Journal of Econometrics 147 (2008) 186–197

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

High dimensional covariance matrix estimation using a factor model[☆]

Jianqing Fan[a], Yingying Fan[b], Jinchi Lv[b,*]

[a] Department of Operations Research and Financial Engineering, Princeton University, United States

[b] Information and Operations Management Department, Marshall School of Business, University of Southern California, United States

ARTICLE INFO

Article history:
Available online 18 September 2008

JEL classification:
C13
C51

Keywords:
Factor model
Diverging dimensionality
Covariance matrix estimation
Asymptotic properties
Portfolio management

ABSTRACT

High dimensionality comparable to sample size is common in many statistical problems. We examine covariance matrix estimation in the asymptotic framework that the dimensionality p tends to ∞ as the sample size n increases. Motivated by the Arbitrage Pricing Theory in finance, a multi-factor model is employed to reduce dimensionality and to estimate the covariance matrix. The factors are observable and the number of factors K is allowed to grow with p. We investigate the impact of p and K on the performance of the model-based covariance matrix estimator. Under mild assumptions, we have established convergence rates and asymptotic normality of the model-based estimator. Its performance is compared with that of the sample covariance matrix. We identify situations under which the factor approach increases performance substantially or marginally. The impacts of covariance matrix estimation on optimal portfolio allocation and portfolio risk assessment are studied. The asymptotic results are supported by a thorough simulation study.

1. Introduction

Covariance matrix estimation is fundamental for almost all areas of multivariate analysis and many other applied problems. In particular, covariance matrices and their inverses play a central role in portfolio risk assessment and optimal portfolio allocation. For example, the smallest and largest eigenvalues of a covariance matrix are related to the minimum and maximum variances of the selected portfolio, respectively, and the eigenvectors are related to optimal portfolio allocation. Therefore, we need a good covariance matrix estimator inverting which does not excessively amplify the estimation error. See Goldfarb and Iyengar (2003) for applications of covariance matrices to portfolio selections and Johnstone (2001) for their statistical implications.

Estimating high-dimensional covariance matrices is intrinsically challenging. For example, in optimal portfolio allocation and portfolio risk assessment, the number of stocks p, which is typically of the same order as the sample size n, can well be in the order of hundreds. In particular, when $p = 200$ there are more than 20,000 parameters in the covariance matrix. Yet, the available sample size is usually in the order of hundreds or a few thousand because longer time series (larger n) increases modeling bias. For instance, by taking daily data of the past three years we have only roughly $n = 750$. So it is hard or even unrealistic to estimate covariance matrices without imposing any structure (see the rejoinder in Fan (2005)).

Factor models have been widely used both theoretically and empirically in economics and finance. Derived by Ross (1976, 1977) using the Arbitrage Pricing Theory (APT) and by Chamberlain and Rothschild (1983) in a large economy, the multi-factor model states that the excessive return of any asset Y_i over the risk-free interest rate satisfies

$$Y_i = b_{i1}f_1 + \cdots + b_{iK}f_K + \varepsilon_i, \quad i = 1, \ldots, p, \tag{1}$$

where $f_1, \ldots, f_K$ are the excessive returns of K factors, b_{ij}, $i = 1, \ldots, p, j = 1, \ldots, K$, are unknown factor loadings, and $\varepsilon_1, \ldots, \varepsilon_p$ are p idiosyncratic errors uncorrelated given $f_1, \ldots, f_K$. The factor models have been widely applied and studied in economics and finance. See, for example, Ross (1976, 1977), Engle and Watson (1981), Chamberlain (1983), Chamberlain and Rothschild (1983), Diebold and Nerlove (1989), Fama and French (1992, 1993), Aguilar and West (2000), Bai (2003), Ledoit and Wolf (2003), Stock and Watson (2005) and references therein. These are extensions of the famous Capital Asset Pricing Model (CAPM) and can be regarded as efforts to approximate the market portfolio in the CAPM.

Thanks to the multi-factor model (1), if a few factors can completely capture the cross-sectional risks, the number of

[☆] Financial support from the NSF under grant DMS-0532370 is gratefully acknowledged. Lv's research was partially supported by National Science Foundation grant DMS-0806030 and the 2008 Zumberge Individual Award from the James H. Zumberge Faculty Research and Innovation Fund at the University of Southern California. We are grateful to the guest editor, Jiti Gao and two anonymous referees for their constructive and helpful comments.

