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

Jianqing Fan
Princeton University
with Alex Furger and Dacheng Xiu
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?
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

![Graph showing correlation between IBM and INTC over time](image)
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
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:** \( Y_t = B f_t + \varepsilon_t \), \( \Sigma_Y = B \Sigma_f B^T + \Sigma_\varepsilon \) \( \Sigma_\varepsilon \) 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)

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).

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

[Diagram showing correlation matrix with non-zero entries highlighted]

- 12 months with $|\text{corr}| \geq 0.15$
- Based on 15 minutes data ($n = 572$) each month
Sorting by GICS (Pre-Crisis: 2004-2006)

None

CAPM

CAPM + FF

CAPM + FF + 9 IF

≥ 12 months with $|\text{corr}| \geq 0.15$

Based on 15 minutes data ($n = 572$) each month

Jianqing Fan (Princeton University)  High-Frequency High-Dimensional Volatility
Sorting by GICS (Crisis: 2007-2009)

\[ \begin{align*}
\text{None} & : \quad \begin{array}{ccccccc}
E & I & D & S & H & F & T & U \\
\end{array} \\
\text{CAPM} & : \quad \begin{array}{ccccccc}
E & I & D & S & H & F & T & U \\
\end{array} \\
\text{CAPM + FF} & : \quad \begin{array}{ccccccc}
E & I & D & S & H & F & T & U \\
\end{array} \\
\text{CAPM + FF + 9 IF} & : \quad \begin{array}{ccccccc}
E & I & D & S & H & F & T & U \\
\end{array}
\end{align*} \]

\[ \begin{align*}
\geq 12 \text{ months with } |\text{corr}| \geq 0.15 & \text{ Based on 15 minutes data (} n = 572 \text{) each month}
\end{align*} \]

Jianqing Fan (Princeton University)  
High-Frequency High-Dimensional Vol
Sorting by GICS (Post-Crisis: 2010-2012)

- None
- CAPM
- CAPM + FF
- CAPM + FF + 9 IF

- \(|\text{corr}| \geq 0.15\)
- Based on 15 minutes data \((n = 572)\) each month
## Average $R^2$ and No. of Significant correlations

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
<th>Pre-Crisis</th>
<th>Crisis</th>
<th>Post-Crisis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Significant Correlations</td>
<td>Strict</td>
<td>Industry Group</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pre-Crisis</strong></td>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPM</td>
<td>19.39% (6.57%)</td>
<td>3900</td>
<td>980</td>
<td>626</td>
</tr>
<tr>
<td>CAPM + FF</td>
<td>21.45% (6.53%)</td>
<td>3714</td>
<td>904</td>
<td>664</td>
</tr>
<tr>
<td>CAPM + FF + 9 IF</td>
<td>28.95% (9.00%)</td>
<td><strong>2156</strong></td>
<td>246</td>
<td><strong>138</strong></td>
</tr>
<tr>
<td><strong>Crisis</strong></td>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPM</td>
<td>31.79% (7.49%)</td>
<td>10986</td>
<td>6170</td>
<td>2798</td>
</tr>
<tr>
<td>CAPM + FF</td>
<td>33.97% (7.68%)</td>
<td>8422</td>
<td>3904</td>
<td>1782</td>
</tr>
<tr>
<td>CAPM + FF + 9 IF</td>
<td>44.52% (10.09%)</td>
<td><strong>2658</strong></td>
<td>354</td>
<td><strong>204</strong></td>
</tr>
<tr>
<td><strong>Post-Crisis</strong></td>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPM</td>
<td>37.12% (9.01%)</td>
<td><strong>13380</strong></td>
<td>6812</td>
<td><strong>3198</strong></td>
</tr>
<tr>
<td>CAPM + FF</td>
<td>39.00% (9.01%)</td>
<td><strong>10604</strong></td>
<td>4346</td>
<td><strong>2274</strong></td>
</tr>
<tr>
<td>CAPM + FF + 9 IF</td>
<td>48.48% (9.94%)</td>
<td><strong>2560</strong></td>
<td>212</td>
<td><strong>110</strong></td>
</tr>
</tbody>
</table>

Jianqing Fan (Princeton University)  
High-Frequency High-Dimensional Vol
High-dimensional Covariance Regularization
**Block thresholding estimator**

1. Estimate the **joint covariance**: \( U_s = (Y_s^T, X_s^T)^T \)

\[
\hat{\Pi} = \frac{1}{t} \left[ \sum_{j=1}^{[t/\Delta n]} (U_{j\Delta} - U_{(j-1)\Delta}) (U_{j\Delta} - U_{(j-1)\Delta})^T \right]
\]

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

\[
\hat{\Gamma} = \hat{\Pi}_{11} - \hat{\Pi}_{12} (\hat{\Pi}_{22})^{-1} \hat{\Pi}_{21},
\]

