High Dimensional Covariance Matrix Estimation Using a Factor Model

Jinchi Lv

Marshall School of Business
University of Southern California

Joint with Jianqing Fan and Yingying Fan

http://www-rcf.usc.edu/∼jinchilv

Radcliffe Workshop, November 10, 2007

Outline

1. Introduction
2. Factor-Model Based Estimation
3. Theoretical Studies & Applications
4. Simulation Studies

Covariance Matrix

Covariance matrix: fundamental & pervades financial econometrics.

- VaR;
- capital requirement & risk management;
- asset pricing;
- portfolio allocation;
- genetic networks & climatology.

Challenge of High Dimensionality

Estimating high-dimensional covariance matrices: challenging.

- 200 stocks;
- 20,200 parameters;
- 3-year daily returns, only about 750 samples.
- Estimating it accurately?!
- high-frequency data?

Dimensionality Reduction

Sample covariance matrix:

- problematic when p is large. (Johnstone 01.)

Dimensionality reduction:

- Factor models. (Engle&Watson, 81; Chamberlain&Rothschild, 83; Diebold&Nerlove, 89; Aguilar&West, 00; Stock&Watson, 05.)
- Sparsity & AR-models. (Bickel&Levina, 06; Pourahmadi, 00; Boik, 02; Wu&Pourahmadi, 03; Huang, Liu, Pourahmadi, 04; Li&Gui, 05.)
- Shrinkage & eigen-method. (Ledoit&Wolf, 04; Stein, 75; Eaton&Taylor, 91,94.)

Motivation: Multi-Factor Model

Multi-factor model: Ross (76) & Chamberlain, Rothschild (83).

Notation:

- $p = p_n$ & $K = K_n$;
- Y_i: excess return;
- $f_1, \cdots, f_K$: factors.

Multi-factor model:

$$Y_i = b_{n,i1} f_1 + \cdots + b_{n,iK} f_K + \varepsilon_i, \ i = 1, \cdots, p.$$

- $\{\varepsilon_i\}$: idiosyncratic, uncorrelated given **f**;
- varies across n.

An Example: Fama-French 3-Factor Model

Fama-French 3-factor model:

- f_1: market portfolio;
- f_2: capitalization,

  $f_2 = 1/3(SV + SN + SG) - 1/3(BV + BN + BG)$;
- f_3: book-to-market ratio,

  $f_3 = 1/2(SV + BV) - 1/3(SG + BG)$.

Model-Based Estimation: A Substitution Estimator

Multi-period

$$\mathbf{y}_t = \mathbf{B}_n \mathbf{f}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \cdots, n.$$

Covariance structure: $\boldsymbol{\Sigma} = \mathbf{B}\mathrm{cov}(\mathbf{f})\mathbf{B}' + \boldsymbol{\Sigma}_0$,

- $\boldsymbol{\Sigma}_0$: diagonal;
- p_n and K_n: growing.

Estimated covariance:

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\mathbf{f})\widehat{\mathbf{B}}' + \widehat{\boldsymbol{\Sigma}}_0.$$

Sample covariance matrix: $\widehat{\boldsymbol{\Sigma}}_{\mathrm{sam}}$.

Questions and Objectives

- Estimation error growing with p_n and K_n?
- Impacts on portfolio allocation & risk management?
- Comparison with the sample covariance?
- When does factor approach gain substantially/marginally?

Choice of Norms

Frobenius norm: not appropriate, e.g. knowing ideally $\mathbf{B} = \mathbf{1}$ and $\mathrm{cov}(\varepsilon) = I_{p_n} \quad \Longrightarrow \quad \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| = p_n |\widehat{\mathrm{var}}(f) - \mathrm{var}(f)|$.

New norm: $\|\mathbf{A}\|_{\mathbf{\Sigma}_n} = p_n^{-1/2} \|\mathbf{\Sigma}_n^{-1/2} \mathbf{A} \mathbf{\Sigma}_n^{-1/2}\|$,

- factor structure & diverging p_n;
- $p_n^{1/2} \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathbf{\Sigma}} = \{\mathrm{tr}[\widehat{\mathbf{\Sigma}} \mathbf{\Sigma}^{-1} - I_{p_n}]^2\}^{1/2}$;
- entropy loss: $\quad \mathrm{tr}(\widehat{\mathbf{\Sigma}} \mathbf{\Sigma}^{-1}) - \log|\widehat{\mathbf{\Sigma}} \mathbf{\Sigma}^{-1}| - p$.

A Surprising Fact

$\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{sam}}$: same rate $O_P(n^{-1/2} p_n K_n)$ under Frobenius norm.

- explicit;
- K_n: constant or slowly growing;

- Factor model does not help on estimating $\boldsymbol{\Sigma}$;
- Same rate in risk management:
  $\xi_n' \boldsymbol{\Sigma}_n \xi_n$, variance of portfolio ξ_n.

