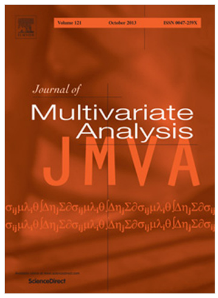

Estimation of two high-dimensional covariance matrices and the spectrum of their ratio

Jun Wen

PII: S0047-259X(17)30545-6
DOI: https://doi.org/10.1016/j.jmva.2018.06.008
Reference: YJMVA 4376

To appear in: *Journal of Multivariate Analysis*

Received date : 8 September 2017

Estimation of two high-dimensional covariance matrices and the spectrum of their ratio

Jun Wen[a]

[a] *Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546*

Abstract

Let $S_{p,1}$, $S_{p,2}$ be two independent $p \times p$ sample covariance matrices with degrees of freedom n_1 and n_2, respectively, whose corresponding population covariance matrices are $\Sigma_{p,1}$ and $\Sigma_{p,2}$, respectively. Knowing $S_{p,1}$, $S_{p,2}$, this article proposes a class of estimators for the spectrum (eigenvalues) of the matrix $\Sigma_{p,2}\Sigma_{p,1}^{-1}$ as well as the pair of the whole matrices $(\Sigma_{p,1}, \Sigma_{p,2})$. The estimators are created based on Random Matrix Theory. Under mild conditions, our estimator for the spectrum of $\Sigma_{p,2}\Sigma_{p,1}^{-1}$ is shown to be weakly consistent and the estimator for $(\Sigma_{p,1}, \Sigma_{p,2})$ is shown to be optimal in the sense of minimizing the asymptotic loss within the class of equivariant estimators as $n_1, n_2, p \to \infty$ with $p/n_1 \to c_1 \in (0, 1)$, $p/n_2 \to c_2 \in (0, 1) \cup (1, \infty)$. Also, our estimators are easy to implement. Even when p is 1000, our estimators can be computed in seconds using a personal laptop.

Keywords: Covariance matrix estimation, High-dimensional asymptotics, Marčenko–Pastur equation, Random matrix theory, Spectrum estimation, Two-sample problem.

1. Introduction

It is widely accepted that covariance matrices play a vital role in various statistical problems. However, in most real life applications, the true (population) covariance matrices are unknown. Therefore, a good estimate of it is much in demand. Traditionally, when the population covariance matrix is needed, what statisticians usually do is to use the sample covariance matrix instead. It is well known that when the dimension p of the covariance matrix is fixed and the sample size n tends to infinity, the sample covariance matrix is a consistent estimator of its population counterpart. However, when the dimension p of the covariance matrix is large, especially when the magnitude of p is comparable to the sample size n, the sample covariance matrix no longer performs as well as it does in the small p large n case; see, e.g., [19] for an illustration.

With the development of Random Matrix Theory (RMT), and especially spectral analysis of random matrices, quite a number of new statistical tools on covariance matrices related problems assuming p and n both large have been proposed, which largely stimulates the exploration of better covariance matrices estimators. One of the important techniques rooted in RMT is the so-called Marčenko–Pastur (or MP for short) equation technique. Initiated from the seminal paper [29], MP equation technique has been extensively studied in recent years, see [3–5, 32, 33, 40]. There has been quite a number of statistical applications resulting from it; see [2, 13, 20–22, 25, 26, 31, 39]. Sometimes not only are people interested in estimating one covariance matrix, but also in estimating two covariance matrices and the spectrum of their ratio; see [7, 10, 16, 17, 23, 24, 27, 28, 36]. In this article, we base our proposed covariance matrices estimators on this MP equation technique with the aim of getting substantial performance improvement over the traditional estimators.

In the following, for a symmetric positive semidefinite matrix A, its square root $A^{1/2}$ is defined as the unique symmetric positive semidefinite matrix such that $A^{1/2}A^{1/2} = A$. For a $p \times p$ matrix M with real eigenvalues $\lambda_1, \ldots, \lambda_p$, we define the spectral distribution (SD) of M as $F^M = p^{-1} \sum_{i=1}^{p} \mathbf{1}(x \le \lambda_i)$, where $\mathbf{1}$ is the indicator function. Denoting the set of the eigenvalues of $\Sigma_{p,2}\Sigma_{p,1}^{-1}$ as $\mathbf{d}_p = \{d_{p,1}, \ldots, d_{p,p}\}$ with $d_{p,1} \le \cdots \le d_{p,p}$, we impose the following three assumptions throughout this article.

Email address: stawenj@nus.edu.sg (Jun Wen)