Principal Component Estimation of a Large Covariance Matrix with High-Frequency Data

Yacine Aït-Sahalia †  Dacheng Xiu ‡

†Department of Economics, Princeton University
‡Booth School of Business, University of Chicago

Workshop on Algorithms for Modern Massive Data Set

June 22, 2016
Motivation

Economic theories, APT by Ross (1976) and the ICAPM by Merton (1973) predict that

- Assets earn risk premia because they are exposed to risk factors
- The co-movement between assets is driven by their exposure to these common risk factors
In the literature, much effort has been devoted to searching for observable empirical proxies for the factors.

In practice, using factors or more precisely characteristics, the covariance matrix provided by MSCI Barra has been the industrial standard for decades.
However, methodologies that rely on daily historical data are developed when markets were, by modern standards, more quaint than quant.

- A long time span is required for estimation, so that factors and covariance matrix based on low frequency data adjust slowly to the time varying investment opportunities.
- Daily returns may not be informative enough about risk.
Figure: S&P 500 Index on October 10th, 2008
Main Contributions

- Using the principal component analysis (PCA), we document a finer structure of a large cross-section of high-frequency equity returns.
- We propose a covariance matrix estimator using this latent structure.
  - It is well-conditioned to give reasonable (and economically feasible) portfolio allocations.
  - It also takes advantage of the established time locality of volatility and correlation.
    - It is simple.
- We also provide a new estimator to determine the number of factors. Factors and their loadings are also estimated.
Main Empirical Findings

- The statistical factors explain more variations than that by economic factors such as the market portfolio, the Fama-French portfolios, as well as the industrial ETF portfolios.
- The residual covariance matrix by the PCA is more sparse and not diagonal.
- The optimal portfolio based on a factor model has better out-of-sample performance than that using the sample covariance matrix.
Main Idea

- Even with intraday data, the curse of dimensionality remains.
  - Ex: 1 month of 15-minute returns for 500 stocks, i.e., 500 - 600 observations per ticker
  - Sample covariance matrix is ill-conditioned

- Why not sample more frequently?
  - Microstructure noise
  - Asynchronicity
We need more structure on the covariance matrix

▶ Sparsity?
  ▶ Entries decay as they move away from the diagonal: Bandable, Block-Bandable, Toeplitz, etc.
  ▶ A small number of nonzero entries on each row/column.
  ▶ Imposing the above structure on the inverse.
  ▶ ... 

▶ Factor model?
  ▶ Classical (or strict) factor model
    ▶ Factors alone may not fully explain the co-movement of asset returns, e.g., Pepsi and Coca-Cola, Intel and IBM
  ▶ Approximate factor model
Figure: Significant Entries of the Sample Covariance Matrix
Figure: Sparsity Pattern of the Regression Residual based on known Economic Factors
Figure: Sparsity Pattern of the Regression Residual Sorted by GICS
We make use of a **Low Rank Plus Sparsity** structure on the covariance matrix $\Sigma$.

**Figure:** Decomposition of the Covariance Matrix
Model Setup and Assumptions

**Assumption 1** $Y_t$ is a $d$-dimensional vector process, $X_t$ is a $r$-dimensional *unobservable factor* process, $Z_t$ is the idiosyncratic component:

$$Y_t = \beta X_t + Z_t,$$

$$X_t = \int_0^t h_s \, ds + \int_0^t \eta_s \, dW_s, \quad Z_t = \int_0^t f_s \, ds + \int_0^t \gamma_s \, dB_s,$$

where all entries of $h$, $\eta$, $f$, and $\gamma$ are bounded uniformly by a locally bounded process $H$.

We pose no assumptions on stationarity or heteroscedasticity. We thereby measure realized covariance $[\cdot, \cdot]$ within a fixed window instead of the usual covariance.
Assumption 2 For any $1 \leq j \leq r$, $1 \leq k \leq d$, and $0 \leq s \leq t$,

$$[Z_{k,s}, X_{j,s}] = 0.$$ 

This leads to the key decomposition:

$$\Sigma_{d \times d} = \begin{bmatrix} \beta_{r \times r} E & \beta_{r \times d}^T \end{bmatrix} + \Gamma_{d \times d},$$

where, without ambiguity

$$\Sigma = \frac{1}{t} \int_0^t c_s ds, \quad \Gamma = \frac{1}{t} \int_0^t g_s ds, \quad \text{and} \quad E = \frac{1}{t} \int_0^t e_s ds.$$ 

