Incorporating GISC and High Frequency Data into Portfolio Allocation and Risk Estimation

Jianqing Fan

Princeton University

with **Alex Furger and Dacheng Xiu**

About this talk

■ Estimation large volatility matrix using high-frequency data

- How to large estimate covariance matrix?
- What is a reasonable covariance structure?
- How to use **high-frequency** data to ensure locality?
- What does the procedure perform?
- How does it impact on portfolio allocation and risk assessment?

About this talk

■ Estimation large volatility matrix using high-frequency data

- How to large estimate covariance matrix?
- What is a reasonable covariance structure?
- How to use **high-frequency** data to ensure locality?
- What does the procedure perform?
- How does it impact on portfolio allocation and risk assessment?

About this talk

■ Estimation large volatility matrix using high-frequency data

- How to large estimate covariance matrix?
- What is a reasonable covariance structure?
- How to use **high-frequency** data to ensure locality?
- What does the procedure perform?
- How does it impact on portfolio allocation and risk assessment?

Outline

1. Introduction
2. Structures of Covariance Matrix
3. High-Frequency High-dim Covariance Estimation
4. Numerical Studies

Introduction

Volatility Matrix

pervades every facet of finance. It is related to

- ★ option pricing
- ★ risk measures and risk-adjusted returns
- ★ securities regularization
- ★ portfolio allocation and proprietary trading
- ★ simulation of complex financial systems

Properties of a Good Covariance Estimator

1. Positive definite.
2. Well-conditioned to give reasonable portfolio allocations and risk management.
3. Time locality and time varying.
4. Transparent or easily understood.

Aim: Propose a simple estimator that achieves them all.

Locality of Volatility and Correlation

- use high-frequency financial data

  ★to localize volatility ★to increase sample size.

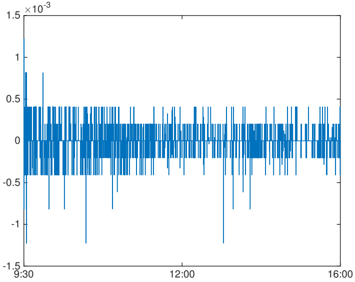
- wide availability of high-frequency data
  - ★ individual stocks, stock indices, ETFs
  - ★ foreign exchange and interest rate futures
  - ★ commodity futures
- our procedure takes advantage of this

High-frequency Data Problems

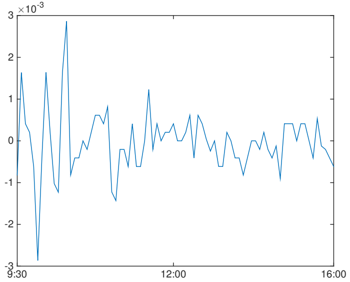
★ Microstructure noise ★Nonsynchronized trading ★Jumps

- Dealt with MSRV, realized kernel, wavelets, bipower, QMLE, pre-averaging, but required **additive** assumptions.
- Subsampling at 5-15 minutes frequency.

one-second returns

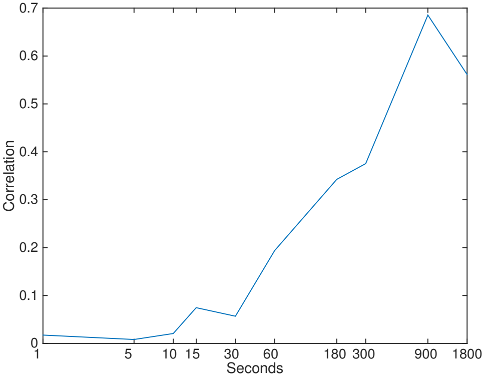


5-minute returns



Impact of Microstructural Noise: Epps Effect

Epps effect: corr between IBM and INTC on 11/14/13



High-dimensional Data Problems

1. High spurious correlations *(Fan, Shao, Zhou, 2015)*
2. Sample covariance matrix can be inconsistency.

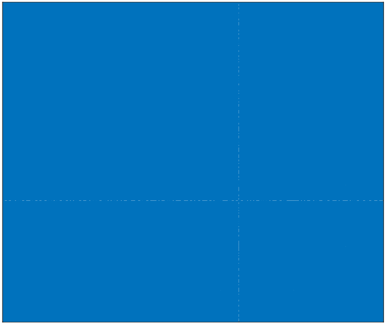
   (Tracy, Widom, Bai, Silverstein, Johnstone, ...)
3. Estimate can easily ill conditioned