Strength of Factor Structure I

Summary:

- $\widehat{\mathbf{\Sigma}}$: invertible;
- Faster rate under norm $\|\cdot\|_{\mathbf{\Sigma}}$ when $K_n = o(\sqrt{p_n})$;
  $K_n = O(1)$: $\widehat{\mathbf{\Sigma}}$ is root-n-consistent when $p_n = O(n)$, whereas $\mathbf{\Sigma}_{\text{sam}}$ is root-n/p_n-consistent;
- Under Frobenius norm, $\widehat{\mathbf{\Sigma}}^{-1}$ has a rate an order p_n/K_n faster than that of $\widehat{\mathbf{\Sigma}}_{\text{sam}}^{-1}$.

Mean-Variance Optimal Portfolio

Mean-Variance Optimal Portfolio (Markowitz, 1952):

$$\min_{\xi \in \mathbf{R}^{p_n}} \xi' \boldsymbol{\Sigma}_n \xi \quad \text{s.t. } \xi' \mathbf{1} = 1 \text{ and } \xi' \boldsymbol{\mu}_n = \gamma_n.$$

- γ_n: expected rate of return;
- closed-form solution, involving $\boldsymbol{\Sigma}_n^{-1}$.

Questions:

- Impact on portfolio allocation?
- Performance of $\boldsymbol{\Sigma}_{\text{sam}}$?

Strength of Factor Structure II

Summary:

- Optimal portfolio: an order p_n/K_n faster;
- Minimum-variance portfolio: same result.

Simulation: Fit Fama-French 3-Factor Model

Fama-French 3-factor model:

- 30 industry portfolios, 5/1/02-8/29/05 ($n = 756$);
- 30 estimated factor loading vectors:

$\mu_{\mathbf{f}}$	$\mathrm{cov}_{\mathbf{f}}$		
0.023558	1.2507	-0.034999	-0.20419
0.012989	-0.034999	0.31564	-0.0022526
0.020714	-0.20419	-0.0022526	0.19303
$\mu_{\mathbf{b}}$	$\mathrm{cov}_{\mathbf{b}}$		
0.78282	0.029145	0.023873	0.010184
0.51803	0.023873	0.053951	-0.006967
0.41003	0.010184	-0.006967	0.086856

- SDs of 30 idiosyn. errors: ave. 0.6608, SD 0.3275 & min 0.1950.

Simulation Design

n & K fixed and p growing:

- Generate $\mathbf{f}$ from $\mathcal{N}\left(\mu_{\mathbf{f}}, \mathrm{cov}_{\mathbf{f}}\right)$, $n = 756$;
- $p \in [16, 1000]$, increment 20;
- Generate $\mathbf{b}_1, \cdots, \mathbf{b}_p$ from $\mathcal{N}\left(\mu_{\mathbf{b}}, \mathrm{cov}_{\mathbf{b}}\right)$;
- Generate $\sigma_1, \cdots, \sigma_p$ from a gamma distribution $G(3.3586, 0.1876)$ conditioned on $[0.1950, \infty)$;
- Generate idiosyn. noise from $\mathcal{N}\left(0, \sigma_i^2\right)$;
- Get pseudo excess returns using $\mathbf{y} = \mathbf{B}\mathbf{f} + \varepsilon$.

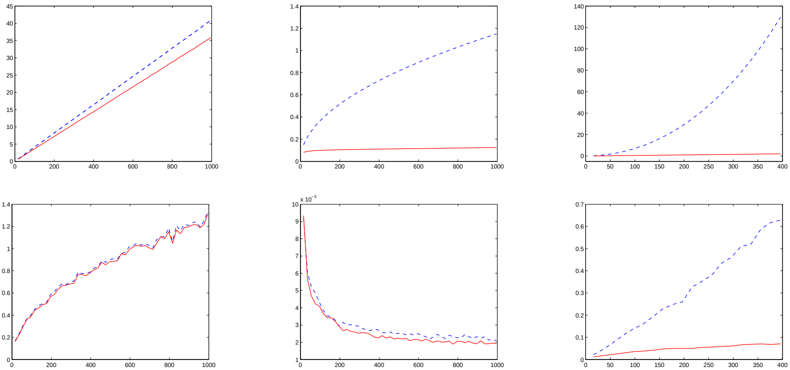
Comparison of Performance

Comparison of $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_{\text{sam}}$ under different measures:

Frobenius norm | norm $\|\cdot\|_{\boldsymbol{\Sigma}}$ | Entropy loss

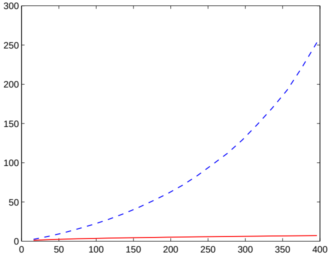
mean over 500 simu.

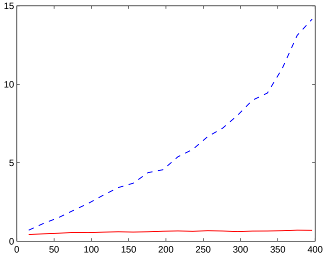
std



Estimation of $\boldsymbol{\Sigma}^{-1}$ under Frobenius Norm

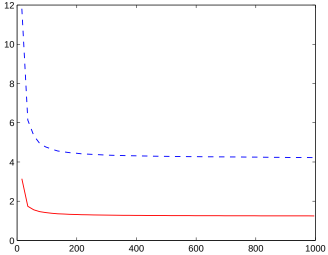
Average and standard deviation of errors under the Frobenius norm over 500 simulations for $\widehat{\boldsymbol{\Sigma}}^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{sam}}^{-1}$ against dimensionality.