[*] Corresponding address: Information and Operations Management Department, Marshall School of Business, University of Southern California, 90089 Los Angeles, CA, United States. Tel.: +1 213 740 6603.

E-mail address: jinchilv@marshall.usc.edu (J. Lv).

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.09.017

parameters in covariance matrix estimation can be significantly reduced. For example, using the Fama-French three-factor model (Fama and French, 1992, 1993), there are $4p$ instead of $p(p+1)/2$ parameters to be estimated. Despite the popularity of factor models in the literature, the impact of dimensionality on the estimation errors of covariance matrices and its applications to optimal portfolio allocation and portfolio risk assessment are poorly understood so, in this paper, determined efforts are made on such an investigation. To make the multi-factor model more realistic, we allow K to grow with the number of assets p and hence with the sample size n. As a result, we also investigate the impact of the number of factors on the estimation of covariance matrices, as well as its applications to optimal portfolio allocation and portfolio risk assessment. To appreciate the derived rates of convergence, we compare them with those without using the factor structure. One natural candidate is the sample covariance matrix. This also allows us to examine the impact of dimensionality on the performance of the sample covariance matrix. We will assume that the factors are observable as in Fama and French (1992, 1993). Our results also provide an important milestone for understanding the performance of factor models with unobservable factors.

The traditional covariance matrix estimator, the sample covariance matrix, is known to be unbiased, and it is invertible when the dimensionality is no larger than the sample size. See, for example, Eaton and Tyler (1994) for the asymptotic spectral distributions of random matrices including sample covariance matrices and their statistical implications. In the absence of prior information about the population covariance matrix, the sample covariance matrix is certainly a natural candidate in the case of small dimensionality, but no longer performs very well for moderate or large dimensionality [see, e.g. Lin and Perlman (1985) and Johnstone (2001)]. Many approaches were proposed in the literature to construct good covariance matrix estimators. Among them, two main directions were taken. One is to remedy the sample covariance matrix and construct a better one by using approaches such as shrinkage and the eigen-method, etc. See, for example, Ledoit and Wolf (2004) and Stein (1975). The other one is to reduce dimensionality by imposing some structure on the data. Many structures, such as sparsity, compound symmetry, and the autoregressive model, are widely used. Various approaches were taken to seek a balance between the bias and variance of covariance matrix estimators. See, for example, Wong et al. (2003), Huang et al. (2006), and Bickel and Levina (2008).

The rest of the paper is organized as follows. Section 2 introduces the estimators of the covariance matrix. In Section 3 we give some basic assumptions and present sampling properties of the estimators. We study the impacts of covariance matrix estimation on optimal portfolio allocation and portfolio risk assessment in Section 4. A simulation study is presented in Section 5, which augments our theoretical study. Section 6 contains some concluding remarks. The proofs of our results are given in the Appendix.

2. Covariance matrix estimation

We always denote by n the sample size, by p the dimensionality, and by $f_1, \ldots, f_K$ the K observable factors, where p grows with sample size n and K increases with dimensionality p. For ease of presentation, we rewrite the factor model (1) in matrix form

$$\mathbf{y} = \mathbf{B}_n \mathbf{f} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{y} = (Y_1, \ldots, Y_p)'$, $\mathbf{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ with $\mathbf{b}_i = (b_{n,i1}, \ldots, b_{n,iK})'$, $i = 1, \ldots, p$, $\mathbf{f} = (f_1, \ldots, f_K)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$. Throughout we assume that $E(\boldsymbol{\varepsilon}|\mathbf{f}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}|\mathbf{f}) = \boldsymbol{\Sigma}_{n,0}$ is diagonal. For brevity of notation, we suppress the first subscript n in some situations where the dependence on n is self-evident.