■ Need to impose a structure

Covariance Structure

Sparsity

Large literature: Bickel & Levina (08), Lam & Fan (09), Cai, Zhou, Zhang, Ledoit, Wolfe ..., **but**



Monthly sample correlation at least 0.15 in at least 12 months in 2007-2009.

Conditional Sparsity

- Large literature in econ and finance on (dynamic) factor models *(Stoch, Watson, Bai, Ng, Forni, Hallin, Lippi, Reichlin, ...)*
- Estimate large covariance *(Fan, Fan, Lv 08; Fan, Liao, Martina, 13)*
- Related to low-rank + sparse: *Candes, Kolchinskii, Wainwright, Negahban, Cai, ...*

Model: $\mathbf{Y}_t = \mathbf{B}\mathbf{f}_t + \varepsilon_t$, $\quad \Sigma_Y = \mathbf{B}\Sigma_f\mathbf{B}^T + \Sigma_\varepsilon \quad$ Σ_ε sparse.

(Sharpe, 63; Lintner, 65; Ross, 76; Chamberlain & Rothschild, 83; Fama & French, 93)

Factors: **observable** or unobservable

Commonly used factors

- **Fama-French factors**: CRSP index, HML, SMB (daily)

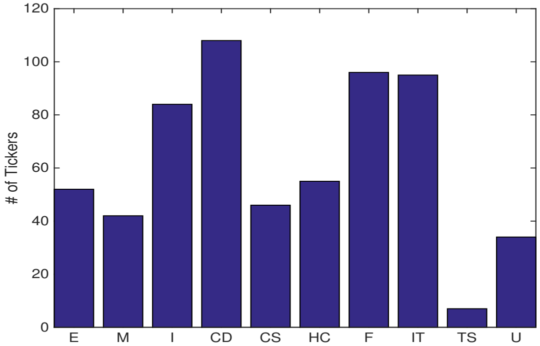
  ★Extended to 5 minutes by *Ait-Sahalia, Kalnina an Xiu (2014)*

  ★Inadequate to capture sector correlation.

- **9 sector factors** proxied by their ETFs: Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Health Care (XLV), Financial (XLF), Information Technology (XLK), Utilities (XLU)

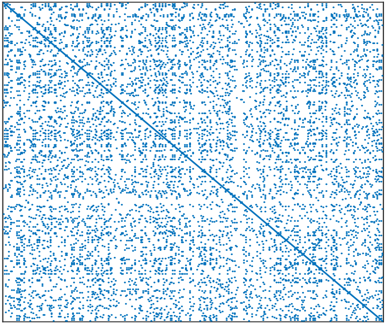
GICS Sector Classification

8-digit codes: ★Digits 1-2: sector. ★Digits 3-4: industry group. ★Digits 5-6: industry. ★Digits 7-8: sub industry.



■**No. of companies in each GICS sector** on Dec. 28th, 2012: Energy (E), Materials (M), Industrials (I), Consumer Discretionary (CD), Consumer Staples (CS), Health Care (HC), Financial (F), Information Technology (IT), Telecommunication Services (TS), Utilities (U).

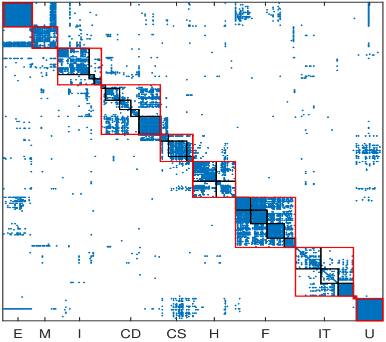
Is Conditional Sparsity Reasonable?



