Journal of Multivariate Analysis 142 (2015) 133–143

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A better approximation of moments of the eigenvalues and eigenvectors of the sample covariance matrix

A. Enguix-González [*], J.L. Moreno-Rebollo, J.M. Muñoz-Pichardo

Department of Statistics and Operations Research, Faculty of Mathematics, University of Seville, C/ Tarfia s/n, 41012 Seville, Spain

ARTICLE INFO

Article history:
Received 23 September 2013
Available online 14 August 2015

AMS subject classifications:
62H25
62F12

Keywords:
Eigenvalues
Eigenvectors
Sample covariance matrix
Principal Component Analysis
Influence Analysis

ABSTRACT

Lawley (Lawley, 1956) obtained an approximation, through the first terms of a series expansion, of certain moments of an eigenvalue of the sample covariance matrix. The aim of this paper is to improve that approximation and to calculate a similar approximation for certain moments of the associated eigenvector. The results have practical applications in certain fields of Statistics, such as Influence Analysis.

1. Introduction

Eigenvalues and eigenvectors of the sample covariance matrix are the basic statistics in some statistical techniques, such as Principal Component Analysis, and their moments play a major role in the inferential process.

However, the exact distributions of the eigenvalues and eigenvectors of the sample covariance matrix remain unknown. In the literature various approximations have been obtained. Lawley [4] obtained the first terms of a series expansion of a sample eigenvalue related to a simple population eigenvalue. Moreover, under the assumption of normality, approximated values for the mean, variance and covariance matrix were obtained from the expansion series. Approximations for the moments of the sample eigenvectors are given in Seber [6].

In this paper, we add a term to the series expansion obtained by Lawley [4] and improve the approximations to their moments. Furthermore, series expansions for the eigenvectors and their moments are obtained. The use of symbolic computation with MAPLE has proved to be very helpful in dealing with the cumbersome calculations.

In Section 2, the notation to be used throughout the paper and some basic results are set out. In Section 3, the procedure applied to obtain the terms of the series expansions is laid out and the expressions of the five first terms are provided. In Section 4, approximations of a number of moments of the eigenvalues, which improve those given by Lawley [4], are obtained. Furthermore, some moments of the eigenvectors are calculated with the same degree of accuracy. In Section 5, a numeric evaluation through a Monte Carlo simulation has been carried out. In the last section, the most relevant conclusions are stated.

[*] Corresponding author.
E-mail address: aenguix@us.es (A. Enguix-González).

http://dx.doi.org/10.1016/j.jmva.2015.08.002
0047-259X/

2. Notation and basic results

Let $\underline{X}$ be a p-dimensional random vector with mean vector $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ and covariance matrix $\mathbf{\Sigma} = (\sigma_{st})$. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of $\mathbf{\Sigma}$, $\underline{\alpha}_1, \underline{\alpha}_2, \ldots, \underline{\alpha}_p$ orthonormal associated eigenvectors, respectively, and $\mathbf{A} = \left[\underline{\alpha}_1\, \underline{\alpha}_2\, \cdots\, \underline{\alpha}_p\right]'$. Let $\underline{Y} = \left(Y_1, \ldots, Y_p\right)' = \mathbf{A}\left(\underline{X} - \underline{\mu}\right)$. The (population) principal components of $\underline{X}$ are $Y_k = \underline{\alpha}_k'\left(\underline{X} - \underline{\mu}\right)$, $k = 1, \ldots, p$. Obviously, the covariance matrix of $\underline{Y}$ is given by $\mathbf{\Sigma}_{\underline{Y}} = \mathbf{A\Sigma A}' = \mathbf{\Lambda} = diag\left\{\lambda_1, \ldots, \lambda_p\right\}$, where *diag* denotes the diagonal matrix.

In practice, $\underline{\mu}$ and $\mathbf{\Sigma}$ are unknown parameters. Therefore, they have to be estimated from a random sample, $\underline{X}_1, \ldots, \underline{X}_n$, with $\underline{X}_i = \left(X_{i1}, \ldots, X_{ip}\right)'$, $i = 1, \ldots, n$. These are usually estimated by the sample mean vector, $\overline{\underline{X}} = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)'$, where $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, $j = 1, \ldots, p$, and the sample covariance matrix $\mathbf{S} = (S_{lm}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\underline{X}_i - \overline{\underline{X}}\right)\left(\underline{X}_i - \overline{\underline{X}}\right)'$, respectively. Furthermore, the eigenvalues, $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_p$, and orthonormal associated eigenvectors, $\underline{\widetilde{\alpha}}_1, \underline{\widetilde{\alpha}}_2, \ldots, \underline{\widetilde{\alpha}}_p$, of $\mathbf{S}$ can be considered as estimators of the population eigenvalues and eigenvectors, respectively. The sample principal components of $\underline{X}_i$ are given by $\widetilde{Y}_{ik} = \underline{\widetilde{\alpha}}_k'\left(\underline{X}_i - \overline{\underline{X}}\right)$, $k = 1, \ldots, p$, $i = 1, \ldots, n$. Moreover, the sample covariance matrix of $\underline{Y}$ is given by $\mathbf{S}_{\underline{Y}} = \mathbf{ASA}' = \mathbf{T} = (T_{lm})$.

Since $\mathbf{T}$ is an orthonormal transformation of $\mathbf{S}$, the eigenvalues of $\mathbf{S}$ coincide with the eigenvalues of $\mathbf{T}$ and the eigenvectors of $\mathbf{T}$ are given by $\underline{\widetilde{\beta}}_k = \mathbf{A}\underline{\widetilde{\alpha}}_k$.

