Statistics and Probability Letters 97 (2015) 41–45

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Approximate ellipsoidal tolerance regions for multivariate normal populations

Malick Mbodj, Thomas Mathew *

Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Received 9 September 2014
Received in revised form 5 November 2014
Accepted 5 November 2014
Available online 15 November 2014

Keywords:
Imhof's approximation
Linear combination of non-central chisquare
Liu, Tang and Zhang approximation
Tolerance factor

ABSTRACT

For the approximate computation of an ellipsoidal tolerance region for multivariate normal distributions, an approximation due to Liu et al. (2008) for the cdf of a positive linear combination of independent non-central chisquare random variables is used, along with Monte Carlo simulation. Such a computational approach is originally due to Krishnamoorthy and Mondal (2006, 2008), and the present work provides an improvement.

1. Introduction

A tolerance region for a multivariate population is a region, computed using a random sample, that will contain a specified proportion (say, p) or more of the population, with a given confidence level (say, $1-\alpha$). The quantity p is referred to as the *content* of the tolerance region. A tolerance region that satisfies such a content and confidence level requirement is simply referred to as a $(p, 1-\alpha)$ tolerance region. Tolerance intervals and tolerance regions are widely used in life testing and reliability studies, quality control problems, laboratory medicine, and environmental monitoring. In the univariate case, tolerance intervals have been well investigated in the literature, for normal as well as for non-normal populations, including non-parametric set ups. Literature in the multivariate case is mostly limited to multivariate normal populations; however, non-parametric tolerance regions have also been constructed. We refer to the book by Krishnamoorthy and Mathew (2009) for a detailed discussion on the topic. The present work is on the derivation of multivariate tolerance regions in the context of multivariate normal populations.

Consider the q-variate normal distribution $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are both unknown, and let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ be a random sample from $N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\bar{\boldsymbol{y}}$ denote the sample mean vector, $\boldsymbol{A}$ denote the sample sum of squares and cross products matrix, and $\boldsymbol{S}$ denote the sample covariance matrix. That is, $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i$, $\boldsymbol{A} = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$, and $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{A}$. Then $\bar{\boldsymbol{y}} \sim N_q(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$, $\boldsymbol{A} \sim W_q(\boldsymbol{\Sigma}, n-1)$, and $\boldsymbol{S} \sim W_q\left(\frac{1}{n-1}\boldsymbol{\Sigma}, n-1\right)$, where $W_q(\boldsymbol{\Sigma}, m)$ denotes a q-dimensional Wishart distribution with scale matrix $\boldsymbol{\Sigma}$ and df $= m$. Note that $\bar{\boldsymbol{y}}$ and $\boldsymbol{A}$ are independently distributed and so are $\bar{\boldsymbol{y}}$ and $\boldsymbol{S}$. If $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then an *ellipsoidal* tolerance region for $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is taken to be of the form

$$\{\boldsymbol{y} : (\boldsymbol{y} - \bar{\boldsymbol{y}})'\boldsymbol{S}^{-1}(\boldsymbol{y} - \bar{\boldsymbol{y}}) \le c\}, \tag{1}$$

* Corresponding author. Tel.: +1 410 455 2418.
E-mail address: mathew@umbc.edu (T. Mathew).

http://dx.doi.org/10.1016/j.spl.2014.11.001
0167-7152/

where the quantity c, referred to as the tolerance factor, is determined so that

$$P_{\bar{\boldsymbol{y}},\boldsymbol{S}}[P_{\boldsymbol{y}}\{(\boldsymbol{y}-\bar{\boldsymbol{y}})'\boldsymbol{S}^{-1}(\boldsymbol{y}-\bar{\boldsymbol{y}}) \le c \mid \bar{\boldsymbol{y}}, \boldsymbol{S}\} \ge p] = 1-\alpha. \tag{2}$$

In other words, the *content* of the tolerance region is p, with confidence level $1-\alpha$.

For computing the tolerance factor c in (2), the article by Krishnamoorthy and Mathew (1999) provides a review and a comparison of the various closed form approximations for c available in the literature, and also suggest some new approximations. It turns out that none of the available approximations is satisfactory for all values of the sample size n, dimension q, content p and confidence level $1-\alpha$. In view of this, Krishnamoorthy and Mondal (2006) investigated a simple and efficient method for computing the ellipsoidal tolerance factor c, and is based on a result due to Imhof (1961) for approximating the cdf of a positive linear combination of independent non-central chisquare random variables. The results are summarized in Krishnamoorthy and Mathew (2009, Chapter 9).

The present work is motivated by the observation that the approximate tolerance factor resulting from the use of Imhof's (1961) approximation is not satisfactory when the sample size n is somewhat close to the dimension q. In fact in such situations, the use of Imhof's (1961) approximation can result in coverage probabilities that are significantly below the nominal level; this will be clear from the numerical results reported later in this article. In our work, we have revisited the problem, and have investigated the accurate computation of the tolerance factor c using an approximation due to Liu et al. (2008); the approximation is once again for computing the cdf of a positive linear combination of independent non-central chisquare random variables. The numerical results reported in this paper show that the use of the Liu et al. (2008) approximation provides a rather accurate approximate tolerance factor regardless of the values of the sample size and the dimension. Two illustrative examples are given in the paper, and extension to the case of a multivariate linear regression model is briefly indicated.