Non-zero Entries of the Residual Correlation Matrix (2007 - 2009) after taking out Fama-French factors

■ $\geq$ 12 months with $|\text{corr}| \geq 0.15$ ■ Based on 15 minutes data ($n = 572$) each month

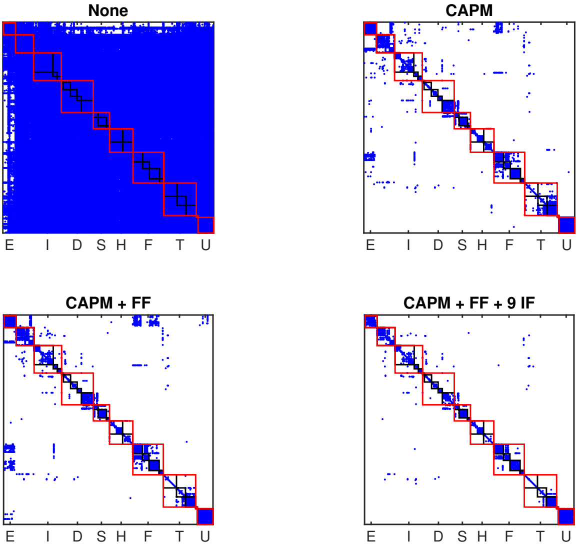
Sorting by GICS



Non-zero Entries of the Residual Correlation Matrix (2007 - 2009) after taking out Fama-French factors

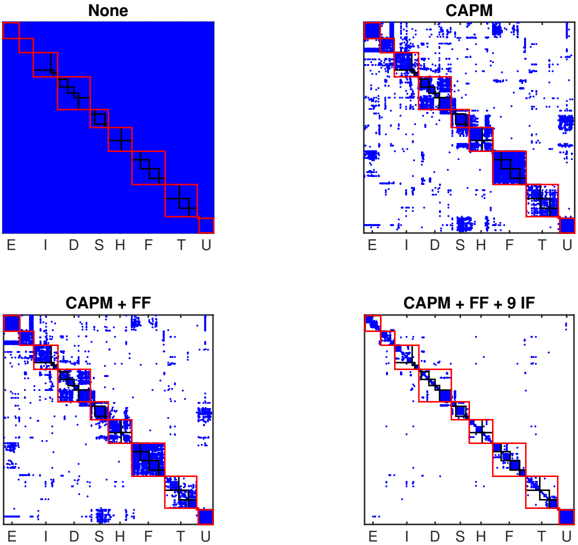
$\geq$ 12 months with $|\text{corr}| \geq 0.15$ Based on 15 minutes data ($n = 572$) each month

Sorting by GICS (Pre-Crisis: 2004-2006)



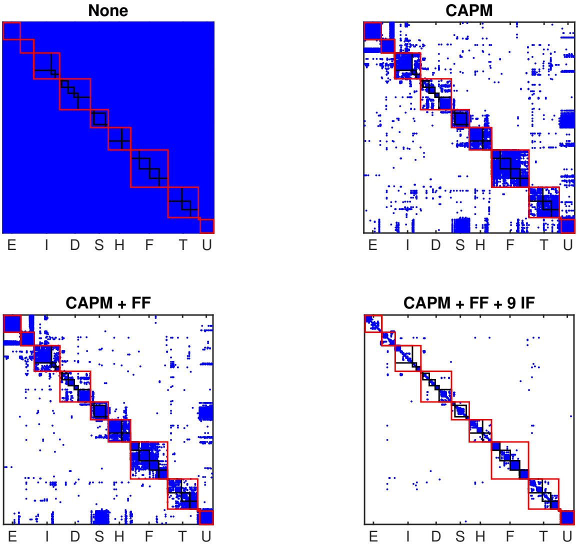
$\geq$ 12 months with $|\text{corr}| \geq 0.15$ Based on 15 minutes data ($n = 572$) each month

Sorting by GICS (Crisis: 2007-2009)



$\geq$ 12 months with $|\text{corr}| \geq 0.15$ Based on 15 minutes data ($n = 572$) each month

Sorting by GICS (Post-Crisis: 2010-2012)



