Journal of Multivariate Analysis 152 (2016) 249–258

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Limitations on detecting row covariance in the presence of column covariance

Peter D. Hoff

Department of Statistical Science, Duke University, United States

ARTICLE INFO

Article history:
Received 14 January 2016
Available online 10 September 2016

AMS subject classifications:
62H15
62H25

Keywords:
Hypothesis test
Invariance
Random matrix
Regression
Separable covariance

ABSTRACT

Many inference techniques for multivariate data analysis assume that the rows of the data matrix are realizations of independent and identically distributed random vectors. Such an assumption will be met, for example, if the rows of the data matrix are multivariate measurements on a set of independently sampled units. In the absence of an independent random sample, a relevant question is whether or not a statistical model that assumes such row exchangeability is plausible. One method for assessing this plausibility is a statistical test of row covariation. Maintenance of a constant type I error rate regardless of the column covariance or matrix mean can be accomplished with a test that is invariant under an appropriate group of transformations. In the context of a class of elliptically contoured matrix-variate regression models (such as matrix normal models), it is shown that there are no non-trivial invariant tests if the number of rows is not sufficiently larger than the number of columns. Furthermore, even if the number of rows is large, there are no non-trivial invariant tests that have power to detect arbitrary row covariance in the presence of arbitrary column covariance. However, biased tests can be constructed that have power to detect certain types of row covariance that may be encountered in practice.

1. Introduction

A canonical statistical model for an observed data matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$ is that the rows of the matrix are i.i.d. realizations from a mean-$\boldsymbol{\mu}$ p-variate normal distribution with covariance Σ. We write this hypothesized model as

$$\mathbf{Y} \sim \mathcal{N}_{n\times p}(\mathbf{1}\boldsymbol{\mu}^\top, \Sigma \otimes \mathbf{I}_n),$$

where $\mathbf{1}$ is the n-vector of all 1 s and "$\otimes$" is the Kronecker product. If the rows represent multivariate measurements on a simple random sample of n units from a population, then the assumption of i.i.d. rows is a valid one (or nearly valid for a large finite population, in the case of sampling without replacement). However, in many analyses, the units are obtained from a convenience sample rather than a random sample. We might then want to entertain an alternative model for the data, such as

$$\mathbf{Y} \sim \mathcal{N}_{n\times p}(\mathbf{1}\boldsymbol{\mu}^\top, \Sigma \otimes \Psi),$$

where Ψ is an unknown $n \times n$ covariance matrix describing dependence and heteroscedasticity among the rows of $\mathbf{Y}$. This alternative model is the so-called matrix normal model; see, e.g., Dawid [3]. Letting $\mathbf{y}_i$ and $\mathbf{y}_{i'}$ be two rows of $\mathbf{Y}$, this model implies that $\text{cov}(\mathbf{y}_i, \mathbf{y}_{i'}) = \psi_{i,i'}\Sigma$.

E-mail address: peter.hoff@duke.edu.

http://dx.doi.org/10.1016/j.jmva.2016.09.003
0047-259X/

Several parametric and nonparametric tests of row dependence in the presence of column dependence were considered in Efron [4] for the case that $p > n$. The parametric tests were based on estimates $\hat{\Psi}$ of Ψ in the matrix normal model. Efron suggested that such tests appear to be promising, but suffer some deficiencies. In particular, the distribution of the proposed estimate $\hat{\Psi}$ of Ψ depends on the unknown value of Σ, a phenomenon that Efron referred to as "leakage". Proceeding with a similar approach, Muralidharan [9] constructed a permutation invariant test using asymptotic approximations in the $p > n$ scenario. This test is conservative, and has power that depends on both Σ and Ψ, that is, it also experiences some leakage.

The issue of leakage suggests the use of invariant tests which, having power functions that do not depend on the parameters of the null model, are leakage-free. In this article, we characterize the invariant tests of $\mathcal{H} : \Psi = \mathbf{I}$ versus $\mathcal{K} : \Psi \neq \mathbf{I}$ in matrix regression models that have a stochastic representation of the form

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^\top + \Psi^{1/2}\mathbf{Z}\Sigma^{1/2},$$

where $\mathbf{X} \in \mathbb{R}^{n\times q}$ is an observed regression matrix, $(\mathbf{B}, \Sigma, \Psi)$ are unknown parameters, and $\mathbf{Z}$ is a random mean-zero error matrix with uncorrelated entries. For notational simplicity, the results in this article are developed for Gaussian random matrices, but as will be discussed, the results hold for a more general class of elliptically contoured matrix distributions, including heavy-tailed and contaminated distributions; see Gupta and Varga [7].