3. Sort by GICS to form blocks.

4. Obtain \( \hat{\Gamma}^S \) by truncating off-block entries of \( \hat{\Gamma} \) to 0s.

5. \( \hat{\Sigma}^S = \hat{\beta} \hat{E} \hat{\beta}^T + \hat{\Gamma}^S \), where \( \hat{E} = \hat{\Pi}_{22} \)
Choice of set thresholding: $S$ can be one of the following:

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

- Positive-Definite? ✓
- Better Conditioned? ✓
- Time Varying? ✓
- Simple!! ✓
Assumptions

- Asset returns and factors follow Itô semimartingales:

  \[ Y_t = \beta X_t + Z_t, \quad X_t = \int_0^t h_s \, ds + \int_0^t \eta_s \, dW_s \]

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

  \[ \implies \Sigma_{d \times d} = \beta_{d \times r} E_{r \times r} \beta_{d \times r}^T + \Gamma_{d \times d}, \]

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

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

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

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

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

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

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

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

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

- In-fill asymptotic for continuous time model.

- **Risk approx:** \[ |\omega^*^T \hat{\Sigma}_S \omega^* - \omega^* \Sigma \omega^*| \leq \| \hat{\Sigma}_S - \Sigma \|_{\text{max}} \| \omega^* \|^2. \]

- **Sparse portfolio alloc.:** \[ \min_{\omega} \omega^T \hat{\Sigma}_S \omega, \quad \text{s.t. } \omega^T 1 = 1, \| \omega \|_1 \leq \gamma. \]

\[ |\hat{\omega}^*^T \hat{\Sigma}_S \hat{\omega}^* - \omega^* \Sigma \omega^*| \leq \| \hat{\Sigma}_S - \Sigma \|_{\text{max}} \gamma^2. \]

- **Min variance:** \( (1^T \Sigma^{-1} 1)^{-1} \)

\[ \left| \left( 1^T \left( \hat{\Sigma}_S \right)^{-1} 1 \right)^{-1} - \left( 1^T \Sigma^{-1} 1 \right)^{-1} \right| \leq 2 \left( 1^T \Sigma^{-1} 1 \right)^{-2} \| \left( \hat{\Sigma}_S \right)^{-1} - \Sigma^{-1} \|, \]
Remarks

- In-fill asymptotic for continuous time model.

- Risk approx: \[ |\omega^* \hat{\Sigma} \hat{\omega}^* - \omega^* \Sigma \omega^*| \leq \|\hat{\Sigma}^S - \Sigma\|_{\text{max}} \|\hat{\omega}^*\|^2 \].

- Sparse portfolio alloc.: \[ \min_\omega \omega^T \hat{\Sigma}^S \omega, \quad \text{s.t.} \quad \omega^T 1 = 1, \|\omega\|_1 \leq \gamma \].

  \[ |\hat{\omega}^* \hat{\Sigma} \hat{\omega}^* - \omega^* \Sigma \omega^*| \leq \|\hat{\Sigma}^S - \Sigma\|_{\text{max}} \gamma^2. \]

- Min variance: \( (1^T \Sigma^{-1} 1)^{-1} \)

\[
\left| \left(1^T (\hat{\Sigma}^S)^{-1} 1 \right)^{-1} - (1^T \Sigma^{-1} 1)^{-1} \right| \leq 2 (1^T \Sigma^{-1} 1)^{-2} \| (\hat{\Sigma}^S)^{-1} - \Sigma^{-1} \|.
\]
Remarks

- **In-fill asymptotic for continuous time model.**

- **Risk approx.** \[ \left| \omega^* \mathbf{1}^T \hat{\Sigma} \hat{\omega}^* - \omega^* \mathbf{1}^T \Sigma \hat{\omega}^* \right| \leq \| \hat{\Sigma} \mathbf{1}^T \Sigma - \Sigma \|_{\text{max}} \| \hat{\omega}^* \|_1^2. \]

- **Sparse portfolio alloc.** \[ \min_{\omega} \omega^T \hat{\Sigma} \mathbf{1}^T \hat{\Sigma} \omega, \quad \text{s.t.} \quad \omega^T \mathbf{1} = 1, \| \omega \|_1 \leq \gamma. \]

- \[ \left| \hat{\omega}^* \mathbf{1}^T \hat{\Sigma} \hat{\omega}^* - \omega^* \mathbf{1}^T \Sigma \hat{\omega}^* \right| \leq \| \hat{\Sigma} \mathbf{1}^T \Sigma - \Sigma \|_{\text{max}} \gamma^2. \]

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

- \[ \left| \left( \mathbf{1}^T \left( \hat{\Sigma} \right)^{-1} \mathbf{1} \right)^{-1} - \left( \mathbf{1}^T \Sigma^{-1} \mathbf{1} \right)^{-1} \right| \leq 2 \left( \mathbf{1}^T \Sigma^{-1} \mathbf{1} \right)^{-2} \| \left( \hat{\Sigma} \right)^{-1} - \Sigma^{-1} \|. \]
Numerical Studies
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):
\[
  d\sigma_{j,t}^2 = \kappa_j (\theta_j - \sigma_{j,t}^2) \, dt + \eta_j \sigma_{j,t} \, d\tilde{W}_{j,t},
\]
leverage \( \mathbb{E}[dW_{j,t} d\tilde{W}_{j,t}] = \rho_j \, dt \).