$\geq$ 12 months with $|\text{corr}| \geq 0.15$ Based on 15 minutes data ($n = 572$) each month

Average R^2 and No. of Significant correlations

Pre-Crisis	R^2	Significant Correlations		
		Strict	Industry Group	Sector
CAPM	19.39% (6.57%)	3900	980	626
CAPM + FF	21.45% (6.53%)	3714	904	664
CAPM + FF + 9 IF	28.95% (9.00%)	**2156**	246	**138**

Crisis	R^2	Significant Correlations		
		Strict	Industry Group	Sector
CAPM	31.79% (7.49%)	10986	6170	2798
CAPM + FF	33.97% (7.68%)	8422	3904	1782
CAPM + FF + 9 IF	44.52% (10.09%)	**2658**	354	**204**

Post-Crisis	R^2	Significant Correlations		
		Strict	Industry Group	Sector
CAPM	37.12% (9.01%)	**13380**	6812	**3198**
CAPM + FF	39.00% (9.01%)	**10604**	4346	**2274**
CAPM + FF + 9 IF	48.48% (9.94%)	**2560**	212	**110**

High-dimensional Covariance Regularization

Block thresholding estimator

1. Estimate the **joint covariance**: $U_s = (\overbrace{Y_s^\mathsf{T}}^{d \text{ returns}}, \overbrace{X_s^\mathsf{T}}^{r \text{ factors}})^\mathsf{T}$

$$\widehat{\Pi} = \frac{1}{t} \sum_{j=1}^{[t/\Delta_n]} (U_{j\Delta} - U_{(j-1)\Delta})(U_{j\Delta} - U_{(j-1)\Delta})^\mathsf{T}$$

2. Compute **market beta** $\beta = \widehat{\Pi}_{12}\widehat{\Pi}_{22}^{-1}$ and **residual variance**

$$\widehat{\Gamma} = \widehat{\Pi}_{11} - \widehat{\Pi}_{12}(\widehat{\Pi}_{22})^{-1}\widehat{\Pi}_{21},$$

3. Sort by GICS to form blocks.
4. Obtain $\widehat{\Gamma}^S$ by truncating off-block entries of $\widehat{\Gamma}$ to 0s.
5. $\widehat{\Sigma}^S = \widehat{\beta}\widehat{\mathrm{E}}\widehat{\beta}^\mathsf{T} + \widehat{\Gamma}^S$, where $\widehat{\mathrm{E}} = \widehat{\Pi}_{22}$

Other variants and Checklist

Choice of set thresholding: S can be one of the folliwng

$$S_1 = \{(i,j) \text{ such that the } ith \text{ and } jth \text{ assets belong to the same sector}\},$$
$$S_2 = \{(i,j) \text{ such that the } ith \text{ and } jth \text{ assets belong to the same industry}\},$$
$$S_3 = \{(i,j) \text{ such that } i = j\}.$$

- Positive-Definite? $\checkmark$
- Better Conditioned? $\checkmark$
- Time Varying? $\checkmark$
- *Simple!!* $\checkmark$

Assumptions

- Asset returns and factors follow Itô semimartingales:

$$\mathbf{Y_t} = \beta \mathbf{X_t} + \mathbf{Z_t}, \qquad X_t = \int_0^t h_s\, ds + \int_0^t \eta_s dW_s$$

$$Z_t = \int_0^t f_s ds + \int_0^t \gamma_s dB_s, \qquad [X, Z^T] = 0.$$

$$\Longrightarrow \Sigma_{d\times d} = \beta_{d\times r} \mathrm{E}_{r\times r} \beta^{\mathsf{T}}_{d\times r} + \Gamma_{d\times d},$$

where $\Sigma = \frac{1}{t}\int_0^t (\beta \eta_s \eta_s^{\mathsf{T}} \beta^{\mathsf{T}} + \gamma_s \gamma_s^T) ds$ and $\Gamma = \frac{1}{t}\int_0^t \gamma_s \gamma_s^{\mathsf{T}} ds$.