The results of this article are primarily negative, illustrating inherent limitations on our ability to detect arbitrary row covariance in the presence of arbitrary column covariance. In the next section, I show that if $n \leq p + q$ then there are no non-trivial invariant tests of $\mathcal{H}$ versus $\mathcal{K}$. In Section 3, I show that if $n > p + q$ then there are no non-trivial unbiased invariant tests. The implication of these results is that, for these matrix regression models, there are no useful invariant tests for arbitrary row covariance in the presence of arbitrary column covariance. On the bright side, one can construct biased invariant tests that have power to detect certain types of row dependence that may be of interest in practice. For example, in Section 4, I obtain the UMP invariant test in a submodel where the eigenvector structure of Ψ is known. This result is used in Section 5 to construct a test that has the ability to detect positive dependence among arbitrary pairs of rows. The use of this test is illustrated on several datasets. In Section 6, I show how the results of the other sections generalize to non-Gaussian models, and discuss some open questions.

2. Invariant test statistics

We are interested in testing $\mathcal{H} : \Psi = \mathbf{I}$ versus $\mathcal{K} : \Psi \neq \mathbf{I}$ in the matrix normal regression model

$$\mathbf{Y} \sim \mathcal{N}_{n\times p}(\mathbf{X}\mathbf{B}^\top, \Sigma \otimes \Psi), \quad \mathbf{B} \in \mathbb{R}^{p\times q},\ \Sigma \in \mathcal{S}_p^+,\ \Psi \in \mathcal{S}_n^+, \tag{1}$$

where $\mathbf{X}$ is a known $n \times q$ matrix with rank $q < n$ and $\mathcal{S}_k^+$ denotes the space of $k \times k$ nonsingular covariance matrices. Let $\mathbf{P} = I - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ so that $\mathbf{R} \equiv \mathbf{P}\mathbf{Y}$ is the matrix of residuals corresponding to the least-squares estimate of $\mathbf{B}$. Then $\mathrm{E}(\mathbf{R}\mathbf{R}^\top|\mathbf{B}, \Sigma \otimes \Psi) = \mathrm{tr}(\Sigma) \times \mathbf{P}\Psi\mathbf{P}$, which suggests the use of $\mathbf{R}\mathbf{R}^\top$ to test whether or not $\Psi = \mathbf{I}$. The problem with such an approach is that, as pointed out by Efron [4], the distribution of $\mathbf{R}\mathbf{R}^\top$ will generally depend on the unknown value of Σ. If the distribution of a test statistic depends on Σ, then maintaining the level of the test for all Σ without sacrificing power is difficult.

With this in mind, we would like to identify test statistics whose distributions under $\mathcal{H}$ do not depend on $\mathbf{B}$ or Σ. To do this, we first note that the model and testing problem are invariant under the group $\mathcal{G}$ of transformations g of the form $g(\mathbf{Y}) = \mathbf{X}\mathbf{C}^\top + \mathbf{Y}\mathbf{A}^\top$ for $\mathbf{C} \in \mathbb{R}^{p\times q}$ and nonsingular $\mathbf{A} \in \mathbb{R}^{p\times p}$: If $\mathbf{Y} \sim \mathcal{N}_{n\times p}(\mathbf{X}\mathbf{B}^\top, \Sigma \otimes \Psi)$, then $g(\mathbf{Y}) \sim \mathcal{N}_{n\times p}[\mathbf{X}(\mathbf{A}\mathbf{B} + \mathbf{C})^\top, \mathbf{A}\Sigma\mathbf{A}^\top \otimes \Psi]$, It follows that the group $\mathcal{G}$ induces a group $\bar{\mathcal{G}}$ of transformations on the parameter space of the form $\bar{g}(\mathbf{B}, \Sigma \otimes \Psi) = (\mathbf{A}\mathbf{B} + \mathbf{C}, \mathbf{A}\Sigma\mathbf{A}^\top \otimes \Psi)$. This group is transitive on the null parameter space, and so any statistic or test function ϕ that is invariant to $\mathcal{G}$, meaning that $\phi\{g(\mathbf{Y})\} = \phi(\mathbf{Y})$ for all $g \in \mathcal{G}$, will have a distribution that does not depend on $\mathbf{B}$ or Σ. In particular, if ϕ is invariant then $\mathrm{E}\{\phi(\mathbf{Y})|\mathbf{B}, \Sigma \otimes \mathbf{I}\}$ is constant in $\mathbf{B}$ and Σ.

2.1. Maximal invariant statistics

Any invariant test function or statistic must depend on $\mathbf{Y}$ only through a statistic that is maximal invariant, that is, an invariant function M of $\mathbf{Y}$ for which $M(\mathbf{Y}_1) = M(\mathbf{Y})$ implies $\mathbf{Y}_1 = g(\mathbf{Y})$ for some $g \in \mathcal{G}$. Therefore, characterizing the class of invariant tests requires that we find a maximal invariant statistic (since all maximal invariant statistics are functions of each other, we only need to find one). One maximal invariant statistic in particular has an intuitive form: Let $\hat{\mathbf{B}}$ be the OLS estimator of $\mathbf{B}$, let

$$\hat{\Sigma} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^\top)^\top(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^\top)/n,$$

and let $\hat{\Sigma}^-$ be the inverse or Moore–Penrose inverse of $\hat{\Sigma}$, depending on whether or not $\hat{\Sigma}$ is full rank. As will be shown below, the $n \times n$ matrix given by

$$M(\mathbf{Y}) = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^\top)\hat{\Sigma}^-(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^\top)^\top/n$$