimension

- Equity Cross Section: over 4000 stocks are traded at NYSE. Each day, NYSE has about 1 billion shares being traded
- Options: contracts with varying excise prices, contracts with varying maturity times (additional dimension for many stocks)
- Order Book: varying depth (additional dimension for every stock)

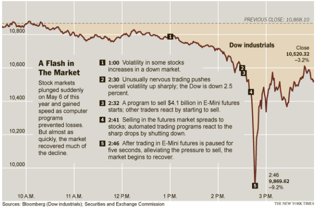
Snapshot of Limit Order Book for E-mini S&P 500



Snapshot from May 1, 2007: horizontal line shows the five best bid prices (red) and five best ask prices (green), while vertical line shows the volume of each quote.

Price movement almost path-continuous, but . . .

Figure: Intraday Sudden Price Movement



(a) 2010 Crash of 2:45pm

(b) 2013 Twitter Crash

Left: (a) On May 6 2010: All major US stock indices plunged and rebounded within about 30 minutes. Dow Jones Industrial Average plunged 998.5 points (about 9%), most within minutes. Graph source: NYT. Right: (b) On Tuesday April 23, 2013: Dow quickly plunged 140 points (about 1%) after a false tweet. The S&P 500 lost $121 billion of its value within minutes. Graph source: CNN money

Data Features & Challenges Including Three Trolls

- Large amount of data (up to about a million observations a day for single security)
- High frequency: observation interval could be less than milliseconds.
- Price movement almost path-continuous, but rare extreme events (jumps) could occur.
- Microstructure noise is more pronounced in high frequency data.
- Random observation times
  - Unequal time interval (not in the setting of time series)
  - Time stamps could be inaccurate when data are from different sources/exchanges
  - Trade times are often endogenous.
- Cross-sectional data (multiple securities): asynchronicity in trade time or quote update time
- Volume could be intentionally split.
- Edge effect in estimators

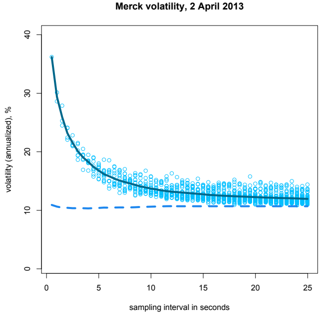
How to Handle the High Frequency Data?

- Direct modeling to take microstructure noise into account
- Hidden semimartingale model
  - observed log stock price: $Y_{t_i} = X_{t_i} + \epsilon_i$,
  - X_t is latent log price, semimartingale, say, Ito process

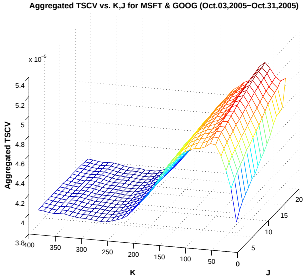
$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s,$$

    B_t is Brownian motion; μ_t and σ_t can be random processes
  - ϵ_i is stationary or iid, or similar
  - there may also be jumps, but not in this version of the paper
- log-price process $X_t = (X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(d)})$ of d stocks
  - spot covariance process: $c_t = \sigma_t \sigma_t^T$
  - if X_t is continuous: quadratic variation $[X, X]_t = \int_0^t c_s ds$.

Correct Bias from Noise and Asynchronicity



(c) Daily volatility



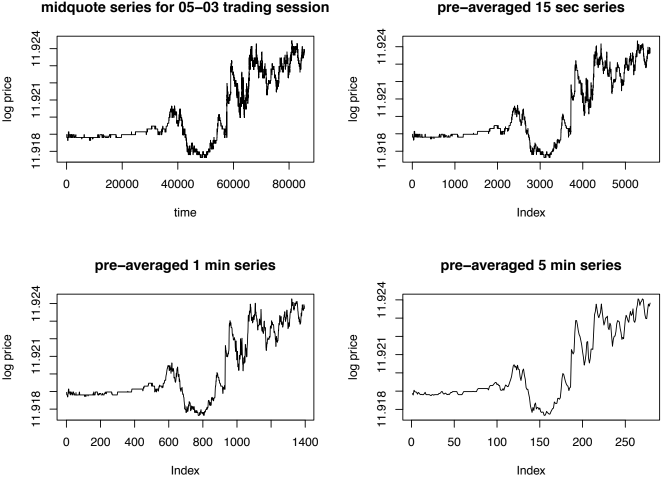
(d) Daily correlation

All estimates are computed from intraday data. Left: (c) Ignoring the microstructure noise over-estimates price volatility. Bias is even more pronounced if one uses ultra high frequency data; Right: (d) Ignoring the noise and/or interpolation under-estimates

How to Handle the High Frequency Data?

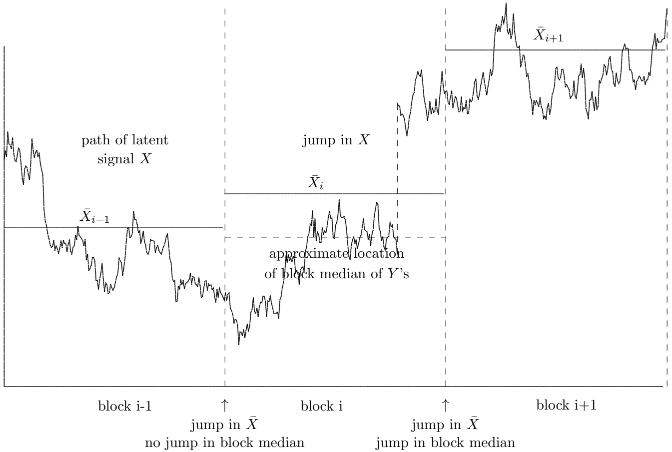
- Data reduction:
  - Smooth (pre-average or pre-medianize) the tick-by-tick data
    - Reduces the size of noise, but complicates the model
    - Induces bias when data are from irregular or asynchronous times (in-depth critique later, if time permits)
    - Pre-averaging pulverizes jumps
  - Estimate the volatility matrix:
    - Pre-averaging best used as an ingredient in estimation, but not off the shelf
    - In this paper, we use the Smoothed TSRV (MZC (2019), “The algebra of two scales estimation"; more later)

Before and After Data Smoothing



By smoothing, you can get ride of most of the noise, but also lose part of the true dispersion of the data. Our methodology quantifies the loss and recovers the true dispersion.

Jump Pulverization: Block Median vs. Block Average



What if the dimension is also high?