- Factors are **pervasive**: $\|d^{-1}\beta^T\beta - B\| = o(1)$, and the residual covariance matrix is block-diagonal: $\Gamma_{S^c} = 0$.

- $r = o(d)$, $\Delta_n r^4 \log(d) = o(1)$.

- $m_d r \sqrt{\Delta_n \log d} = o(1)$, where m_d is the largest nonzero

Asymptotic Properties

- **Theorem 1**: For any $S = S_1, S_2, S_3$, without any assumption on m_d:

$$||\widehat{\Sigma}^S - \Sigma||_{\max} = O_p\left(\sqrt{r^4 \Delta_n \log d}\right).$$

- **Theorem 2**: With sparsity assumption on m_d:

$$||(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}|| = O_p\left(m_d \sqrt{r^2 \Delta_n \log d}\right).$$

- Comparable rates with i.i.d. case.

Remarks

- In-fill asymptotic for continuous time model.

- **Risk approx**: $\left|\omega^{*T}\widehat{\Sigma}^S\widehat{\omega}^* - \omega^*\Sigma\omega^*\right| \le ||\widehat{\Sigma}^S - \Sigma||_{\max}||\widehat{\omega}^*||_1^2.$

- **Sparse portfolio alloc.**: $\min_\omega \omega^T\widehat{\Sigma}^S\omega, \quad \text{s.t. } \omega^T 1 = 1, \|\omega\|_1 \le \gamma.$

$$\left|\widehat{\omega}^{*T}\widehat{\Sigma}^S\widehat{\omega}^* - \omega^*\Sigma\omega^*\right| \le ||\widehat{\Sigma}^S - \Sigma||_{\max}\gamma^2.$$

- **Min variance**: $\left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}$

$$\left|\left(\mathbf{1}^T\left(\widehat{\Sigma}^S\right)^{-1}\mathbf{1}\right)^{-1} - \left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}\right| \le 2\left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-2}||\left(\widehat{\Sigma}^S\right)^{-1} - \Sigma^{-1}||,$$

Remarks

- In-fill asymptotic for continuous time model.
- **Risk approx**: $\left|\omega^{*T}\widehat{\Sigma}^S\widehat{\omega}^* - \omega^*\Sigma\omega^*\right| \le ||\widehat{\Sigma}^S - \Sigma||_{\max}||\widehat{\omega}^*||_1^2.$
- **Sparse portfolio alloc.**: $\min_\omega \omega^T\widehat{\Sigma}^S\omega, \quad \text{s.t. } \omega^T 1 = 1, \|\omega\|_1 \le \gamma.$

$$\left|\widehat{\omega}^{*T}\widehat{\Sigma}^S\widehat{\omega}^* - \omega^*\Sigma\omega^*\right| \le ||\widehat{\Sigma}^S - \Sigma||_{\max}\gamma^2.$$

- **Min variance**: $\left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}$

$$\left|\left(\mathbf{1}^T\left(\widehat{\Sigma}^S\right)^{-1}\mathbf{1}\right)^{-1} - \left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}\right| \le 2\left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-2}||\left(\widehat{\Sigma}^S\right)^{-1} - \Sigma^{-1}||,$$

Remarks

- In-fill asymptotic for continuous time model.
- **Risk approx**: $\left|\omega^{*T}\widehat{\Sigma}^S\widehat{\omega}^* - \omega^*\Sigma\omega^*\right| \le ||\widehat{\Sigma}^S - \Sigma||_{\max}||\widehat{\omega}^*||_1^2.$
- **Sparse portfolio alloc.**: $\min_\omega \omega^T\widehat{\Sigma}^S\omega, \quad \text{s.t. } \omega^T 1 = 1, \|\omega\|_1 \le \gamma.$

$$\left|\widehat{\omega}^{*T}\widehat{\Sigma}^S\widehat{\omega}^* - \omega^*\Sigma\omega^*\right| \le ||\widehat{\Sigma}^S - \Sigma||_{\max}\gamma^2.$$