2. Computation of the approximate tolerance factor

It is easy to show that c satisfying (2) does not depend on the unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In order to see this, consider the transformations $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu})$, $\bar{\boldsymbol{u}} = \Sigma^{-1/2}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})$, and $\boldsymbol{V} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}$. Then $\boldsymbol{u}$, $\bar{\boldsymbol{u}}$ and $\boldsymbol{V}$ are independent having the distributions

$$\boldsymbol{u} \sim N_q(\mathbf{0}, \mathbf{I}_q), \qquad \bar{\boldsymbol{u}} \sim N_q\left(\mathbf{0}, \frac{1}{n}\mathbf{I}_q\right), \qquad \boldsymbol{V} \sim W_q\left(\frac{1}{n-1}\mathbf{I}_q, n-1\right). \tag{3}$$

Eq. (2) can now be re-written as

$$P_{\bar{\boldsymbol{u}},\boldsymbol{V}}[P_{\boldsymbol{u}}\{(\boldsymbol{u}-\bar{\boldsymbol{u}})'\boldsymbol{V}^{-1}(\boldsymbol{u}-\bar{\boldsymbol{u}}) \le c \mid \bar{\boldsymbol{u}}, \boldsymbol{V}\} \ge p] = 1-\alpha. \tag{4}$$

Since the distribution of $(\boldsymbol{u}, \bar{\boldsymbol{u}}, \boldsymbol{V})$ does not depend on any unknown parameters, we conclude that c satisfying the condition (2), equivalently the condition (4), is free of any unknown parameters. However an explicit analytical form is not available for c, and we have to estimate c using Monte Carlo simulation, or use a suitable approximation.

An algorithm for the Monte Carlo estimation of c is provided in Krishnamoorthy and Mathew (2009, Section 9.4). The authors note that the values of c so obtained are rather unstable, especially when the sample size is small. Consequently, the coverage probability of the resulting tolerance region can be somewhat different from the nominal confidence level of $1-\alpha$. In order to ease the computational burden, Krishnamoorthy and Mondal (2006) used a chisquare approximation due to Imhof (1961) to approximate the inner probability in (4), followed by a simulation to compute the probability with respect to the joint distribution of $(\bar{\boldsymbol{u}}, \boldsymbol{V})$. A representation for (4), involving the eigenvalues of the Wishart matrix $\boldsymbol{V}$, was used for this purpose. In order to give this representation, let $\boldsymbol{Q}$ be a $q \times q$ orthogonal matrix such that $\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q} = \text{diag}(l_1, \ldots, l_q)$, where $l_i's$ are the eigenvalues of the matrix $\boldsymbol{V}$, with $l_1 > l_2 > \cdots > l_q$. Now let us consider $\boldsymbol{Q}'\boldsymbol{u} = \boldsymbol{z} = (Z_1, Z_2, \ldots, Z_q)'$ and $\boldsymbol{Q}'\bar{\boldsymbol{u}} = \bar{\boldsymbol{z}} = (\bar{Z}_1, \bar{Z}_2, \ldots, \bar{Z}_q)'$, so that Eq. (4) becomes

$$P_{\bar{\boldsymbol{z}},\boldsymbol{l}}\left[P_{\boldsymbol{z}}\left\{\sum_{i=1}^{q}\frac{(Z_i-\bar{Z}_i)^2}{l_i} \le c \mid \bar{\boldsymbol{z}}, \boldsymbol{l}\right\} \ge p\right] = 1-\alpha \tag{5}$$

where $\boldsymbol{l} = (l_1, \ldots, l_q)'$. Note that $\boldsymbol{z} = \boldsymbol{Q}'\boldsymbol{y} \sim N_q(\mathbf{0}, \mathbf{I}_q)$, $\bar{\boldsymbol{z}} = \boldsymbol{Q}'\bar{\boldsymbol{u}} \sim N_q(\mathbf{0}, \frac{1}{n}\mathbf{I}_q)$, and $\boldsymbol{z}, \bar{\boldsymbol{z}}$ and $\boldsymbol{l}$ are independently distributed. Furthermore, conditionally given $\bar{\boldsymbol{z}}$, $(Z_i-\bar{Z}_i)^2 \sim \chi_1^2(\bar{Z}_i^2)$, a non-central chisquare distribution with df $= 1$, and non-centrality parameter $\bar{Z}_i^2$, and these non-central chisquare distributions are independent for $i = 1, \ldots, q$. Here is the approximation due to Liu et al. (2008), that we shall use for the approximate computation of the tolerance factor:

Lemma 1. *For $i = 1, 2, \ldots, m$, let $\chi^2_{h_i}(\delta_i)$ denote independent non-central chisquare random variables with* df $= h_i$ *and non-centrality* parameter $= \delta_i$. *Consider the linear combination $\sum_{i=1}^m \lambda_i \chi^2_{h_i}(\delta_i)$ and define $c_u = \sum_{i=1}^m \lambda_i^u h_i + u\sum_{i=1}^m \lambda_i^u \delta_i$, $u = 1, 2, 3$. Then*

$$P\left[\sum_{i=1}^{m}\lambda_i\chi^2_{h_i}(\delta_i) > t\right] \approx P\left[\chi_f^2(\delta) > t^*\sqrt{2(f+2\delta)} + (f+\delta)\right],$$