\( \kappa = (3, 4, 5) \), \( \theta = (0.09, 0.04, 0.06) \), \( \eta = (0.3, 0.4, 0.3) \), \( \rho = (-0.6, -0.4, -0.250) \), and \( b = (0.05, 0.03, 0.02) \).

Noise: \( dZ_{i,t} = \gamma_j^T dB_{i,t} \)

10 blocks (of size 50 \( \times \) 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**: \( \beta_1 \sim U[0.25, 2.25] \), and \( \beta_2, \beta_3 \sim U[-0.5, 0.5] \).

**Dynamic of Factors**: \( dX_{j,t} = b_j dt + \sigma_{j,t} dW_{j,t} \) with stochastic volatility (CIR):
\[
\begin{align*}
    d\sigma_{j,t}^2 &= \kappa_j (\theta_j - \sigma_{j,t}^2) dt + \eta_j \sigma_{j,t} d\tilde{W}_{j,t}, \\
    \kappa &= (3, 4, 5), & \theta &= (0.09, 0.04, 0.06), & \eta &= (0.3, 0.4, 0.3), \\
    \rho &= (-0.6, -0.4, -0.250), & \text{and } b &= (0.05, 0.03, 0.02).
\end{align*}
\]

**Noise**: \( dZ_{i,t} = \gamma_i^T dB_{i,t} \)

10 blocks (of size 50 \( \times \) 50): with-block corr = 0.25, vol \( \sim U[0.25, 2.25] \).
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: $dX_{j,t} = b_j dt + \sigma_{j,t} d\tilde{W}_{j,t}$ with stochastic volatility (CIR):
$d\sigma_{j,t}^2 = \kappa_j (\theta_j - \sigma_{j,t}^2) dt + \eta_j \sigma_{j,t} d\tilde{W}_{j,t}$, leverage $\mathbb{E}[dW_{j,t} d\tilde{W}_{j,t}] = \rho_j dt$.

$k = (3, 4, 5)$, $\theta = (0.09, 0.04, 0.06)$, $\eta = (0.3, 0.4, 0.3)$, $\rho = (-0.6, -0.4, -0.250)$, and $b = (0.05, 0.03, 0.02)$.

Noise: $dZ_{i,t} = \gamma_{i,t} dB_{i,t}$

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

<table>
<thead>
<tr>
<th>I</th>
<th>Frequency</th>
<th>5 sec</th>
<th>15 sec</th>
<th>30 sec</th>
<th>1 min</th>
<th>5 min</th>
<th>15 min</th>
<th>30 min</th>
<th>65 min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>|(|\hat{\Sigma} - \Sigma||_\text{max}</td>
<td>59.29</td>
<td>20.09</td>
<td>10.21</td>
<td>5.22</td>
<td>1.18</td>
<td>0.52</td>
<td>0.39</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>|(|\hat{(\Sigma^S)}^{-1} - \Sigma^{-1}|</td>
<td>6.62</td>
<td>6.59</td>
<td>6.55</td>
<td>6.46</td>
<td>5.86</td>
<td>5.13</td>
<td>4.76</td>
<td>25.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II</th>
<th>Frequency</th>
<th>5 sec</th>
<th>15 sec</th>
<th>30 sec</th>
<th>1 min</th>
<th>5 min</th>
<th>15 min</th>
<th>30 min</th>
<th>65 min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>|(|\hat{\Sigma} - \Sigma||_\text{max}</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.08</td>
<td>0.14</td>
<td>0.20</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>|(|\hat{(\Sigma^S)}^{-1} - \Sigma^{-1}|</td>
<td>0.28</td>
<td>0.49</td>
<td>0.70</td>
<td>1.03</td>
<td>2.68</td>
<td>6.06</td>
<td>11.82</td>
<td>36.49</td>
</tr>
</tbody>
</table>

⭐️ I: With ⭐️ II: Without microstructure noise
**Minimum variance portfolios**


**Output**: Risk of monthly rebalanced portfolios

Jianqing Fan (Princeton University)  
High-Frequency High-Dimensional Vol
Number of Stocks

![Graphs showing the number of stocks selected in the optimized portfolios for different models: None, CAPM, CAPM + FF, and CAPM + FF + 9 IF. Each graph plots the number of stocks used against exposure constraint, with different line styles and colors representing different criteria.]

Figure: Number of Stocks Selected in the Optimized Portfolios

Jianqing Fan (Princeton University)
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