Journal of Multivariate Analysis 151 (2016) 151–166

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

High-dimensional inference on covariance structures via the extended cross-data-matrix methodology

Kazuyoshi Yata, Makoto Aoshima *

Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan

ARTICLE INFO

Article history:
Received 20 March 2015
Available online 3 August 2016

AMS subject classifications:
primary 62H15
62H12
secondary 62H10

Keywords:
Correlations test
Graphical modeling
Large p, small n
Partial correlation
Pathway analysis
RV-coefficient

ABSTRACT

Tests of the correlation matrix between two subsets of a high-dimensional random vector are considered. The test statistic is based on the extended cross-data-matrix methodology (ECDM) and shown to be unbiased. The ECDM estimator is also proved to be consistent and asymptotically Normal in high-dimensional settings. The authors propose a test procedure based on the ECDM estimator and evaluate its size and power, both theoretically and numerically. They give several applications of the ECDM estimator and illustrate the performance of the test procedure using microarray data.

1. Introduction

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a random sample of size $n \geq 4$ from a p-variate distribution. We are interested here in situations where the data dimension, p, is very high compared to the sample size n.

For each $j \in \{1, \ldots, n\}$, write $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^\top, \boldsymbol{x}_{2j}^\top)^\top$, where for $i \in \{1, 2\}$, $\boldsymbol{x}_{ij} \in \mathbb{R}^{p_i}$ with $p_1 \in \{1, \ldots, p-1\}$ and $p_2 = p - p_1$. Assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ have unknown mean vector, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$, and unknown covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^\top & \boldsymbol{\Sigma}_2 \end{pmatrix} \geq \boldsymbol{0}.$$

In other words, for all $j \in \{1, \ldots, n\}$ and $i \in \{1, 2\}$,

$$\mathrm{E}(\boldsymbol{x}_{ij}) = \boldsymbol{\mu}_i, \qquad \mathrm{var}(\boldsymbol{x}_{ij}) = \boldsymbol{\Sigma}_i, \qquad \mathrm{cov}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \mathrm{E}(\boldsymbol{x}_{1j}\boldsymbol{x}_{2j}^\top) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top = \boldsymbol{\Sigma}_*.$$

For all $i \in \{1, 2\}$ and $k \in \{1, \ldots, p_i\}$, the kth diagonal element σ_{ik} of $\boldsymbol{\Sigma}_i$ is assumed to be strictly positive. Then, for all $j \in \{1, \ldots, n\}$,

$$\mathrm{corr}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \boldsymbol{P} = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{1p_1})^{-1/2}\boldsymbol{\Sigma}_*\mathrm{diag}(\sigma_{21}, \ldots, \sigma_{2p_2})^{-1/2}.$$

In this paper, we consider the problem of testing the hypotheses

$$\mathcal{H}_0 : \boldsymbol{P} = \boldsymbol{0} \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{P} \neq \boldsymbol{0} \tag{1}$$

in high-dimensional settings. When $(p_1, p_2) = (p-1, 1)$ or $(1, p-1)$, testing (1) amounts to testing correlation coefficients.

* Corresponding author.
E-mail address: aoshima@math.tsukuba.ac.jp (M. Aoshima).

http://dx.doi.org/10.1016/j.jmva.2016.07.011
0047-259X/

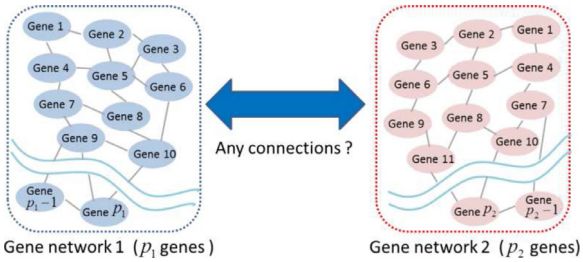

Fig. 1. Relevance of hypotheses (1) illustrated in the context of gene networks.

Aoshima and Yata [1] proposed a statistic for the latter problem and Yata and Aoshima [19] improved this test statistic by using a method called the *extended cross-data-matrix methodology* (ECDM). However, tests on the correlation matrix are equally important, e.g., in pathway analysis or graphical modeling for high-dimensional data. One possible application pertains to the construction of gene networks, as portrayed in Fig. 1.