Let $(\mathbf{f}_1, \mathbf{y}_1), \ldots, (\mathbf{f}_n, \mathbf{y}_n)$ be n independent and identically distributed (i.i.d.) samples of $(\mathbf{f}, \mathbf{y})$. We introduce here some notation used throughout the paper. Let

$$\boldsymbol{\Sigma}_n = \mathrm{cov}(\mathbf{y}), \quad \mathbf{X} = (\mathbf{f}_1, \ldots, \mathbf{f}_n),$$
$$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \quad \text{and} \quad \mathbf{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n).$$

Under model (2), we have

$$\boldsymbol{\Sigma}_n = \mathrm{cov}(\mathbf{B}_n\mathbf{f}) + \mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{B}_n \mathrm{cov}(\mathbf{f}) \mathbf{B}_n' + \boldsymbol{\Sigma}_{n,0}. \tag{3}$$

A natural idea for estimating $\boldsymbol{\Sigma}_n$ is to plug in the least-squares estimators of $\mathbf{B}_n$, $\mathrm{cov}(\mathbf{f})$, and $\boldsymbol{\Sigma}_{n,0}$. Therefore, we have a substitution estimator

$$\hat{\boldsymbol{\Sigma}}_n = \widehat{\mathbf{B}}_n \widehat{\mathrm{cov}}(\mathbf{f}) \widehat{\mathbf{B}}_n' + \hat{\boldsymbol{\Sigma}}_{n,0}, \tag{4}$$

where $\widehat{\mathbf{B}}_n = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$ is the matrix of estimated regression coefficients, $\widehat{\mathrm{cov}}(\mathbf{f}) = (n-1)^{-1}\mathbf{X}\mathbf{X}' - \{n(n-1)\}^{-1}\mathbf{X}\mathbf{1}\mathbf{1}'\mathbf{X}'$ is the sample covariance matrix of the factors $\mathbf{f}$, and

$$\hat{\boldsymbol{\Sigma}}_{n,0} = \mathrm{diag}\left(n^{-1}\widehat{\mathbf{E}}\widehat{\mathbf{E}}'\right)$$

is the diagonal matrix of $n^{-1}\widehat{\mathbf{E}}\widehat{\mathbf{E}}'$ with $\widehat{\mathbf{E}} = \mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X}$ the matrix of residuals. If the factor model is not employed, then we have the sample covariance matrix estimator

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{sam}} = (n-1)^{-1}\mathbf{Y}\mathbf{Y}' - \{n(n-1)\}^{-1}\mathbf{Y}\mathbf{1}\mathbf{1}'\mathbf{Y}'. \tag{5}$$

Ledoit and Wolf (2003) propose an interesting idea of combining the single-index ($K = 1$, CAPM) model based estimation of the covariance matrix with the sample covariance matrix to improve the estimate of the covariance matrix. It aims at a trade-off between the bias and variance of the two estimated covariance matrices for practical applications.

In the paper we mainly aim to provide a theoretical understanding of the factor model with a diverging dimensionality and growing number of factors for the purpose of covariance matrix estimation, but not to compare with other popular estimators. Throughout the paper, we always contrast the performance of the covariance matrix estimator $\hat{\boldsymbol{\Sigma}}$ in (4) with that of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathrm{sam}}$ in (5). The paper also provides a theoretical study on the two estimators used in the procedure of Ledoit and Wolf (2003). With prior information of the true factor structure, the substitution estimator $\hat{\boldsymbol{\Sigma}}$ is expected to perform better than $\hat{\boldsymbol{\Sigma}}_{\mathrm{sam}}$. However, this has not been formally shown, especially when $p \to \infty$ and $K \to \infty$, and this is not always true. In addition, exact properties of this kind are not well understood. As the problem is important for portfolio management, determined efforts are devoted in this regard.

3. Sampling properties

In this section we study the sampling properties of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{sam}}$ with growing dimensionality and number of factors. We give some basic assumptions in Section 3.1. The sampling properties are presented in Section 3.2.

In the presence of diverging dimensionality, one should carefully choose appropriate norms for large matrices in different situations. We first introduce some notation. We always denote by $\lambda_1(\mathbf{A}), \ldots, \lambda_q(\mathbf{A})$ the q eigenvalues of a $q \times q$ symmetric matrix $\mathbf{A}$ in decreasing order. For any matrix $\mathbf{A} = (a_{ij})$, its Frobenius norm is given by

$$\|\mathbf{A}\| = \left\{\mathrm{tr}(\mathbf{A}\mathbf{A}')\right\}^{1/2}. \tag{6}$$