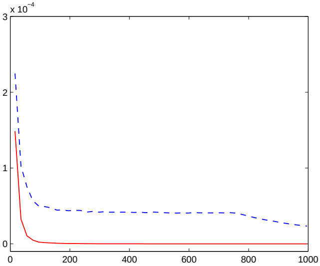




Impact on Portfolio Allocation

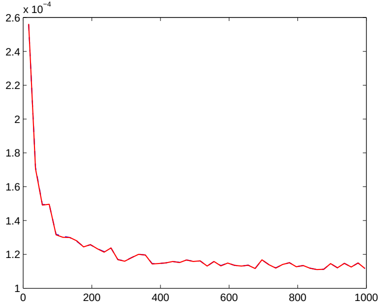
Left: MSEs of estimated variances of optimal portfolios, $\gamma_n = 10\%$;
Right: MSEs of estimated minimum variances over 500 simulations.





Impact on Risk Management

MSEs of estimated variances of the equally-weighted portfolio over 500 simulations.



Convergence Rates

Theorem 1. Under some regularity conditions and the Frobenius norm, we have $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = O_P(n^{-1/2} p_n K_n)$ and

$$\max_{1 \le k \le p_n} |\lambda_k(\widehat{\boldsymbol{\Sigma}}) - \lambda_k(\boldsymbol{\Sigma})| = o_P\{(p_n^2 K_n^2 \log n / n)^{1/2}\};$$

$\widehat{\boldsymbol{\Sigma}}_{\text{sam}}$ has the same rates.

Theorem 2. If $p_n = n^{\alpha}$ and $K_n = n^{\alpha_1}$, then $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\boldsymbol{\Sigma}} = O_P(n^{-\beta/2})$ with $\beta = \min(1 - 2\alpha_1, 2 - \alpha - \alpha_1)$, whereas $\widehat{\boldsymbol{\Sigma}}_{\text{sam}}$ has rate $O_P(n^{-\beta_1/2})$ with $\beta_1 = 1 - \max(\alpha, 3\alpha_1/2, 3\alpha_1 - \alpha)$.

Theorem 3. Under the Frobenius norm, $\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\| = o_P\{(p_n^2 K_n^4 \log n / n)^{1/2}\}$, an order p_n/K_n smaller than $\|\widehat{\boldsymbol{\Sigma}}_{\text{sam}}^{-1} - \boldsymbol{\Sigma}^{-1}\|$.

Asymptotic Normality

Theorem 4. Asymptotic normality of $\widehat{\boldsymbol{\Sigma}}$ has been derived to facilitate statistical inferences, whereas in general $\widehat{\boldsymbol{\Sigma}}_{\text{sam}}$ may have no asymptotic normality of the same kind when $p_n \to \infty$.

Impacts on Portfolio Management

Theorem 5 (Optimal portfolio).

$$\left|\widehat{\xi}_n' \widehat{\boldsymbol{\Sigma}}_n \widehat{\xi}_n - \xi_n' \boldsymbol{\Sigma}_n \xi_n\right| = o_P\{(p_n^4 K_n^4 \log n/n)^{1/2}\},$$

whereas the rate using $\boldsymbol{\Sigma}_{\text{sam}}$ is an order p_n/K_n worse.

Theorem 6 (Minimum-variance portfolio).

$$\left|\widehat{\xi}_{ng}' \widehat{\boldsymbol{\Sigma}}_n \widehat{\xi}_{ng} - \xi_{ng}' \boldsymbol{\Sigma}_n \xi_{ng}\right| = o_P\{(p_n^4 K_n^4 \log n/n)^{1/2}\},$$

whereas the rate using $\boldsymbol{\Sigma}_{\text{sam}}$ is an order p_n/K_n worse.

Theorem 7. Given a portfolio ξ_n with $\xi_n' \mathbf{1} = 1$ and $\xi_n = O(1)\mathbf{1}$, we have

$$\left|\xi_n' \widehat{\boldsymbol{\Sigma}}_n \xi_n - \xi_n' \boldsymbol{\Sigma}_n \xi_n\right| = o_P\{(p_n^4 K_n^2 \log n/n)^{1/2}\};$$

$|\xi_n' \widehat{\boldsymbol{\Sigma}}_{\text{sam}} \xi_n - \xi_n' \boldsymbol{\Sigma}_n \xi_n|$ has the same rate. Moreover, if no short position, rate is $o_P\{(p_n^2 K_n^2 \log n/n)^{1/2}\}$.

Conclusions

- We propose and study the use of the factor model to estimate high-dimensional covariance matrix.
- When dimensionality is high, the factor-model based estimator
  - significantly outperforms the sample covariance particularly in estimating the inverse;
  - significantly outperforms the sample covariance in portfolio allocation;
  - does not improve the performance of risk management.