where for $0 \leq s \leq t$,

$$c_s = \beta e_s \beta^T + g_s, \quad g_s = \gamma_s \gamma_s^T, \quad \text{and} \quad e_s = \eta_s \eta_s^T.$$
To complete the model, we need an additional assumption on the residual covariance matrix $\Gamma$. We define

$$m_d = \max_{1 \leq i \leq d} \sum_{1 \leq j \leq d} 1\{\Gamma_{ij} \neq 0\}$$  \hspace{1cm} (3)$$

and impose the so-called sparsity assumption on $\Gamma$, i.e., $\Gamma$ cannot have too many non-zero elements.

**Assumption 3** When $d \to \infty$, the degree of sparsity of $\Gamma$, $m_d$, grows at a rate which satisfies

$$d^{-a} m_d \to 0$$  \hspace{1cm} (4)$$

where $a$ is some positive constant.
To ensure identification, we also need the so called pervasive assumption:

**Assumption 4** $\Omega$ is a positive-definite covariance matrix, with distinct eigenvalues bounded away from 0. Moreover,

$$\| d^{-1} \beta^T \beta - I_r \| = o(1), \text{ as } d \to \infty.$$
Identification

**Theorem** Suppose Assumptions 1, 2, 3 with $a = 1/2$, and 4 hold. Also, assume that $\|E\|_{\text{MAX}} \leq K$, $\|\Gamma\|_{\text{MAX}} \leq K$ almost surely. Then $r$, $\beta E \beta^T$, and $\Gamma$ can be identified as $d \to \infty$. That is, if $d$ is sufficiently large, $\bar{r} = r$, where 

$$
\bar{r} = \arg \min_{1 \leq j \leq d} (d^{-1} \lambda_j + jd^{-1/2} m_d) - 1,
$$

and $\{\lambda_j, 1 \leq j \leq d\}$ are the eigenvalues of $\Sigma$. Moreover, suppose $\{\xi_j, 1 \leq j \leq d\}$ are the corresponding eigenvectors of $\Sigma$, we have

$$
\left\| \sum_{j=1}^{\bar{r}} \lambda_j \xi_j \xi_j^T - \beta E \beta^T \right\|_{\text{MAX}} \leq Kd^{-1/2} m_d, \quad \text{and}
$$

$$
\left\| \sum_{j=\bar{r}+1}^{d} \lambda_j \xi_j \xi_j^T - \Gamma \right\|_{\text{MAX}} \leq Kd^{-1/2} m_d.
$$
Estimation Procedure

To fix ideas, let $\Delta_i^n X = X_i \Delta_n - X_{(i-1)\Delta_n}$, for $1 \leq i \leq n = [t/\Delta_n]$. Our estimator is built on the principal component analysis of the sample covariance matrix estimator. Denote

$$\hat{\Sigma} = \frac{1}{t} \sum_{i=1}^{n} (\Delta_i^n X)(\Delta_i^n X)^\top.$$

Suppose that $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_d$ are the simple eigenvalues of $\hat{\Sigma}$, and that $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_d$ are the corresponding eigenvectors.
Once we have an estimator of $r$, $\hat{r}$, we can estimate $\Gamma$ as:

$$\hat{\Gamma} = \sum_{j=\hat{r}+1}^{d} \lambda_j \hat{\xi}_j \hat{\xi}^\top_j$$

Then we need to impose sparsity so as to improve the estimates. Instead of using thresholding, we use the following assumption to avoid tuning parameters:

**Assumption 5** $\Gamma$ is a block diagonal matrix, and the set of its non-zero entries, denoted by $S$, is known prior to the estimation.
Our covariance matrix estimator $\hat{\Sigma}^S$ is given by

$$\hat{\Sigma}^S = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{\Gamma}^S,$$

(5)

where $\hat{r}$ is an estimator of $r$ discussed below,

$$\hat{\Gamma} = \sum_{j=\hat{r}+1}^{d} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T, \quad \text{and} \quad \hat{\Gamma}^S = (\hat{\Gamma}_{ij} 1_{(i,j) \in S}),$$

(6)

The residual covariance matrix estimator $\hat{\Gamma}^S$ is a by-product.
Determine the Number of Factors

To determine the number of factors, we propose the following estimator using a penalty function:

\[ \hat{r} = \arg \min_{1 \leq j \leq r_{\text{max}}} \left( d^{-1} \lambda_j(\hat{\Sigma}) + j \times g(n, d) \right) - 1. \]

where \( g(n, d) \to 0 \), and \( g(n, d) \left( (\Delta_n \log d)^{1/2} + d^{-1} m_d \right)^{-1} \to \infty \).