- **Min variance**: $\left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}$

$$\left|\left(\mathbf{1}^T\left(\widehat{\Sigma}^S\right)^{-1}\mathbf{1}\right)^{-1} - \left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}\right| \le 2\left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-2}||\left(\widehat{\Sigma}^S\right)^{-1} - \Sigma^{-1}||,$$

Numerical Studies

Simulation Designs

Parameters: $d = 500$ assets, $r = 3$ factors, $N_{sim} = 100$

Factor Model: $dY_{i,t} = \sum_{j=1}^{r} \beta_{i,j} dX_{j,t} + dZ_{i,t}$

Loadings: $\beta_1 \sim \mathcal{U}[0.25, 2.25]$, and $\beta_2, \beta_3 \sim \mathcal{U}[-0.5, 0.5]$.

Dynamic of Factors: $dX_{j,t} = b_j dt + \sigma_{j,t} dW_{j,t}$ with stochastic volatility (CIR):
$d\sigma^2_{j,t} = \kappa_j(\theta_j - \sigma^2_{j,t})dt + \eta_j \sigma_{j,t} d\widetilde{W}_{j,t}$, leverage $\mathbb{E}[dW_{j,t} d\widetilde{W}_{j,t}] = \rho_j dt$.
$\kappa = (3, 4, 5)$, $\theta = (0.09, 0.04, 0.06)$, $\eta = (0.3, 0.4, 0.3)$,
$\rho = (-0.6, -0.4, -0.250)$, and $b = (0.05, 0.03, 0.02)$.

Noise: $dZ_{i,t} = \gamma_i^T dB_{i,t}$

10 blocks (of size 50×50): with-block corr = 0.25, vol$\sim \mathcal{U}[0.25, 2.25]$.

Simulation Designs

Parameters: $d = 500$ assets, $r = 3$ factors, $N_{sim} = 100$

Factor Model: $dY_{i,t} = \sum_{j=1}^{r} \beta_{i,j} dX_{j,t} + dZ_{i,t}$

Loadings: $\beta_1 \sim \mathcal{U}[0.25, 2.25]$, and $\beta_2, \beta_3 \sim \mathcal{U}[-0.5, 0.5]$.

Dynamic of Factors: $dX_{j,t} = b_j dt + \sigma_{j,t} dW_{j,t}$ with stochastic volatility (CIR):

$d\sigma^2_{j,t} = \kappa_j(\theta_j - \sigma^2_{j,t})dt + \eta_j \sigma_{j,t} d\widetilde{W}_{j,t}$, leverage $\mathbb{E}[dW_{j,t} d\widetilde{W}_{j,t}] = \rho_j dt$.

$\kappa = (3, 4, 5)$, $\theta = (0.09, 0.04, 0.06)$, $\eta = (0.3, 0.4, 0.3)$,

$\rho = (-0.6, -0.4, -0.250)$, and $b = (0.05, 0.03, 0.02)$.

Noise: $dZ_{i,t} = \gamma_i^T dB_{i,t}$

10 blocks (of size 50×50): with-block corr = 0.25, vol $\sim \mathcal{U}[0.25, 2.25]$.

Simulation Designs

Parameters: $d = 500$ assets, $r = 3$ factors, $N_{sim} = 100$

Factor Model: $dY_{i,t} = \sum_{j=1}^{r} \beta_{i,j} dX_{j,t} + dZ_{i,t}$

Loadings: $\beta_1 \sim \mathcal{U}[0.25, 2.25]$, and $\beta_2, \beta_3 \sim \mathcal{U}[-0.5, 0.5]$.

Dynamic of Factors: $dX_{j,t} = b_j dt + \sigma_{j,t} dW_{j,t}$ with stochastic volatility (CIR):

$d\sigma_{j,t}^2 = \kappa_j(\theta_j - \sigma_{j,t}^2)dt + \eta_j \sigma_{j,t} d\widetilde{W}_{j,t}$, leverage $\mathbb{E}[dW_{j,t} d\widetilde{W}_{j,t}] = \rho_j dt$.

$\kappa = (3, 4, 5)$, $\theta = (0.09, 0.04, 0.06)$, $\eta = (0.3, 0.4, 0.3)$,

$\rho = (-0.6, -0.4, -0.250)$, and $b = (0.05, 0.03, 0.02)$.

