Journal of Multivariate Analysis 148 (2016) 160–172

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Spectral analysis of the Moore–Penrose inverse of a large dimensional sample covariance matrix

Taras Bodnar [a], Holger Dette [b], Nestor Parolya [c,*]

[a] Department of Mathematics, Stockholm University, SE-10691 Stockholm, Sweden

[b] Department of Mathematics, Ruhr University Bochum, D-44870 Bochum, Germany

[c] Institute of Empirical Economics, Leibniz University Hannover, D-30167 Hannover, Germany

ARTICLE INFO

Article history:
Received 26 October 2015
Available online 24 March 2016

AMS 2010 subject classifications:
60B20
60F05
60F15
60F17
62H10

Keywords:
CLT
Large-dimensional asymptotics
Moore–Penrose inverse
Random matrix theory

ABSTRACT

For a sample of n independent identically distributed p-dimensional centered random vectors with covariance matrix $\mathbf{\Sigma}_n$ let $\tilde{\mathbf{S}}_n$ denote the usual sample covariance (centered by the mean) and $\mathbf{S}_n$ the non-centered sample covariance matrix (i.e. the matrix of second moment estimates), where $p > n$. In this paper, we provide the limiting spectral distribution and central limit theorem for linear spectral statistics of the Moore–Penrose inverse of $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$. We consider the large dimensional asymptotics when the number of variables $p \to \infty$ and the sample size $n \to \infty$ such that $p/n \to c \in (1, +\infty)$. We present a Marchenko–Pastur law for both types of matrices, which shows that the limiting spectral distributions for both sample covariance matrices are the same. On the other hand, we demonstrate that the asymptotic distribution of linear spectral statistics of the Moore–Penrose inverse of $\tilde{\mathbf{S}}_n$ differs in the mean from that of $\mathbf{S}_n$.

1. Introduction

Many statistical, financial and genetic problems require estimates of the inverse population covariance matrix which are often constructed by inverting the sample covariance matrix. Nowadays, the modern scientific data sets involve the large number of sample points which is often less than the dimension (number of features) and so the sample covariance matrix is not invertible. For example, stock markets include a large number of companies which is often larger than the number of available time points; or the DNA can contain a fairly large number of genes in comparison to a small number of patients. In such situations, the Moore–Penrose inverse or pseudoinverse of the sample covariance matrix can be used as an estimator for the precision matrix (see, e.g., Srivastava [15], Kubokawa and Srivastava [7], Hoyle [6], Bodnar et al. [4]).

In order to better understand the statistical properties of estimators and tests based on the Moore–Penrose inverse in high-dimensional settings, it is of interest to study the asymptotic spectral properties of the Moore–Penrose inverse, for example convergence of its linear spectral statistics (LSS). This information is of great interest for high-dimensional statistics because more efficient estimators and tests, which do not suffer from the "curse of dimensionality" and do not reduce the number of dimensions, may be constructed and applied in practice. Most of the classical multivariate procedures are based on the central limit theorems assuming that the dimension p is fixed and the sample size n increases. However, it has been

* Corresponding author.
E-mail addresses: taras.bodnar@math.su.se (T. Bodnar), holger.dette@ruhr-uni-bochum.de (H. Dette), nestor.parolya@ewifo.uni-hannover.de (N. Parolya).

http://dx.doi.org/10.1016/j.jmva.2016.03.001
0047-259X/

pointed out by numerous authors that this assumption does not yield precise distributional approximations for commonly used statistics, and that better approximations can be obtained considering scenarios where the dimension tends to infinity as well (see, e.g., Bai and Silverstein [2] and references therein). More precisely, under the high-dimensional asymptotics we understand the case when the sample size n and the dimension p tend to infinity, such that their ratio p/n converges to some positive constant c. Under this condition the well-known Marchenko–Pastur equation as well as Marchenko–Pastur law were derived (see, Marčenko and Pastur [8], Silverstein [14]).