In Bai and Ng (2002), their objective function can be written as

\[ d^{-1} \sum_{k=j+1}^{d} \lambda_k(\hat{\Sigma}) + \text{penalty}. \]
Equivalently, our estimator can be written as

\[ \hat{\Sigma}^S = t^{-1} F G G^T F^T + \hat{\Gamma}^S, \quad \hat{\Gamma} = t^{-1} (\mathcal{Y} - F G) (\mathcal{Y} - F G)^T, \tag{7} \]

and

\[ \hat{\Gamma}^S = (\hat{\Gamma}_{ij} 1_{(i,j) \in S}), \tag{8} \]

where \( \mathcal{Y} = (\Delta_1^n Y, \Delta_2^n Y, \ldots, \Delta_n^n Y) \) is a \( d \times n \) matrix, \( G = (g_1, g_2, \ldots, g_n) \) is \( \hat{r} \times n \), \( F = (f_1, f_2, \ldots, f_d)^T \) is \( d \times \hat{r} \), and \( F \) and \( G \) solve the least-squre problem:

\[
(F, G) = \arg \min_{F \in \mathcal{M}_{d \times \hat{r}}, G \in \mathcal{M}_{\hat{r} \times n}} \| \mathcal{Y} - F G \|_F^2
\]

subject to

\[ d^{-1} F^T F = I_{\hat{r}}, \quad G G^T \text{ is an } \hat{r} \times \hat{r} \text{ diagonal matrix.} \]
Our theory is based on the dual in-fill and diverging dimensionality asymptotics with the number of factors being finite. That is, $\Delta_n \to 0$, $d \to \infty$, and $r$ is fixed but unknown.
We first establish the consistency of $\hat{r}$.

**Theorem** Suppose Assumptions 1, 2, 3 with $a = 1$, and 4 hold. Suppose that $\Delta_n \log d \to 0$, $g(n, d) \to 0$, and $g(n, d) \left( (\Delta_n \log d)^{1/2} + d^{-1} m_d \right)^{-1} \to \infty$, we have $\mathbb{P}(\hat{r} = r) \to 1$. 
**Theorem** Suppose Assumptions 1, 2, 3 with \( a = 1/2, 4 \) and 5 hold. Suppose that \( \Delta_n \log d \to 0 \). Suppose further that \( \hat{r} \to r \) with probability approaching 1, then we have

\[
\left\| \hat{\Gamma}^S - \Gamma \right\|_{\text{MAX}} = O_p \left( (\Delta_n \log d)^{1/2} + d^{-1/2} m_d \right).
\]

Moreover, we have

\[
\left\| \hat{\Sigma}^S - \Sigma \right\|_{\text{MAX}} = O_p \left( (\Delta_n \log d)^{1/2} + d^{-1/2} m_d \right).
\]
Inverse of the Covariance Matrix

**Theorem** Suppose Assumptions 1 - 5 hold. Suppose 
\(d^{-1/2}m_d = o(1), \Delta_n \log d = o(1),\) and \(\hat{r} \to r\) with probability approaching 1, then we have

\[
\left\| \hat{\Gamma}^S - \Gamma \right\| = O_p \left( m_d (\Delta_n \log d)^{1/2} + d^{-1/2} m_d^2 \right).
\]

If in addition, \(d^{-1/2} m_d^2 = o(1)\) and \(m_d (\Delta_n \log d)^{1/2} = o(1),\) then \(\lambda_{\text{min}}(\hat{\Sigma}^S)\) is bounded away from 0 with probability approaching 1, and

\[
\left\| (\hat{\Sigma}^S)^{-1} - \Sigma^{-1} \right\| = O \left( m_d^3 \left( (\Delta_n \log d)^{1/2} + d^{-1/2} m_d \right) \right).
\]
**Theorem** Suppose Assumptions 1-4 hold. Suppose $d^{-1/2} m_d = o(1)$, $\Delta_n \log d = o(1)$, and $\hat{r} \rightarrow r$ with probability approaching 1, then there exists a $r \times r$ matrix $H$, such that with probability approaching 1, $H$ is invertible,

$$\|HH^T - I_r\| = \|H^TH - I_r\| = o_p(1),$$

and more importantly,

$$\|F - \beta H\|_{\text{MAX}} = O_p \left( (\Delta_n \log d)^{1/2} + d^{-1/2} \frac{m_d}{d^{1/2}} \right),$$

and

$$\|G - H^{-1} \chi\| = O_p \left( (\Delta_n \log d)^{1/2} + d^{-1/2} \frac{m_d}{d^{1/2}} \right).$$
Monte Carlo Simulations