- Recall:
  - log-price process $X_t = (X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(d)})$ of d stocks
  - spot covariance process: $c_t = \sigma_t \sigma_t^\intercal$ is $d \times d$
- Further data reduction:
  - Principal Component Analysis (PCA): find the eigenvalues and eigenvectors of $\hat{c}_t$:
    $\lambda_t^{(j)}$, $1 \leq j \leq q$: eigenvalues of c_t in non-ascending order
    $\gamma_t^{(j)}$, $1 \leq j \leq q$: corresponding eigenvectors
  - Factor Analysis (FA):
    Factor model with time-varying factor loadings
    $dX_t = \mathbf{B}_t d\mathbf{F}_t + dZ_t$
    High dimension of X_t and low dimension of $\mathbf{F}_t$:
    $(\gamma_{t-}^{(j)})^\intercal dX_t \approx (\lambda_{t-}^{(j)})^{\frac{1}{2}} dF_t^{(j)}$: j^{th} factor or PC
  - Regression on observed or estimated factors

Standard Chain from Data to Factors: Five Trolls

data → spot covariance matrix c → PCA → factor analysis

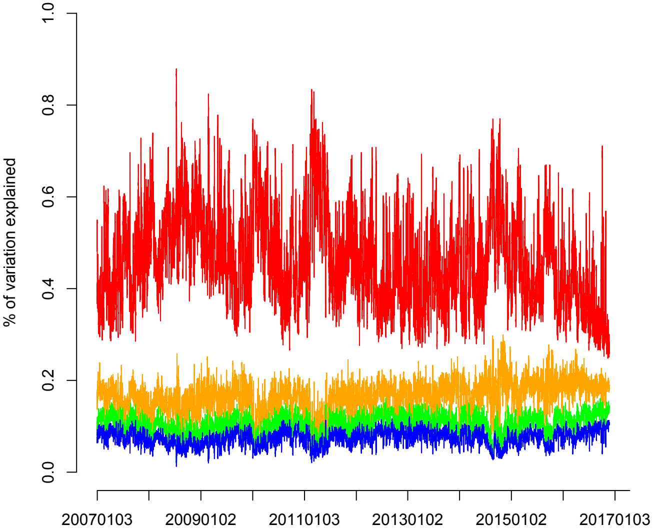
- PCA → factor analysis: Requires high dimension. PCs with large eighenvalues can be taken as factors. (Chamberlain and Rothschild (1983), Connor and Korajczyk (1986), Stock and Watson (1998, 2002), Bai and Ng (2002), Fan, Liao and Mincheva (2013) (POET), Pelger (2017), many others).
- covariance matrix c → PCA: Pearson (1901), Hotelling (1933), ... , Aït-Sahalia and Xiu (2018) (AX).
- data → spot covariance matrix c: The tricky part for high frequency data: noise, asynchronicity, edge effects, ... An imprecise estimator can mess up your PCs and factors.

In this paper

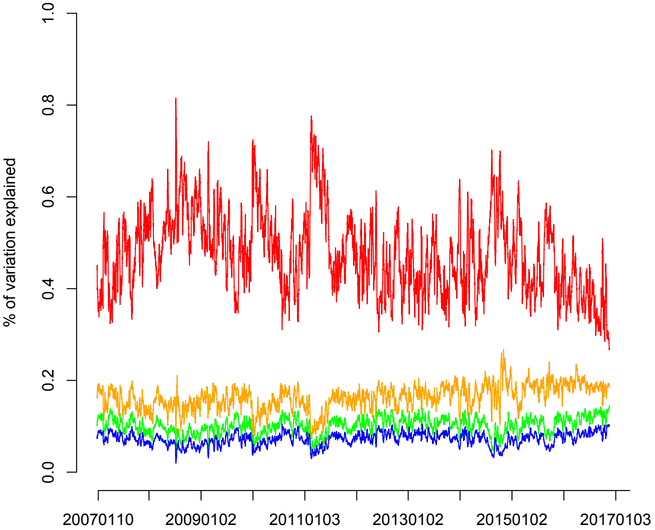
- data $\rightarrow$ estimated spot covariance matrix $\hat{c}_t$: Updated semi-frequently, every ΔT seconds. In empirical data: $\Delta T =$2500 seconds. Nine such periods in one trading day from 9:45 am to 4 pm New York time.
  Method: $\hat{c}_t =$ S-TSRV matrix. Uses all data in time interval. Takes account of noise, asynchronicity, and, in particular, edge effects. (More about this later.)
- covariance matrix $c \rightarrow$ PCA: Follow AX; mostly assume that big eigenvalues are simple.
- PCA $\rightarrow$ factor analysis: Follow POET, but the high frequency data makes the problem simpler (more about this later)

First some graphs, then back to theory

Percentage of the Total Variation Explained by Principal Components (1 day rolling mean)



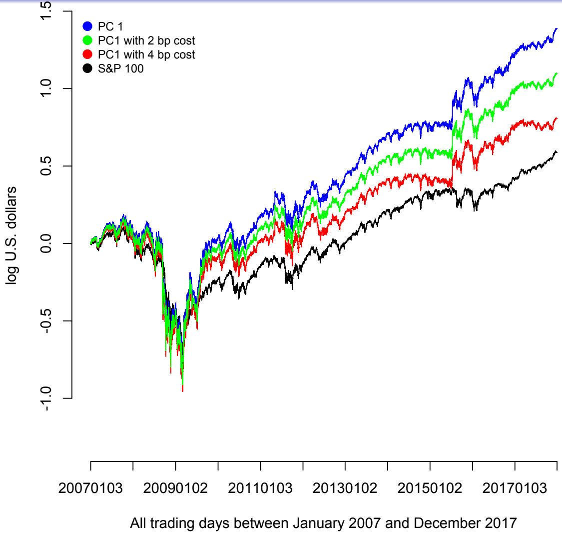
Percentage of the Total Variation Explained by Principal Components (1 week rolling mean)



Why rolling Mean?