Here, we consider testing partial correlation coefficients. When $\boldsymbol{\Sigma} > \mathbf{0}$, write

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}_1 & \boldsymbol{\Omega}_* \\ \boldsymbol{\Omega}_*^\top & \boldsymbol{\Omega}_2 \end{pmatrix} = (\omega_{ij}),$$

where, for $i \in \{1, 2\}$, $\boldsymbol{\Omega}_i$ is the corresponding $p_i \times p_i$ matrix. Here, (m_{ij}) denotes a matrix whose (i, j)th element is m_{ij}. When $i \neq j$, $-\omega_{ij}(\omega_{ii}\omega_{jj})^{-1/2}$ is the (i, j)th partial correlation coefficient; see, e.g., Drton and Perlman [5]. We denote the partial correlation coefficient matrix by

$$\boldsymbol{P}_\Omega = -\mathrm{diag}(\omega_{11}, \ldots, \omega_{p_1p_1})^{-1/2}\boldsymbol{\Omega}_*\mathrm{diag}(\omega_{p_1+1p_1+1}, \ldots, \omega_{pp})^{-1/2}$$

and note that the test of the hypotheses

$$\mathcal{H}_0 : \boldsymbol{P}_\Omega = \mathbf{0} \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{P}_\Omega \neq \mathbf{0}$$

is equivalent to the test of hypotheses (1) since $\boldsymbol{\Omega}_* = \mathbf{0}$ is equivalent to $\boldsymbol{\Sigma}_* = \mathbf{0}$.

Drton and Perlman [5] and Wille et al. [16] considered pathway analysis or graphical modeling of microarray data by testing an individual partial correlation coefficient. For example, Wille et al. [16] analyzed gene networks of microarray data with $p = 834$ ($p_1 = 39$ and $p_2 = 795$) and $n = 118$. In contrast, Hero and Rajaratnam [8] considered correlation screening procedures for high-dimensional data by testing correlations. Lan et al. [10] and Zhong and Chen [20] considered tests of regression coefficient vectors in linear regression models. As for tests of independence, see, among others, Fujikoshi et al. [7], Hyodo et al. [9], Srivastava and Reid [13], and Yang and Pan [17]. Also, one may refer to Székely and Rizzo [14,15] for distance correlation.

In Section 2, we set the notation and state several assumptions required for the construction of our high-dimensional correlation test of hypotheses (1). In Section 3, we produce a test statistic for this problem by using the ECDM methodology and show the unbiasedness of the ECDM estimator. We also show that the ECDM estimator is consistent and asymptotically Normal when $p \to \infty$ and $n \to \infty$. In Section 4, we propose a test procedure for (1) by the ECDM estimator and evaluate its asymptotic size and power when $p \to \infty$ and $n \to \infty$ theoretically and numerically. In Section 5, we give several applications of the ECDM estimator. Finally, we demonstrate how the test procedure performs in practice using microarray data.

2. Assumptions

In this section, we lay out the basic assumptions for the construction of our test of hypotheses (1). The eigenvalue decomposition of $\boldsymbol{\Sigma}$ is denoted by $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^\top$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\boldsymbol{\Sigma}$, while $\boldsymbol{H}$ is an orthogonal matrix of the corresponding eigenvectors.

For all $j \in \{1, \ldots, n\}$, let $\boldsymbol{x}_j = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{z}_j + \boldsymbol{\mu}$, where $\mathrm{E}(\boldsymbol{z}_j) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{z}_j) = \boldsymbol{I}_p$, the identity matrix of dimension p. Note that if $\boldsymbol{x}_j$ is Gaussian, the elements of $\boldsymbol{z}_j$ form a random sample from the standard Normal distribution, $\mathcal{N}(0, 1)$. We assume that, for all $j \in \{1, \ldots, n\}$,

$$\boldsymbol{x}_j = \boldsymbol{\Gamma}\boldsymbol{w}_j + \boldsymbol{\mu}, \tag{2}$$

where $\boldsymbol{\Gamma}$ is a $p \times q$ matrix for some $q > 0$ such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{\Sigma}$, and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ form a random sample, so that for every $j \in \{1, \ldots, n\}$, $\boldsymbol{w}_j = (w_{1j}, \ldots, w_{qj})^\top$, $\mathrm{E}(\boldsymbol{w}_j) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{w}_j) = \boldsymbol{I}_q$. Let $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1^\top, \boldsymbol{\Gamma}_2^\top)^\top$, where for $i \in \{1, 2\}$, $\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{i1}, \ldots, \boldsymbol{\gamma}_{iq})$ with $\boldsymbol{\gamma}_{ij} \in \mathbb{R}^{p_i}$, so that $\boldsymbol{x}_{ij} = \boldsymbol{\Gamma}_i\boldsymbol{w}_j + \boldsymbol{\mu}_i$. Note that

$$\boldsymbol{\Sigma}_* = \boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_2^\top = \sum_{r=1}^{q} \boldsymbol{\gamma}_{1r}\boldsymbol{\gamma}_{2r}^\top.$$

Also note that Eq. (2) includes the case where $\boldsymbol{\Gamma} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{w}_j = \boldsymbol{z}_j$. For all $r \in \{1, \ldots, q\}$, let $\mathrm{var}(w_{rj}^2) = M_r$ and assume that $\limsup_{p\to\infty} M_r < \infty$.