- We sample 100 paths from a continuous-time $r$-factor model of $d$ assets specified as:

$$
\begin{align*}
  dY_{i,t} &= \sum_{j=1}^{r} \beta_{i,j} dX_{j,t} + dZ_{i,t}, \\
  dX_{j,t} &= b_{j}dt + \sigma_{j,t}d\tilde{W}_{j,t}, \\
  dZ_{i,t} &= \gamma_{i}^{T}dB_{i,t},
\end{align*}
$$

We allow for time-varying $\sigma_{j,t}$ which evolves according to the following system of equations:

$$
    d\sigma_{j,t}^{2} = \kappa_{j}(\theta_{j} - \sigma_{j,t}^{2})dt + \eta_{j}\sigma_{j,t}d\tilde{W}_{j,t}, \quad j = 1, 2, \ldots, r,
$$

where $\tilde{W}_{j}$ is a standard Brownian motion with

$$
\mathbb{E}[dW_{j,t}d\tilde{W}_{j,t}] = \rho_{j}dt.
$$
\( d = 500 \) and \( r = 3 \).

In the cross-section, we sample \( \beta_1 \sim U[0.25, 1.75] \), and sample \( \beta_2, \beta_3 \sim N(0, 0.5^2) \).

The variances on the diagonal of \( \Gamma \) are uniformly generated from \([0.05, 0.20]\), with constant within-block correlations from \( U[0.10, 0.50] \) for each block. In total, there are 20 blocks (of random sizes) on the diagonal of the residual covariance matrix.

We add a Gaussian noise with mean zero and variance \( 0.001^2 \) to the simulated log prices before censoring.