- 2500 seconds based estimators too variable relative to bias
- When trading: rolling mean $\Longrightarrow$ less trading cost:
  - 9 period (one day) rolling mean means that only about $(1/9)^{\text{th}}$ of portfolio is updated every 2500 seconds
  - 45 period (one week) rolling mean means that only about $(1/45)^{\text{th}}$ of portfolio is updated every 2500 seconds
- Both phenomena documented by plots (a few slides later)
- Overnight position uses same weights as period 1 next day (based on data from trading periods ending at 4 pm on the preceeding day)

Unsupervised Learning in Intraday Data: PC1 Portfolio vs. S&P 100: 1 Day Rolling Mean Eigenvector



Eigenvalues and -vectors

- $\hat{c}_{T_i}$ Estimated covariance matrix of log returns of prices of 70 stocks from the S&P 100.
- One every 2500 seconds.
- 9 x 2769 trading days = 24921 periods of 2500 sec: T_i has $i = 1, \cdots, 24921$
- $\hat{c}_{T_i}$ is TSRV based on 5 second averages. 12,460,500 periods of 5 sec
- $\hat{\lambda}_{T_i}$: largest eigenvalue of $\hat{c}_{T_i}$
- $\hat{\gamma}_{T_i}$ corresponding eigenvector
- Similar for higher order eigenvalues, -vectors
- If one could trade $X_t = \log S_t$, the $\mathrm{PC_k}$ portfolio would have P/L = $\sum_i^T (\hat{\gamma}_{T_{i-1}})^\intercal (X_{T_i} - X_{T_{i-1}})$
- But this is not possible

The PC Portfolios

- The $\mathrm{PC_k}$ portfolio (log scale): log P/L is

*
$$\log w_T = \log w_0 + \sum_{i=1}^{T} \log\left(1 + \left(\hat{\gamma}_{T_{i-1}}\right)^{\intercal} r_{T_i}\right) \text{ where}$$

- $\hat{\gamma}_{T_{i-1}}$ is $\mathrm{k^{th}}$ eigenvector
- r_{T_i} is a vector with $\mathrm{j^{th}}$ element $r_{T_i}^{(j)} = (S_{T_i}^{(j)} - S_{T_{i-1}}^{(j)})/S_{T_{i-1}}^{(j)}$
- $r_{T_i}^{(j)}$ are the returns on stocks $S^{(j)}$, $j = 1, \cdots, d$
- Trading algorithm invests fraction

*
$$\delta_{i-1} = \sum_{j=1}^{d} \hat{\gamma}_{T_{i-1}}^{(j)}$$

  of $w_{T_{i-1}}$ in stocks in the period from T_{i-1} to T_i
- Fraction $1 - \delta_{i-1}$ kept in cash
- Holds $w_{T_{i-1}} \hat{\gamma}_{T_{i-1}}^{(j)} / S_{T_{i-1}}^{(j)}$ units of stock $S^{(j)}$ in this time period
- Interest rates on cash taken to be zero (nearly the case)
- Algorithm is implementable (but need to add trading cost)

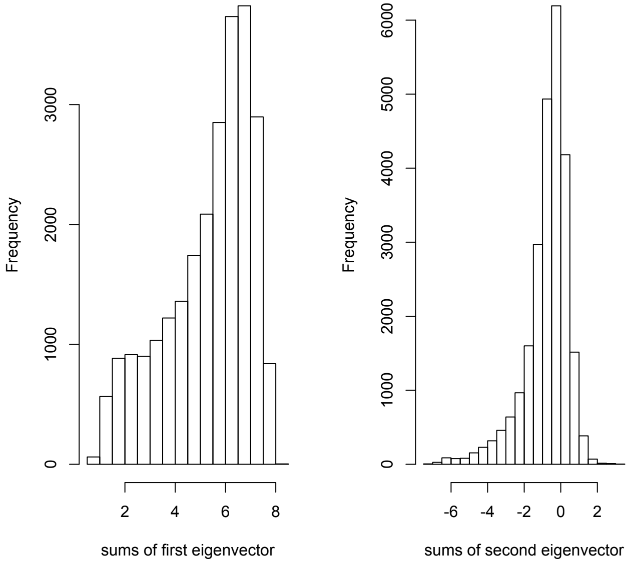
Log vs. non-log Scale

- Apparent paradox:
  - PCA is carried out on log returns of prices: $X_t^{(j)} = \log S_t^{(j)}$
  - Trading is carried out with returns $r_{T_i}^{(j)} = (S_{T_i}^{(j)} - S_{T_{i-1}}^{(j)})/S_{T_{i-1}}^{(j)}$
- Necessity:
  - $X_t^{(j)}$ are approximately additive, suitable for PCA
  - For trading: cannot add log prices, need $r_{T_i}^{(j)}$
- Validity under continuous paths (no jumps):
  - $r_{T_i}^{(j)} = X_{T_i} - X_{T_{i-1}} +$ Itô correction term
  - PCA valid (in a medium term sense) despite correction term due to Girsanov's Theorem
  - The correction term usually improves performance of trading algorithm
- Jumps:
  - May not be desirable to include a large jump that has already occurred in the near past
  - Issue of infinitely many small jumps: unresolved

How similar can PC1 be to the Value Weighted Index?

- A priori: if only one factor driving the market: there is a covariance matrix argument embedded in the argument for holding the VW index such as S&P 100 (Markowitz (1952, 1959), Sharpe (1964), Lintner (1965), Black (1972))
- However:
  - Few people believe that there is only one factor driving the market
  - Other problem in practice: To resemble an index, portfolio needs to be self financing
  - Equivalent statement: Need to standardize first eigenvector to sum to one: $\delta_{i-1} \equiv 1$
  - Referee 2 thought it could not be done: (1) the sum could be close to zero, (2) there could be lots of negative portfolio weights, and (3) the "PC weights do not aggregate to 1 (their 2-norm is)"
- And yet, it can be done, as we shall see presently
- A suggestion that PCA may provide a suitable index when VW argument is not available? (Commodities, etc)

PC1 is unlike other PCs: Sum of eigenvector >> 0



What about negative portfolio weights?

- Negative fraction of the first eigenvector $\hat{\gamma}^{(1)}_{T_{i-1}}$:

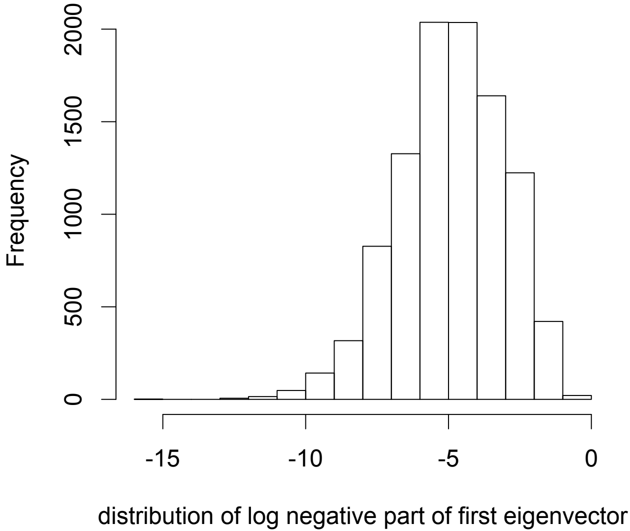
$$n_i = \sum_{j=1}^{d} \left(\hat{\gamma}^{(1,j)}_{T_{i-1}}\right)^{-} / \sum_{j=1}^{d} \hat{\gamma}^{(1,j)}_{T_{i-1}} \text{ where } x^{-} = \max(-x, 0),$$

