Journal of Multivariate Analysis 114 (2013) 378–388

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Comparison of confidence intervals for correlation coefficients based on incomplete monotone samples and those based on listwise deletion

K. Krishnamoorthy

Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA

ARTICLE INFO

Article history:
Received 9 February 2012
Available online 13 August 2012

AMS subject classifications:
62H12
62H15

Keywords:
Coverage probability
Lower triangular invariant
Missing at random
Monotone samples

ABSTRACT

Inferential procedures for estimating and comparing normal correlation coefficients based on incomplete samples with a monotone missing pattern are considered. The procedures are based on the generalized variable (GV) approach. It is shown that the GV methods based on complete or incomplete samples are exact for estimating or testing a simple correlation coefficient. Procedures based on incomplete samples for comparing two overlapping dependent correlation coefficients are also proposed. For both problems, Monte Carlo simulation studies indicate that the inference based on incomplete samples and those based on samples after listwise or pairwise deletion are similar, and the loss of efficiency by ignoring additional data is not appreciable. The proposed GV approach is simple, and it can be readily extended to other problems such as the one of estimating two non-overlapping dependent correlations. The results are illustrated using two examples.

1. Introduction

The Pearson product–moment correlation is the most popular measure of association between two continuous random variables. Assuming normality, several authors have addressed the problem of estimating or testing correlation coefficients in various contexts, and provided solutions based on large sample theory. If the underlying distribution is bivariate normal, then an exact t procedure is available to test if the population correlation coefficient ρ is significantly different from zero. To test a non-zero value of ρ, the test based on [5]'s z transformation of the sample correlation coefficient is commonly used. Fisher's approach is reasonably accurate for moderate samples, and standard software packages use this approach to find confidence limits (CLs) for ρ. There is an exact method, which produces uniformly most accurate confidence intervals (CIs), available in the literature (see [2, Section 4.2]). However, this exact method is not popular because of computational complexity.

Fisher's z transformation for the one-sample case can be readily extended to the problem of testing two independent correlation coefficients, but the test cannot be transformed into a procedure for setting CLs for the difference between correlation coefficients. Olkin and Finn [16,17]) proposed a normal based asymptotic method that can be used for testing as well as for obtaining CIs. In general, the procedures given in the literature are based on asymptotic theory, and simulation studies by Krishnamoorthy and Xia [11] indicated that such asymptotic procedures are, in some cases, not satisfactory even for large samples.

In this article, we consider inferential procedures for correlation coefficients based on missing data. Missing data arises, for example, during data gathering and recording, when the experiment involves a group of individuals over a period of time like in clinical trials or in a planned experiment where the variables that are expensive to measure are collected only from a subset of a sample. To ignore the missingness mechanism, we assume that the data are missing at random (MAR). Lua and Copas [13] noted that inference from the likelihood method is valid if and only if the missing data mechanism is

E-mail address: krishna@louisiana.edu.

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.08.003

MAR. For formal definition and exposition of MAR or missing completely at random (MCAR), we refer to [12, Section 1.3], and [8]. There are a few missing patterns considered in the literature, but the incomplete data with monotone pattern is common, and it is convenient for making inference. For the multivariate normal case, Anderson [1] gives a simple approach to derive the maximum likelihood estimates (MLEs) and present them for a special case. Some invariance properties of the MLEs enable us to develop finite sample inferential procedures for the mean and the covariance matrix of a multivariate normal distribution. See the articles by Krishnamoorthy and Pannala [9,10], Hao and Krishnamoorthy [7], the recent articles by Chang and Richards [3,4], and the references therein.

Although several papers address the problems of making inference on a multivariate normal mean vector and covariance matrix, the problem of estimating or testing a correlation coefficient with missing data is seldom addressed in the literature. Our online review indicates that commonly used software packages use the standard approach after deleting the records for subjects with missing observations. This standard practice is simple but does not utilize the information of the additional data. So it is of interest to assess the loss of efficiency of the standard approach by comparing the results of the methods that utilize the additional data.