The data are then censored using Poisson sampling, where the number of observations for each asset is drawn from a truncated log-normal distribution.
<table>
<thead>
<tr>
<th>MAX</th>
<th>Freq</th>
<th>(|\hat{\Sigma}^5 - \Sigma|_{\text{MAX}})</th>
<th>(|\hat{\Sigma}^5 - \Sigma|_{\text{PCA}})</th>
<th>(|\Sigma^{-1} - \hat{\Sigma}^{-1}|_{\text{REG}})</th>
<th>(|\Sigma^{-1} - \hat{\Sigma}^{-1}|_{\text{PCA}})</th>
<th># Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.004</td>
<td>0.010</td>
<td>2.063</td>
<td>2.063</td>
<td>0.51</td>
<td>1.42</td>
</tr>
<tr>
<td>15</td>
<td>0.007</td>
<td>0.011</td>
<td>0.785</td>
<td>0.785</td>
<td>0.88</td>
<td>1.55</td>
</tr>
<tr>
<td>30</td>
<td>0.009</td>
<td>0.012</td>
<td>0.404</td>
<td>0.404</td>
<td>1.23</td>
<td>1.72</td>
</tr>
<tr>
<td>60</td>
<td>0.014</td>
<td>0.015</td>
<td>0.229</td>
<td>0.217</td>
<td>1.81</td>
<td>2.03</td>
</tr>
<tr>
<td>15</td>
<td>300</td>
<td>0.031</td>
<td>0.032</td>
<td>0.137</td>
<td>0.112</td>
<td>4.30</td>
</tr>
<tr>
<td>300</td>
<td>900</td>
<td>0.054</td>
<td>0.054</td>
<td>0.078</td>
<td>0.070</td>
<td>8.33</td>
</tr>
<tr>
<td>15</td>
<td>1800</td>
<td>0.078</td>
<td>0.079</td>
<td>0.084</td>
<td>0.084</td>
<td>13.44</td>
</tr>
<tr>
<td>1800</td>
<td>3900</td>
<td>0.112</td>
<td>0.113</td>
<td>0.116</td>
<td>0.116</td>
<td>26.22</td>
</tr>
<tr>
<td>3900</td>
<td>4680</td>
<td>0.124</td>
<td>0.124</td>
<td>0.124</td>
<td>0.125</td>
<td>31.01</td>
</tr>
<tr>
<td>4680</td>
<td>11700</td>
<td>0.195</td>
<td>0.196</td>
<td>0.196</td>
<td>0.196</td>
<td>116.23</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>0.004</td>
<td>0.012</td>
<td>2.063</td>
<td>2.063</td>
<td>0.63</td>
</tr>
<tr>
<td>15</td>
<td>0.007</td>
<td>0.013</td>
<td>0.785</td>
<td>0.785</td>
<td>1.10</td>
<td>2.51</td>
</tr>
<tr>
<td>30</td>
<td>0.010</td>
<td>0.014</td>
<td>0.403</td>
<td>0.403</td>
<td>1.57</td>
<td>2.66</td>
</tr>
<tr>
<td>60</td>
<td>0.014</td>
<td>0.017</td>
<td>0.247</td>
<td>0.222</td>
<td>2.21</td>
<td>2.89</td>
</tr>
<tr>
<td>25</td>
<td>300</td>
<td>0.032</td>
<td>0.033</td>
<td>0.148</td>
<td>0.118</td>
<td>5.46</td>
</tr>
<tr>
<td>300</td>
<td>900</td>
<td>0.055</td>
<td>0.055</td>
<td>0.080</td>
<td>0.071</td>
<td>10.90</td>
</tr>
<tr>
<td>15</td>
<td>1800</td>
<td>0.078</td>
<td>0.078</td>
<td>0.084</td>
<td>0.083</td>
<td>18.51</td>
</tr>
<tr>
<td>3900</td>
<td>1800</td>
<td>0.116</td>
<td>0.116</td>
<td>0.116</td>
<td>0.117</td>
<td>37.71</td>
</tr>
<tr>
<td>4680</td>
<td>3900</td>
<td>0.128</td>
<td>0.128</td>
<td>0.130</td>
<td>0.131</td>
<td>47.72</td>
</tr>
<tr>
<td>11700</td>
<td>0.199</td>
<td>0.200</td>
<td>0.200</td>
<td>0.201</td>
<td>315.69</td>
<td>314.05</td>
</tr>
</tbody>
</table>

Table: Simulation Results
Data

- We collect from the TAQ database intraday observations of the S&P 500 index constituents from January 2004 to December 2012.
- There are in total 736 stocks.
- We subsample returns of each asset every 15 minutes. The overnight returns are excluded to avoid dividend issuances and stock splits.
- We collect the Global Industrial Classification Standard (GICS) codes from the Compustat database.
  - The code is split into 4 groups of 2 digits
  - 1-2 describe the company’s sector; digits 3-4 describe the industry group; digits 5-6 describe the industry; digits 7-8 describe the sub-industry.
We also make use of the observable factors constructed at high-frequency, including:

- The market portfolio
- SMB, HML, MOM
- 9 industry SDPR ETFs, including Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Health Care (XLV), Financial (XLF), Information Technology (XLK), and Utilities (XLU).
The Number of Factors

1 Component

4 Components

10 Components

13 Components

Figure: Sparsity Pattern of the PCA Residual
Figure: Sparsity Pattern of the Regression Residual
Figure: Estimates of the Number of Factors
In-Sample $R^2$ Comparison

Figure: In-Sample $R^2$ Comparison
Out-of-Sample Portfolio Allocation Study

In this section, we bring together our covariance estimates to an empirical test-drive. We consider the following constrained portfolio allocation exercise:

\[
\min_{w} w^{T} \hat{\Sigma}^{S} w, \quad \text{subject to } \omega^{T}1 = 1, \|\omega\|_{1} \leq \gamma, \quad (9)
\]

where \(\|\omega\|_{1} \leq \gamma\) imposes an exposure constraint.

▶ When \(\gamma = 1\), the optimal portfolio allows no short-sales, i.e., all portfolio weights are non-negative.

▶ When \(\gamma\) is small and binding, the optimal portfolio is sparse, i.e., many weights are zero.

▶ When \(\gamma\) is no longer binding, the optimal portfolio coincides with the global minimum variance portfolio.
Figure: Out-of-Sample Risk of the Portfolio

For comparison, the annualized volatility of the equal-weight portfolio is 17.89%.
Figure: Out-of-Sample Risk of the Portfolio against Exposure Constraint and Number of Factors