- For different amounts of averaging:

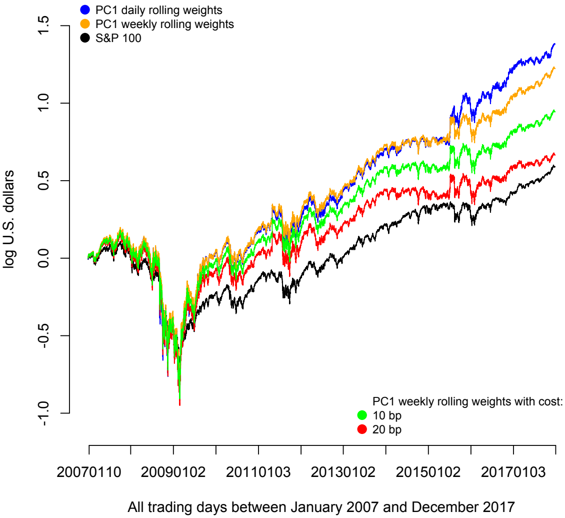
	n_i over 11 years 2007-2017		
first eigenvector	mean	95th %ile	max
daily rolling	0.011	0.067	0.538
weekly rolling	0.0012	0.0053	0.0778
not rolling			1480.94

- Histogram of $\log(n_i)$: next page
- If needed, build limits on the negative part into the portfolio selection, or even calculation of the eigenvector

Log Negative Part of 1 Day Rolling Eigenvector



Allowing for higher Trading Cost in PC1: 5 Days Rolling Mean Eigenvector

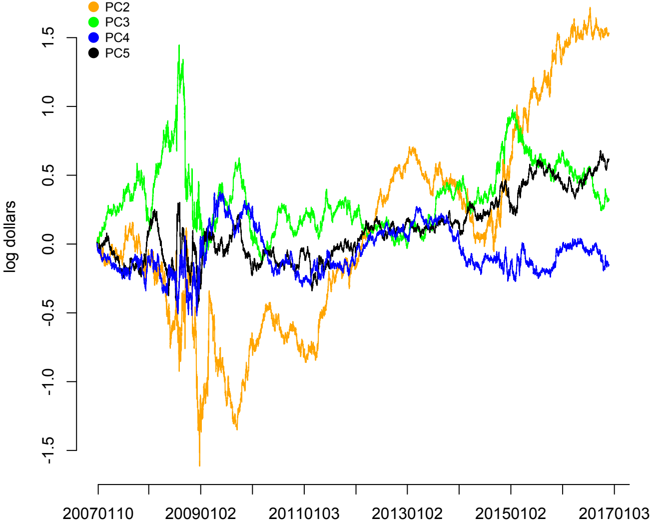


Basic Financial Measures for PC1

	S&P 100	PC1 daily rolling	PC1 weekly rolling
annual returns	5.3%	12.5%	11.1%
cumulative returns	58.8%	138.0%	122.2%
annual volatility	15.6%	24.3%	23.2%
Sharpe ratio	34.0%	51.4%	47.8%
Sortino ratio	43.5%	72.0%	67.7%
daily turnover	0	58.3%	11.2%
maximum drawdown	56.2%	65.3%	65.5%
alpha	0	0	0
beta	1	1.44	1.40

Annual returns. Volatilities were computed using the S-TSRV, and similarly for the semi-variances that go into the Sortino ratio. For the computation of alpha and beta, S&P 100 (OEF) is used as market proxy, and monthly returns have been used in the regression. For all the three series, the maximum drawdown occurred at market close on 5 March, 2009.

Higher Order PC Portfolios



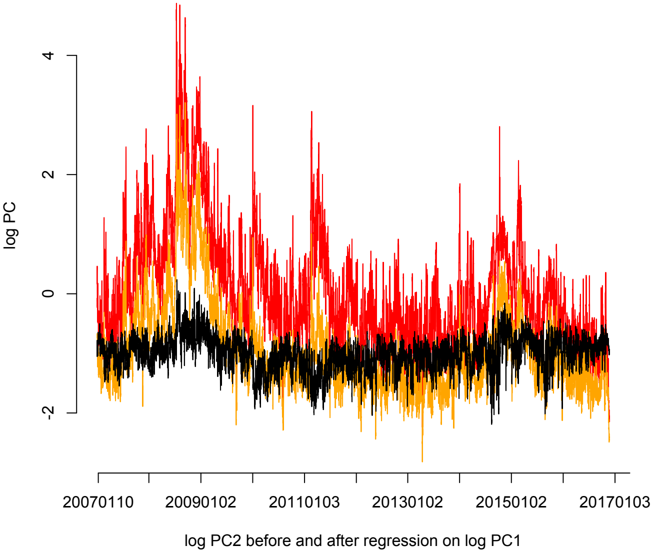
Higher Order PC Portfolios

- Higher order PCs are different from PC1: δ_{i-1} straddle zero
- This is natural: if PC1 is close to the “market", then the higher order PCs should be close to market-neutral
- How to standardize the eigenvectors?
- Some form of constraint on leverage?
- Preceding plot uses eigenvectors with norm one
- Sign: for higher order eigenvectors: “continuity method":

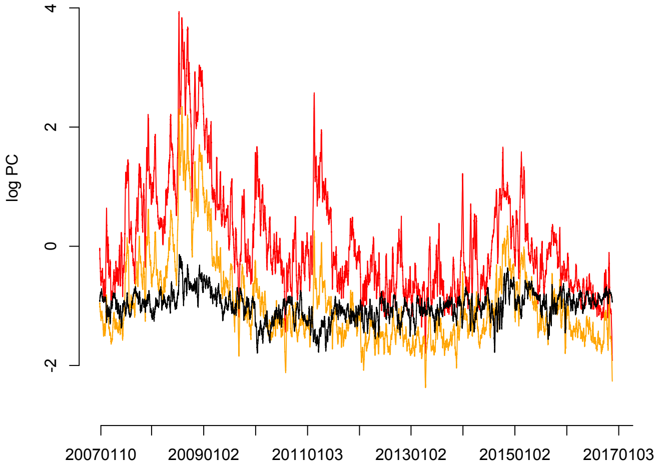
$$\text{assign sign}(\hat{\gamma}_{T_i}^{(h)}) \text{ so that sign}\{(\hat{\gamma}_{T_i}^{(h)})^\intercal \hat{\gamma}_{T_{i-1}}^{(h)}\} \geq 0.$$