While most authors in random matrix theory investigate spectral properties of the sample covariance matrix $\mathbf{S}_n = 1/n \sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top$ (here $\mathbf{y}_1, \ldots, \mathbf{y}_n$ denotes a sample of i.i.d. p-dimensional random column vectors with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_n$), Pan [10] studies the differences occurring if $\mathbf{S}_n$ is replaced by its centered version $\tilde{\mathbf{S}}_n = 1/n \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top$ (here $\bar{\mathbf{y}}$ denotes the sample mean of $\mathbf{y}_1, \ldots, \mathbf{y}_n$). Very recently, Zheng et al. [17] have clarified this issue using a *substitution principle* but it is still not clear if the same method is applicable to the Moore–Penrose inverse of the centered sample covariance matrix. Corresponding (asymptotic) spectral properties for the inverse of $\mathbf{S}_n$ have also been recently derived by Zheng et al. [16] in the case $p < n$, which correspond to the case $c < 1$. The aim of the present paper is to close a gap in the literature and focussing on the case $c \in (1, \infty)$. We investigate the differences in the asymptotic spectral properties of Moore–Penrose inverses of centered and non-centered sample covariance matrices. In particular we provide the limiting spectral distribution and the central limit theorem (CLT) for linear spectral statistics of the Moore–Penrose inverse of the sample covariance matrix.

In Section 2 we present the Marchenko–Pastur equation together with a Marchenko–Pastur law for the Moore–Penrose inverse of the sample covariance matrix. Section 3 is divided into two parts: the first one is dedicated to the CLT for the LSS of the pseudoinverse of the non-centered sample covariance matrix while the second part covers the case when the sample covariance matrix is a centered one. While the limiting spectral distributions for both sample covariance matrices are the same, it is shown that the asymptotic distribution of LSS of the Moore–Penrose inverse of $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ differ. Finally, some technical details are given in the Appendix.

2. Preliminaries and the Marchenko–Pastur equation

Throughout this paper we use the following notations and assumptions:

- For a symmetric matrix $\mathbf{A}$ we denote by $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_p(\mathbf{A})$ its ordered eigenvalues and by $F^{\mathbf{A}}(t)$ the corresponding empirical distribution function (e.d.f.), that is
$$F^{\mathbf{A}}(t) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}\{\lambda_i(\mathbf{A}) \leq t\},$$
where $\mathbb{1}\{\cdot\}$ is the indicator function.
- (A1) Let $\mathbf{X}_n$ be a $p \times n$ matrix which consists of independent and identically distributed (i.i.d.) real random variables with zero mean and unit variance.
- (A2) For the latter matrix $\mathbf{X}_n = (X_{ij})_{i=1,\ldots,p}^{j=1,\ldots,n}$ we assume additionally that $E(X_{11}^{4+\delta}) < \infty$ for some $\delta > 0$.
- By
$$\mathbf{Y}_n = \mathbf{\Sigma}_n^{\frac{1}{2}}\mathbf{X}_n$$
we define a $p \times n$ observation matrix with independent columns with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_n$.[1] It is further assumed that neither $\mathbf{\Sigma}_n^{\frac{1}{2}}$ nor $\mathbf{X}_n$ are observable.
- The centered and non-centered sample covariance matrix are denoted by
$$\tilde{\mathbf{S}}_n = \frac{1}{n}(\mathbf{Y}_n - \bar{\mathbf{y}}\mathbf{1}^\top)(\mathbf{Y}_n - \bar{\mathbf{y}}\mathbf{1}^\top)^\top = \frac{1}{n}\mathbf{Y}_n\mathbf{Y}_n^\top - \bar{\mathbf{y}}\bar{\mathbf{y}}^\top$$
$$\mathbf{S}_n = \frac{1}{n}\mathbf{Y}_n\mathbf{Y}_n^\top = \frac{1}{n}\mathbf{\Sigma}_n^{\frac{1}{2}}\mathbf{X}_n\mathbf{X}_n^\top\mathbf{\Sigma}_n^{\frac{1}{2}},$$
where $\mathbf{1}$ denotes the n-dimensional column vector of ones and $\bar{\mathbf{y}} = 1/n \sum_{i=1}^n \mathbf{y}_i$. The corresponding e.d.f.'s are given by $F^{\tilde{\mathbf{S}}_n}$ and $F^{\mathbf{S}_n}$, respectively.
- The Moore–Penrose inverse of a $p \times n$ matrix $\mathbf{A}$ is denoted by $\mathbf{A}^+$ and by definition must satisfy the following four criteria (see, e.g., Horn and Johnsohn [5])
  (i) $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$,
  (ii) $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$,
  (iii) $\mathbf{A}\mathbf{A}^+$ is symmetric,
  (iv) $\mathbf{A}^+\mathbf{A}$ is symmetric.

[1] We could easily include the population mean vector into the model but it will only make the formulas for weak convergence more complex not the analysis itself.