Noise: $dZ_{i,t} = \gamma_i^T dB_{i,t}$

10 blocks (of size 50×50): with-block corr = 0.25, vol$\sim \mathcal{U}[0.25, 2.25]$.

Microstructure Noise and Nonsynchronized trading

Microstructural noise: add $N(0, 0.005^2)$ to the log-prices

Nonsynchronized trading:

- ★ Generate data at 1-second frequency for 6.5 hours;
- ★ Get number N_i from log-normal with $\mu = 2500$ and $\sigma = 0.8$ truncated at 500 and 23400;
- ★ Subsample N_i data at random as the trading prices of the i^{th} asset.

Length of Data: one month.

Simulation Results

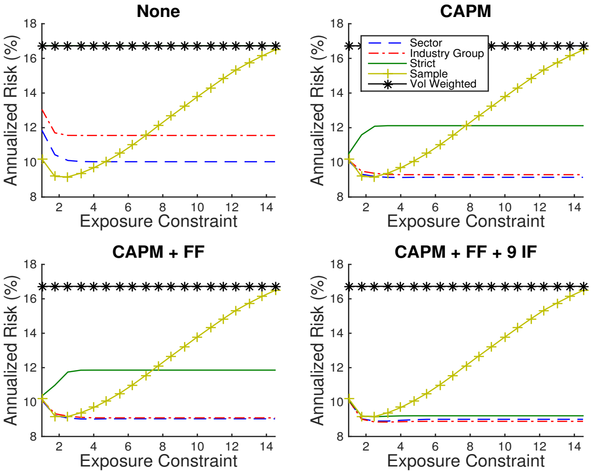
	Frequency	5 sec	15 sec	30 sec	1 min	5 min	15 min	30 min	65 min
I	$\|\widehat{\Sigma}-\Sigma\|_{\max}$	59.29	20.09	10.21	5.22	1.18	**0.52**	0.39	0.38
	$\|(\widehat{\Sigma}^S)^{-1}-\Sigma^{-1}\|$	6.62	6.59	6.55	6.46	5.86	5.13	4.76	25.07
II	$\|\widehat{\Sigma}-\Sigma\|_{\max}$	0.01	0.02	0.03	0.04	0.08	0.14	0.20	0.30
	$\|(\widehat{\Sigma}^S)^{-1}-\Sigma^{-1}\|$	0.28	0.49	0.70	1.03	2.68	6.06	11.82	36.49

★I: With ★II: Without microstructure noise

Minimum variance portfilios

Period: Jan 2007 – Dec. 2012.

Output: Risk of monthly rebalanced portfolios



Number of Stocks

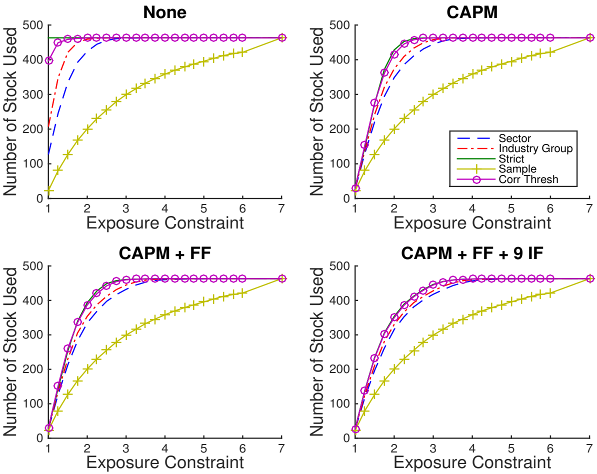


Figure: Number of Stocks Selected in the Optimized Portfolios

Conclusion

★ Propose a covariance matrix estimator for a large panel of high-frequency stock returns.

- It is positive-definite.
- It does not have any tuning parameters.
- It allows time-varying volatilities and correlations.

★ Empirically, this estimator achieves a much lower out-of-sample risk than many alternatives.

★ Establish rates of convergence for continuous time finance model.

The End

Thank  *You*