- Relationship to Fama-French
- Other additional data: volume, text (Tracy Ke and others)
- Volatility of drift (close to observed AVAR approach)

Relationship between Eigenvalue One and Two



Relationship between Eigenvalue One and Two: 45 period (one week) average



And now for some theory

- From Covariance Matrix $\hat{c}_t$ to PCA
- From PCA to realized POET
- From Data to Covariance Matrix $\hat{c}_t$

From Covariance Matrix $\hat{c}_t$ to PCA

- $\lambda_t^{(j)}$, $1 \leq j \leq q$: eigenvalues of c_t in non-ascending order and $\gamma_t^{(j)}$, $1 \leq j \leq q$: corresponding eigenvectors,
- $\lambda_t^{(j)}, \gamma_t^{(j)}$ on form $F(c_t)$, where F are analytic functions
- Spot estimators $\hat{\lambda}_t^{(j)}, \hat{\gamma}_t^{(j)}$ on form $F(\hat{c}_t)$ (AX)
- Integrated quantities by accumulation of spot estimators
- Analysis different from AX because of more complicated $\hat{c}_t$ (three trolls, epecially edge effect)
- Correction term similar to AX, but also containing effect of noise
- SRC, Corrected integrated quantities have consistency, asymptotic normality:

$$a_n^{-1}\left(\tilde{V}\left(\Delta T_n, X; F\right) - \int_0^{\mathcal{T}} F\left(c_s\right) ds\right) \xrightarrow{\mathcal{L}} W_{\mathcal{T}},$$

where $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-3/2}\Delta T_n \to \infty$ as $n \to \infty$.

Illustration of complexity

Table: Error Size Comparison under Different Choices of ΔT_n and a_n

	Types of Error				
	R^{Discrete}	$R^{\text{Spot-V}}$	$R^{\text{Spot-B}}$	$E\left(R^{\text{Spot-B}}\right) - \varphi^{\text{Bias}}_{\Delta T_n}$	$R^{\text{Expansion}}$
$\Delta T_n \to 0$ and $\inf_n a_n^{-1} \Delta T_n > 0$	$O_p(\Delta T_n)$	$O_p(\Delta T_n)$	$O_p(\Delta T_n)$	$o_p(\Delta T_n)$	$O_p\left(\Delta T_n^2\right)$
$a_n^{-1} \Delta T_n \to 0$ and $a_n^{-3/2} \Delta T_n \to \infty$	$O_p(\Delta T_n)$	$O_p(a_n)$	$O_p\left(a_n^2 \Delta T_n^{-1}\right)$	$O_p\left(a_n^4 \Delta T_n^{-2}\right) = o_p(a_n)$	$O_p\left(a_n^3 \Delta T_n^{-1}\right)$
$\sup_n a_n^{-3/2} \Delta T_n < \infty$ and $a_n^{-2} \Delta T_n \to \infty$	$O_p(\Delta T_n)$	$O_p(a_n)$	$O_p\left(a_n^2 \Delta T_n^{-1}\right)$	$O_p\left(a_n^4 \Delta T_n^{-2}\right)$	$O_p\left(a_n^3 \Delta T_n^{-1}\right)$

R^{Discrete}: Discretization error: $R^{\text{Spot-V}}$ and $R^{\text{Spot-B}}$: Martingale term and bias term . $R^{\text{Expansion}}$: Aggregated remainder term. $E\left(R^{\text{Spot-B}}\right) - \varphi^{\text{Bias}}_{\Delta T_n}$: the bias term contributed by the edge effect in covariance estimator. $\varphi^{\text{Bias}}_{\Delta T_n}$ is due to irregular sampling and microstructure noise.

From PCA to realized POET

- Factor Model with time-varying factor loadings:

$$\underbrace{dX_t}_{d\times 1} = \underbrace{\mathbf{B}_t}_{d\times q}\underbrace{d\mathbf{F}_t}_{q\times 1} + \underbrace{dZ_t}_{d\times 1} \text{ with } \langle \mathbf{F}, Z\rangle_t \equiv 0 \quad (1)$$

- $c_t = \mathbf{B}_t c_t^{\mathbf{F}} \mathbf{B}_t^{\intercal} + \mathbf{s}_t$ where $c_t^{\mathbf{F}} = \langle \mathbf{F}, \mathbf{F}\rangle_t'$ and $\mathbf{s}_t = \langle Z, Z\rangle_t'$
- Normalization: $c_t^{\mathbf{F}} = \mathbb{I}_q$ and $\mathbf{B}_t^{\intercal}\mathbf{B}_t$ is diagonal.
- Easiest interpretation: Can choose wlog

$$\underbrace{\mathbf{B}_t}_{d\times q} = \underbrace{\mathbf{G}_t}_{d\times q}\underbrace{\mathbf{L}_t^{1/2}}_{q\times q}$$

  with $\mathbf{L}_t$ is diagonal, $\mathbf{G}_t^T\mathbf{G}_t = \mathbb{I}_q$, so $\mathbf{B}_t^T\mathbf{B}_t = \mathbf{L}_t$
- Factors with scale:
  $\mathbf{L}_t^{1/2} d\mathbf{F}_t = ((L_t^{(11)})^{1/2} dF_t^{(1)}, \cdots, (L_t^{(qq)})^{1/2} dF_t^{(q)})$,
  approximately replicated by trading strategy
- $c_t = \mathbf{B}_t\mathbf{B}_t^{\intercal} + \mathbf{s}_t$
- Two approaches:
  - $\mathbf{s}_t$ is block diagonal (AX)
  - $\mathbf{s}_t$ is sparse (POET)

