Large Covariance Matrix Estimation in Approximate Factor Models

Yuan Liao *

Department of Mathematics, University of Maryland

Keywords: Panel Data, **Financial Econometrics**, high dimensionality, covariance matrix, sparse estimation, thresholding, cross-sectional correlation, unobservable factors

Abstract

Due to the abundance of high dimensional data in modern econometric applications, the estimation of a large covariance matrix for panel data has become an important question. We consider the following factor model:

$$y_{it} = b_i' f_t + u_{it}, \quad i \leq N, t \leq T$$

where f_t is a fixed dimension vector of common factors, which may or may not be observable; b_i is the factor loading vector, and u_{it} is the idiosyncratic component. Depending on the application problem, y_{it} might be the only observable in the model. The goal is to estimate the $N \times N$ large covariance $\Sigma_y = \text{cov}(y_t)$ and $\Sigma_u = \text{cov}(u_t)$. Modern applications usually require N be probably larger than T.

When $N > T$ the classical sample covariance estimator is well known to be degenerate. Hence most of the literature has focused on the cross sectional independence among the idiosyncratic, in the sense that $\{u_{it}\}$ are independent across i, which leads to a diagonal covariance Σ_u. As a result, classical approaches could only handle either $N < T$ case or the cross sectional independence. However, the cross sectional independence has been rejected by many testing procedures in empirical studies on either financial or economic data. Also, it rules out the approximate factor model (Chamberlain and Rothschild 1983).

This paper answers the question on how to estimate the large covariances when $N > T$, allowing the presence of cross sectional dependence, that is, after the common factors are taken out, the idiosyncratic components might still be correlated (not even weakly). The basic assumption is that Σ_u is a sparse covariance, with many

yliao123@umd.edu, Mathematics Building University of Maryland, College Park, MD 20742-4015

off-diagonal components being quite small. But still, there can be $o(\min\{\sqrt{N}, \sqrt{T}\})$ large entries in each row of Σ_u. Therefore, this approach enables us to combine the merits of the methods based on the sparsity and the approximate factor model. We estimate Σ_u using the adaptive thresholding technique, taking into account the fact that direct observations of the idiosyncratic components and the common factors are both unavailable. It is shown that the resulting covariance estimators are positive definite with probability approaching one, and

$$\|\hat{\Sigma}_u - \Sigma_u\|_o = O_p(c_{N,T} m_N) = \|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_o$$

$$\|\hat{\Sigma}_y^{-1} - \Sigma_y^{-1}\|_o = O_p(c_{N,T} m_N),$$

with $\|.\|_o$ denoting the operator norm. Here m_N is the maximum number of "big" entries in each row of Σ_u, assumed to be $o(c_{N,T}^{-1})$. When the common factors are observable,

$$c_{N,T} = \sqrt{\frac{\log N}{T}};$$

when the common factors are not observable,

$$c_{N,T} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}.$$

In particular, the rate is optimal (achieves the minimax rate uniformly in a class of sparse covariance) when $N \log N \gg T$. In addition, when the common factors are not observable, the effect of estimating the factors vanishes when $N \log N \gg T$, and we can treat the factors as if they are observable. Thus the high dimensionality helps in estimating the unknown factors, a claim by Bai (2003).

The asymptotic results are also verified by extensive simulation studies.

This work has many important applications in econometrics. Two examples will be presented.

Example 1: Optimal portfolio allocation Markowitz (1952) defines the mean variance optimal portfolio of N risky assets with expected returns μ and covariance matrix Σ_y as the solution of the allocation vector ξ to the following minimization problem

$$\min_{\xi \in \mathbb{R}^N} \xi' \Sigma_y \xi, \quad \text{s.t. } \sum_{i=1}^{N} \xi_i = 1, \xi' \mu = \gamma$$

where γ is the expected return of the portfolio. Let e denote a N dimensional vector of ones. The optimal solution is given by

$$\xi^* = c_1 \Sigma_y^{-1} e + c_2 \Sigma_y^{-1} \mu$$

for some constants c_1 and c_2 that depend on Σ_y^{-1}, μ and γ (see, e.g. Cochrane (2001), Campbell et al. (1997)). When $N > T$, estimating Σ_y^{-1} is essential for estimating the optimal portfolio. By assuming a factor structure on Σ_y in the Fama French three factor model, we can estimate Σ_y^{-1} with a rate $c_{N,T}m_N$ as introduced before under the operator norm, which is good enough to consistently estimate ξ^*.

Example 2: Testing high dimensional CAPM We test the mean variance efficiency based on a multi-factor model

$$y_{it} = \alpha_i + b_i' f_t + u_{it},$$

which is an extension of the classical CAPM by Sharpe (1964). The null hypothesis of mean variance efficiency is given by

$$H_0 : \alpha_i = 0, i \leq N.$$

The classical Gibbons Ross and Shanken (1989) test is based on

$$c_f \hat{\alpha}' \Sigma_u^{-1} \hat{\alpha},$$

where c_f is a constant that depends on the common factors, $\hat{\alpha}$ is the OLS estimator of α when the factors are observable. A feasible test should replace Σ_u^{-1} by a consistent estimator, which is challenging when $N > T$. Hence the literature (e.g., Sentana 2009) focuses on $N < T$ case, and typical choices are $T = 60$ monthly data and the number of assets $N = 10$ or 25. However, the CAPM should hold for all tradeable assets, not just a small fraction of them. By assuming Σ_u to be sparse, that is, $m_N = o(c_{N,T}^{-1})$, our result shows that

$$\|\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_o = o_p(1),$$

which can produce an operational test statistic even when $N > T$.

NOTE:

1. This is a joint work with Jianqing Fan and Martina Mincheva
2. The main part of this talk is based on the materials of two papers, which are downloadable from the author's homepage, entitled:

   "High Dimensional Covariance Matrix Estimation in Approximate Factor Models"

   "Large Covariance Estimation by Thresholding Principal Orthogonal Complements"