By using the orthonormal transformation of $\mathbf{S}$, $\mathbf{T} = \mathbf{ASA}'$, and taking into account that $\mathbf{T}$ converges almost surely to $\mathbf{\Lambda} = diag\left\{\lambda_1, \ldots, \lambda_p\right\}$, Lawley [4] calculated the first terms of the series expansion of an eigenvalue of $\mathbf{S}$. Assuming that $\lambda_1, \ldots, \lambda_p > 0$, if the sample size is sufficiently large and λ_k is simple, the convergent series expansion of the sample eigenvalue $\widetilde{\lambda}_k$ is given by

$$\widetilde{\lambda}_k = \lambda_k + (T_{kk} - \lambda_k) - \sum_{l\neq k}\frac{T_{lk}^2}{\lambda_l - \lambda_k} - (T_{kk} - \lambda_k)\sum_{l\neq k}\frac{T_{lk}^2}{(\lambda_l - \lambda_k)^2} + \sum_{l\neq k}\sum_{m\neq k}\frac{T_{lk}\,(T_{lm} - \lambda_l\delta_{lm})\,T_{mk}}{(\lambda_l - \lambda_k)\,(\lambda_m - \lambda_k)} + R_k, \tag{1}$$

where δ represents Kronecker's delta and R_k is a sum of terms with a product of four or more factors of the type $T_{lm} - \lambda_l\delta_{lm}$.

Since the rest of a convergent series tends to zero, $\widetilde{\lambda}_k$ can be approximated by $\widetilde{\lambda}_k - R_k$. From $\widetilde{\lambda}_k - R_k$ approximations to the moments of $\widetilde{\lambda}_k$ can be obtained.

Assuming that $\underline{X} \sim \mathcal{N}_p\left(\underline{\mu}, \mathbf{\Sigma}\right)$ and λ_k is a simple eigenvalue, Lawley [4] obtains:

$$E\left(\widetilde{\lambda}_k\right) = \lambda_k - \frac{1}{n-1}\sum_{j\neq k}\frac{\lambda_j\lambda_k}{\lambda_j - \lambda_k} + O\left(n^{-2}\right),$$

$$var\left(\widetilde{\lambda}_k\right) = \frac{2\lambda_k^2}{n-1}\left[1 - \frac{1}{n-1}\sum_{j\neq k}\frac{\lambda_j^2}{\left(\lambda_j - \lambda_k\right)^2}\right] + O\left(n^{-3}\right),$$

$$cov\left[\widetilde{\lambda}_j, \widetilde{\lambda}_k\right] = \frac{2}{(n-1)^2}\frac{\lambda_j^2\lambda_k^2}{\left(\lambda_j - \lambda_k\right)^2} + O\left(n^{-3}\right), \quad j \neq k.$$

Regarding the eigenvectors associated to simple eigenvalues, it is known that, see Seber [6],

$$E\left(\underline{\widetilde{\alpha}}_k\right) = \underline{\alpha}_k + O\left(n^{-1}\right),$$

$$var\left(\underline{\widetilde{\alpha}}_k\right) = \frac{\lambda_k}{n-1}\sum_{j\neq k}\frac{\lambda_j}{\left(\lambda_j - \lambda_k\right)^2}\underline{\alpha}_j\underline{\alpha}_j' + O\left(n^{-2}\right),$$

$$cov\left(\underline{\widetilde{\alpha}}_j, \underline{\widetilde{\alpha}}_k\right) = -\frac{1}{n-1}\frac{\lambda_j\lambda_k}{\left(\lambda_j - \lambda_k\right)^2}\underline{\alpha}_j\underline{\alpha}_k' + O\left(n^{-2}\right), \quad \text{for } j \neq k.$$

3. Series expansion of eigenvalues and eigenvectors

In this section, the necessary tools to obtain the series expansions of the sample eigenvalues and eigenvectors are provided, and the first five terms are explicitly obtained.

Let $\mathcal{D}_p$ be the set of symmetric positive definite matrices of dimension p. Let $\mathbf{\Sigma}_0 \in \mathcal{D}_p$ and let $\mathcal{N}\left(\mathbf{\Sigma}_0\right) \subseteq \mathcal{D}_p$ be a neighborhood of $\mathbf{\Sigma}_0$. Let $\lambda_k^{\mathbf{\Sigma}_0}, \underline{\alpha}_k^{\mathbf{\Sigma}_0}$, $k = 1, \ldots, p$, be functions satisfying:

- $\lambda_k^{\mathbf{\Sigma}_0} : \mathcal{N}\left(\mathbf{\Sigma}_0\right) \longrightarrow \mathbb{R}^+$; $\underline{\alpha}_k^{\mathbf{\Sigma}_0} : \mathcal{N}\left(\mathbf{\Sigma}_0\right) \longrightarrow \mathbb{R}^p$, $k = 1, \ldots, p$.
- If $\mathbf{\Sigma} \in \mathcal{N}\left(\mathbf{\Sigma}_0\right)$, then $\lambda_1^{\mathbf{\Sigma}_0}\left(\mathbf{\Sigma}\right) > \lambda_2^{\mathbf{\Sigma}_0}\left(\mathbf{\Sigma}\right) > \cdots > \lambda_p^{\mathbf{\Sigma}_0}\left(\mathbf{\Sigma}\right) > 0$.