Pervasiveness and PCA → factor analysis

- Recall c_t: eigenvalues $\{\lambda_t^{(j)}\}_{1\leq j\leq q}$ (in non-ascending order) and corresponding eigenvectors $\{\gamma_t^{(j)}\}_{1\leq j\leq q}$
- Let $\mathbf{B}_t\mathbf{B}_t^\intercal$ have eigenvalues $\{\mathfrak{l}_t^{(j)}\}_{1\leq j\leq q}$ (in non-ascending order) and corresponding eigenvectors $\{\mathfrak{g}_t^{(j)}\}_{1\leq j\leq q}$
- $\mathbf{L}_t = \mathrm{diag}(\mathfrak{l}_t^{(1)}\cdots\mathfrak{l}_t^{(q)})$ and $\mathbf{G}_t = (\mathfrak{g}_t^{(1)}\cdots\mathfrak{g}_t^{(q)})$
- Assume, for all t, that all eigenvalues of the $q\times q$ matrix $d^{-1}\mathbf{B}_t^\intercal\mathbf{B}_t = d^{-1}\mathbf{L}_t$ are distinct and bounded away from 0 and ∞ as $d\to\infty$. (Pervasiveness.)
  Then
  - for $1\leq j\leq q$:
    - $|\lambda_t^{(j)} - \mathfrak{l}_t^{(j)}| \leq ||\mathbf{s}_t||$, and
    - $||\gamma_t^{(j)} - \mathfrak{g}_t^{(j)}|| = O\left(d^{-1}\,||\mathbf{s}_t||\right)$
  - and for $j>q$: $|\lambda_t^{(j)}| \leq \|\mathbf{s}_t\|$
- Proof similar to Fan *et al.* (2013). Weyl's theorem, etc

POET becomes simpler in the high frequency setup

- $c_t = \mathbf{B}_t \mathbf{B}_t^\intercal + \mathbf{s}_t$, where $\mathbf{s}_t = \langle Z, Z \rangle'_t$
- Constrained least squares (CLS): Go back to original data matrix and find residuals for given $\mathbf{B}_t \rightarrow$ residual sum of squares (RSS)
- In this case: no need, $RSS = \mathrm{trace}(\mathbf{s}_t)$ (or its estimate)
- CLS gives: $\mathbf{B}_t = \arg\min_{\mathbf{B}_t \in \mathbb{R}^{d \times q}} \mathrm{trace}(\mathbf{s}_t)$
- In other words: $\mathbf{B}_t = \Gamma_t \Lambda_t^{1/2}$ (or its estimators), where $\Lambda_t =$diag $(\lambda_t^{(1)}, \lambda_t^{(2)}, \ldots, \lambda_t^{(q)})$ and $\Gamma_t = (\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(q)})$
- $\lambda_t^{(j)}$, $\gamma_t^{(j)}$ from spectral decomposition of c_t
- $L_t = \Lambda_t$ and $G_t = \Gamma_t$
- Estimation of q: eyeball, or penalized criterion function
- Sparsity enters when estimating $\mathbf{s}_t$, and possibly modified estimate of $\mathbf{c}_t$
- Consistency, convergence rates (see paper)

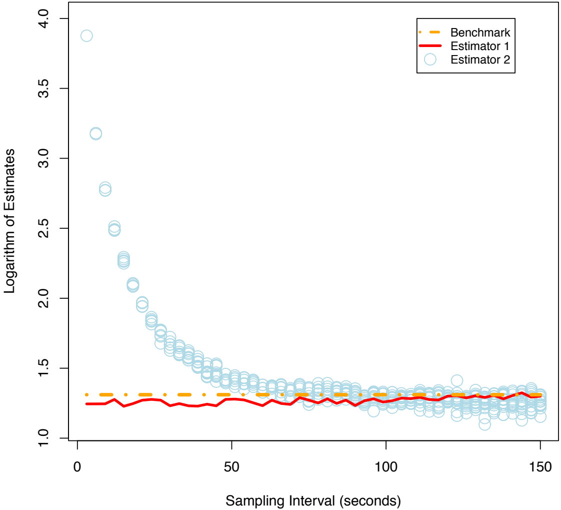
The Smoothed TSRV (S-TSRV)

- Need $\hat{c}_t$
- Synchronous grid $\{0 = \tau_{n,0} < \tau_{n,1} < \cdots < \tau_{n,N} = \mathcal{T}\}$
- $\bar{Y}_i^{(r)}$ = prevaveraged price in each interval $(\tau_{n,i-1}, \tau_{n,i}]$
- Pair (J, K) of scales, $J << K$; set $b = K + J$
- Tapered K-scale variation: $K\left[\widetilde{\bar{Y}^{(r)}, \bar{Y}^{(s)}}\right]_t^{(K)} =$
$\left(\frac{1}{2}\sum_{i=1}^{b-K} + \sum_{i=b-K+1}^{N^*(t)-b} + \frac{1}{2}\sum_{i=N^*(t)-b+1}^{N^*(t)-K}\right)\left(\bar{Y}_{i+K}^{(r)} - \bar{Y}_i^{(r)}\right)\left(\bar{Y}_{i+K}^{(s)} - \bar{Y}_i^{(s)}\right)$
where $N^*(t) = \max\{1 \le i \le N : \tau_{n,i} \le t\}$
- Two-scales construction

$$\widehat{\langle X^{(r)}, X^{(s)} \rangle}_t = \frac{1}{(1 - b/N)(K - J)}\left\{K\left[\widetilde{\bar{Y}^{(r)}, \bar{Y}^{(s)}}\right]_t^{(K)} - J\left[\widetilde{\bar{Y}^{(r)}, \bar{Y}^{(s)}}\right]_t^{(J)}\right\}.$$

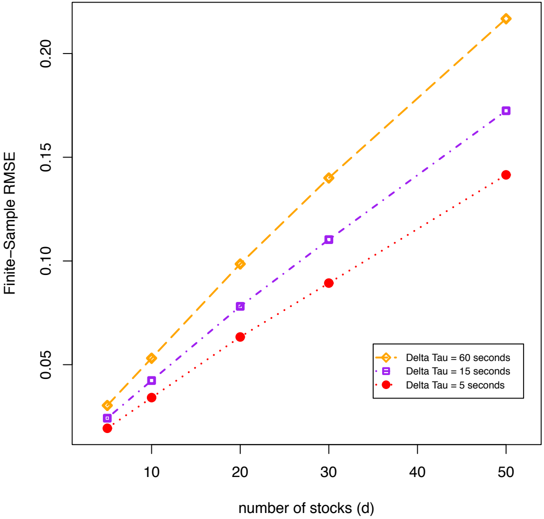
- Why?