In this article, we provide a generalized variable (GV) method for making inference on a simple correlation coefficient and for comparing two dependent correlation coefficients based on incomplete samples with a monotone pattern. The proposed approach is similar to the one for the complete sample case given in [11], but here we show that the GV solutions to the one-sample problems are exact for complete or incomplete samples. Furthermore, the GV approach for incomplete samples can be readily extended to test or interval estimating the difference between two independent correlations, the difference between two overlapping dependent correlations [15] and the difference between two non-overlapping dependent correlation coefficients [16,14].

The rest of the article is organized as follows. In Section 2, we describe the MLE for the normal covariance matrix $\boldsymbol{\Sigma}$, and develop generalized pivotal quantities (GPQs) for the elements of $\boldsymbol{\Sigma}$. In Section 3, we develop a GPQ for the simple correlation coefficient ρ as a function of the GPQs of the elements of $\boldsymbol{\Sigma}$, and outline inferential procedures based on the GPQ. In Section 4, we extend the results of Section 3 to compare two overlapping dependent correlation coefficients. In Section 5, we compare the expected widths of CIs based on the complete pairs and of those based on incomplete samples to assess the gain in precision by using the additional data. An illustrative example and an example based on simulated data are provided in Section 6. Some concluding remarks and applications of the GV approach to other correlation problems are given in Section 7.

2. Generalized pivotal quantity for a normal correlation matrix

Let $\mathbf{X}$ be p-variate normal random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let the correlation matrix based on $\boldsymbol{\Sigma}$ be

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}.$$

Consider a monotone sample of n_1 subjects with the following pattern:

$$\begin{array}{lllllll} X_{11} & \ldots & X_{1n_p} & \ldots & X_{1n_2} & \ldots & X_{1n_1} \\ X_{21} & \ldots & X_{2n_p} & \ldots & X_{2n_2} & & \\ \vdots & \vdots & & & & & \\ X_{p1} & \ldots & X_{pn_p}. & & & & \end{array} \tag{1}$$

Note that there are n_i observations available on the ith component, $i = 1, \ldots, p$ and we assume that $n_1 \geq n_2 \geq \cdots \geq n_p$. That is, measurements on the first component are available for all n_1 subjects, measurements on the first two components are available only for n_2 subjects, and so on. Let $\widehat{\boldsymbol{\Sigma}}$ be the MLE of $\boldsymbol{\Sigma}$ based on sample (1). Write

$$\widehat{\boldsymbol{\Sigma}} = \mathbf{W}\mathbf{W}', \tag{2}$$

where $\mathbf{W}$ be the Cholesky factor $\widehat{\boldsymbol{\Sigma}}$ with positive diagonal elements. Let $\boldsymbol{\theta} = (\theta_{ij})$ be the Cholesky factor of $\boldsymbol{\Sigma}$. Since the MLE $\widehat{\boldsymbol{\Sigma}}$ is invariant under a lower triangular transformation as well as under location transformation, the distribution of $\boldsymbol{\theta}^{-1}\mathbf{W}$ does not depend on any unknown parameters.

To find a GPQ for $\boldsymbol{\theta} = (\theta_{ij})$, let $\mathbf{w}$ be an observed value of $\mathbf{W}$ defined in (2). Then

$$\mathbf{w}(\boldsymbol{\theta}^{-1}\mathbf{W})^{-1} = \mathbf{w}\mathbf{V}^{-1} = \mathbf{A} = (a_{ij}) \text{ is a GPQ for } \boldsymbol{\theta} = (\theta_{ij}), \quad i \geq j. \tag{3}$$

Notice that $\mathbf{A}$ is a lower triangular matrix with $a_{ij} = 0$ for $i < j$. Furthermore, for a given $\mathbf{w}$, the distribution of $\mathbf{A}$ does not depend on any unknown parameters. The element a_{ij} is a GPQ for θ_{ij} for $i \geq j$. Also, if $h(\boldsymbol{\theta})$ is a real valued function of $\boldsymbol{\theta}$, then $h(\mathbf{A})$ is a GPQ for $h(\boldsymbol{\theta})$. For example, the percentiles of $h(\mathbf{A})$ can be used to construct CIs for $h(\boldsymbol{\theta})$. For more details and numerous applications of the GV approach, see the book by Weerahandi [19], and for details in the present context see [11].