The importance of getting $\hat{c}_t$ right



Signature Plot for the Estimates of Integrated Largest Eigenvalue on Logarithmic Scale (simulated data)

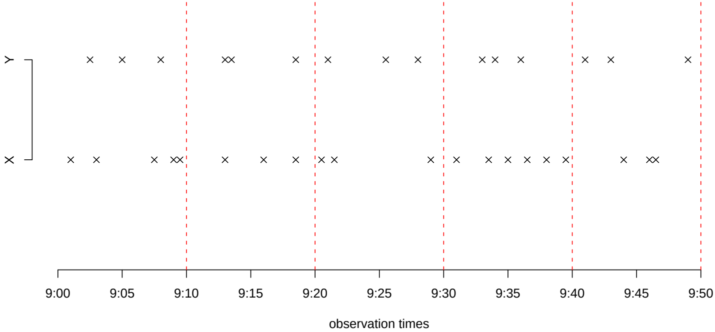
Smaller Intervals are Better under S-TSRV: RMSE of Integrated Largest Eigenvalue Estimates



Some more Theory, about Covariance Estimation

- How pre-averaging helps, but also creates problems
- How a two-scales construction on top of pre-averaging yields a well behaved estimator (S-TSRV)

Pre-Averaging as a Potential Synchronization Device



Simple mean (or weighted mean) of each process in each time interval. For the stock data in this paper, we use 5 second time intervals.

Pre-Averaging can Reduce the Impact of Noise

- Observed log stock price: $Y_{t_j} = X_{t_j} + \epsilon_j, \; t_j \in [0, T]$
- $dX_t = \mu_t dt + \sigma_t dB_t$
- Block average for block i, $[\tau_i, \tau_{i+1})$:

$$\bar{Y}_i = \frac{1}{M_i} \sum_{\tau_i \le t_j < \tau_{i+1}} Y_{t_j}$$

- Reduction of size of noise through block averages:

$$\begin{aligned} \bar{Y}_i &= \bar{X}_i + \bar{\epsilon}_i \\ &= \bar{X}_i + O_p(M_i^{-1/2}) \\ &\overset{?}{\approx} X_{\tau_i} + O_p(M_i^{-1/2}) \end{aligned}$$

- A form of data cleaning
- Analogy to trading: splitting a large order
- However: $\overset{?}{\approx}$ is not innocuous when times are irregular

Characterizing Irregular Times Inside a Single Block

- Let t_{j_0}: first $t_j \in [\tau_{i-1}, \tau_i)$ and set

$$I_i = \begin{cases} \frac{M_i - j}{M_i} \text{ with probability } \frac{\Delta t_{j_0+j}}{\Delta \tau_i} \\ 1 \text{ with probability } \frac{t_{j_0} - \tau_{i-1}}{\Delta \tau_i} \\ 0 \text{ with probability } \frac{\tau_i - t_{j_0+M_i-1}}{\Delta \tau_i} \end{cases}$$

  where $j = 1, 2, ..., M_i - 1$ and $\Delta t_{j_0+j} = t_{j_0+j} - t_{j_0+j-1}$

- Preaveraged $RV = \sum_i (\Delta \bar{X}_i)^2$ will depend on:

$$E(I_i) = \sum_{t_j \in (\tau_{i-1}, \tau_i]} \frac{M_i - j}{M_i} \frac{\Delta t_{j_0+j}}{\Delta \tau_i} + \frac{t_{j_0} - \tau_{i-1}}{\Delta \tau_i}$$

$$E(I_i^2) = \sum_{t_j \in (\tau_{i-1}, \tau_i]} \left(\frac{M_i - j}{M_i} \right)^2 \frac{\Delta t_{j_0+j}}{\Delta \tau_i} + \frac{t_{j_0} - \tau_{i-1}}{\Delta \tau_i}$$

- You have to use the exact times t_j to get $E(I_i)$ and $E(I_i^2)$.

Cases where Irregular Times are Innocuous

- For equidistant observations:

$$E(I_i) = \frac{1}{2} \text{ and } E(I_i^2) = \frac{1}{3}$$

- For times distributed by an inhomogenous Poisson process:

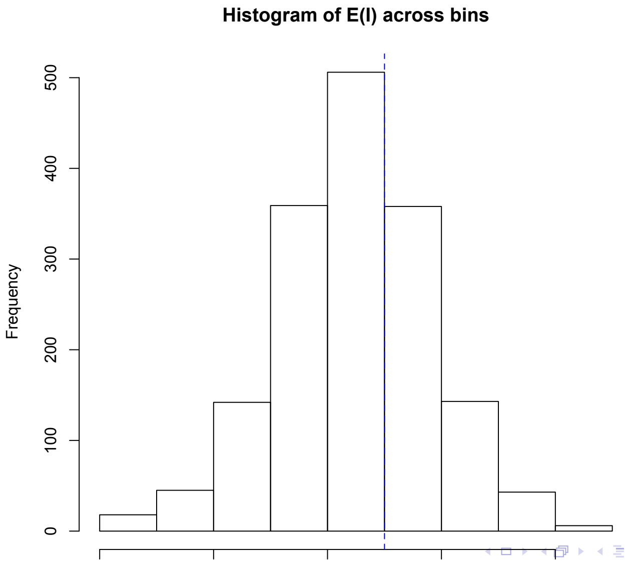
$$E(I_i) \approx \frac{1}{2} \text{ and } E(I_i^2) \approx \frac{1}{3} \qquad (2)$$

- For benign irregularity: observation times that are a fixed transformation of an equidistant grid:

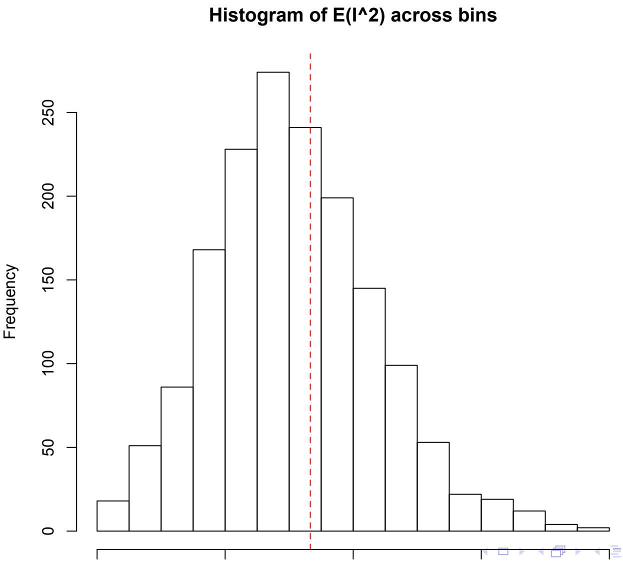
$$t_{n,j} = F(i/n) \text{ and } F \text{ is independent of } n$$

  - (1) is also true.
  - Benign irregularity is close to (and implies) contiguity to equidistant times.
- For general irregularity, however, you have to use the exact times t_j to get $E(I_i)$ and $E(I_i^2)$.

Histograms of Irregular Times



Histograms of Irregular Times



Sum of Squares under Irregular Times

- Suppose for simplicity: $\sigma_t^2 \approx \sigma_{\tau_{i-1}}^2$ (by contiguity, this is a smaller order problem) and that the τ_i are non-random
- Obtain

$$E[(\Delta \bar{X}_{i+1})^2 \mid \mathcal{F}_{\tau_{i-1}}] \approx \sigma_{\tau_{i-1}}^2 \Delta\tau_i (E(1-I_i)^2) + \sigma_{\tau_i}^2 \Delta\tau_{i+1} E((I_{i+1})^2).$$

- RV of preaveraged signal:

$$\sum_i (\Delta \bar{X}_{i+1})^2 \approx \sum_i \sigma_{\tau_{i-1}}^2 \Delta\tau_i E(2I_i^2 - 2I_i + 1)$$

good news: $\xrightarrow{p} \frac{2}{3} \int_0^T \sigma_t^2 dt$ under benign irregularity

bad news: $\xrightarrow{p}$??? for more general irregularity of times

- Pre-averaging CANNOT mitigate the effect of irregular times. – How to proceed next?

Estimated Moments of Irregular Times

- For $E(I_i)$ over the 1620 bins during the day of May 1, 2007, on S&P 500 E-mini

Min.	1*stQu.*	*Median*	*Mean*	3*rdQu.*	*Max.*
0.0000	0.3621	0.4511	0.4476	0.5364	0.8685

- For $E(I_i^2)$ over the 1620 bins,

Min.	1*stQu.*	*Median*	*Mean*	3*rdQu.*	*Max.*
0.0000	0.2159	0.2984	0.3066	0.3865	0.7958

- For $E(2I_i^2 - 2I_i + 1)$ over the 1620 bins.

Min.	1*stQu.*	*Median*	*Mean*	3*rdQu.*	*Max.*
0.5293	0.6715	0.7114	0.7184	0.7583	1.0000

Twoscales Estimation as Repair of Pre-Averaging

- Smoothed TSRV (S-TSRV):
  Preaveraging+tapering+two-scales realized variance
- The problem from the irregular times disappears as an algrebraic identity
- A good side effect: Effectively data can be synchronized between different series. In later application, for example, we pre-average transactions, and each quotes series, in 15 second intervals
- Particularly small edge effects. Of great importance when estimating spot quantities. (Which is the case here.)
- Hard to analyze (AVAR, etc). Original background for MZ paper on “Observed AVAR"

A Tapered and Smoothed TSRV (S-TSRV)

- Two scales, $K > J$. $N =$ number of intervals $[\tau_{i-1}, \tau_i)$
- Set single K-scale volatility as

$$K\widetilde{[\bar{Y}, \bar{Y}]}^{(K)} = \frac{1}{2}\sum_{i=1}^{J}(\bar{Y}_{i+K} - \bar{Y}_i)^2 + \sum_{i=J+1}^{N-b}(\bar{Y}_{i+K} - \bar{Y}_i)^2 + \frac{1}{2}\sum_{i=N-b+1}^{N-K}(\bar{Y}_{i+K} - \bar{Y}_i)^2.$$

where $b = K + J$

- $J\widetilde{[\bar{Y}, \bar{Y}]}^{(J)}$ is similar, by switching J and K
- Overall J, K scales TSRV:

$$\widehat{\langle X, X \rangle} = \frac{1}{(1 - b/N)(K - J)}\left\{K\widetilde{[\bar{Y}, \bar{Y}]}^{(K)} - J\widetilde{[\bar{Y}, \bar{Y}]}^{(J)}\right\}.$$

- Benefit of tapering: Complete elimination of edge effect due to (noise)^2: reduction in worst-case edge effect

Main Theorem (Exact Algebraic Reduction of TSRV)

- One-Scale Representation:

$$\widehat{\langle X, X\rangle} = \frac{1}{(1-b/N)(K-J)} \sum_{i=J+1}^{N-K} (X_{\tau_{i+K-J}} - X_{\tau_i})^2 + \underbrace{\text{martingale terms}}_{O_p(\sqrt{(K-J)/N})} + \underbrace{\text{edge effect}}_{o_p(\sqrt{(K-J)/N})}$$

- $\widehat{\langle X, X\rangle} \xrightarrow{p} [X, X]$: consistency under irregular times
- Edge effect:

$$\frac{1}{(1-b/N)(K-J)} \left\{ \left(-\sum_{J+1}^{K} + \sum_{N-K+1}^{N-J} \right) \left(\frac{1}{2}(\eta_i - \eta_i')\Delta X_{\tau_i} + \eta_i (X_{\tau_{i-1}} - X_{\tau_{i-J}}) \right) + \frac{1}{2}(X_{\tau_K} - X_{\tau_J})^2 + \frac{1}{2}(X_{\tau_{N-J}} - X_{\tau_{N-K}})^2 \right\}$$

where $\eta_i = \bar{X}_i - X_{\tau_{i-1}} + \bar{\epsilon}_i$ and $\eta_i' = X_{\tau_i} - \bar{X}_i - \bar{\epsilon}_i$.

Explanation of Theorem

- Martingale terms:
  - U-statistics if $E(\bar{\epsilon}_{i+J} \mid \mathcal{F}_i) = 0$ and under the statistical equivalent martingale measure of X
  - Contribute only to variance
- J must be set large enough to avoid bias
- Optimal choice for variance: $N = O(n^{1/2})$ and $K = O(1)$. The rate of convergence is $O_p(N^{-1/2}) = O_p(n^{-1/4})$, the best attainable is order $O_p(n^{-1/2})$
- The effect of the irregular times DOES show up in the variance, as with regular TSRV.

Conclusions

- A vast amount of data: high frequency, high dimension
- Building blocks: AX and POET
- POET is easier in high frequency: Residuals not required to be orthogonal
- PCA and Factor Estimation depend crucially on the quality of the estimator of the underlying spot covariance matrix
- The spot covariance matrix gets more precise when facing three trolls: Financial prices have ...
  - “error" (microstructure noise)
  - asynchronous observation
  - edge effects in estimators
- First PC very close to value weighted index: theoretically plausible, form of validation
- PC possible export to indices for for non-equity securities
- PC2 related to Fama-French factors