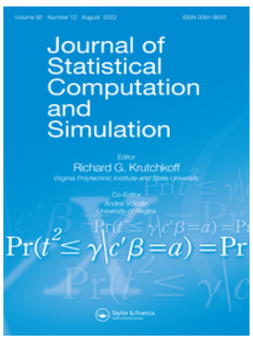

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Computation of tolerance ellipses for bivariate and trivariate normal populations

W. Liu, F. Bretz, A. J. Hayter & S. Kiatsupaibul

To cite this article: W. Liu, F. Bretz, A. J. Hayter & S. Kiatsupaibul (2022): Computation of tolerance ellipses for bivariate and trivariate normal populations, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2022.2076091

To link to this article: https://doi.org/10.1080/00949655.2022.2076091

Published online: 27 Jun 2022.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2022.2076091

Computation of tolerance ellipses for bivariate and trivariate normal populations

W. Liu[a], F. Bretz[b], A. J. Hayter[c] and S. Kiatsupaibul[d]

[a]S3RI and School of Mathematics, University of Southampton, Southampton, UK; [b]Novartis Pharma AG, Basel, Switzerland; [c]Department of Statistics and Operations Technology, University of Denver, Denver, CO, USA; [d]Department of Statistics, Chulalongkorn University, Bangkok, Thailand

ABSTRACT

We consider the computation of the critical constants c of a p-content $(1-\alpha)$-confidence tolerance ellipse for bivariate and trivariate normal populations which are probably the most used multivariate normal distributions in applications. There are only approximation methods available in the statistical literature for computing c. Without knowing the accurate value of c, it is impossible to assess the accuracy of the approximation methods. In this paper, a new method is given to compute c. Although the method is also based on Monte-Carlo simulation, it allows c to be computed as accurate as required if the number of simulation replications is sufficiently large. The `R` codes provided allow easy implementation of the method. Furthermore, the method allows the accuracy of the available approximation methods for bivariate and trivariate normal distributions to be assessed.

ARTICLE HISTORY
Received 6 April 2022
Accepted 7 May 2022

KEYWORDS
Multivariate normal distribution; Monte-Carlo simulation; tolerance regions

1. Introduction

Tolerance intervals/regions were introduced first by Wilks [1]. Refs. [2–6] are excellent references on tolerance intervals/regions. The R package `tolerance` of Young [7,8] allows the computation of many tolerance intervals/regions. The focus of this paper is on the tolerance regions for a multivariate normal population (cf. Ref. [5], Chapter 9).

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample from a population having a q-dimensional multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$, where both $\boldsymbol{\mu}$ and Σ are unknown. Denote the sample mean $\bar{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$ and the sample variance–covariance matrix $\mathbf{V} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'/(n-1)$. It is well known (cf. Ref. [9]) that $\bar{\mathbf{x}} \sim N(\boldsymbol{\mu}, \Sigma/n)$, $(n-1)\mathbf{V} \sim W(\Sigma, n-1)$, and $\bar{\mathbf{x}}$ and $\mathbf{V}$ are independent; here $W(\Sigma, n-1)$ denotes the Wishart distribution (cf. Ref. [[9], p. 255]).

Let $\mathbf{y}$ be a future sample observation, independent of the random sample $\mathbf{X}$, from the same population $N(\boldsymbol{\mu}, \Sigma)$. The p-content $(1-\alpha)$-confidence ellipsoidal tolerance region for the population is a subset in R^q of the form

$$R_c(\mathbf{X}) = \left\{\mathbf{x} \in R^q : (\mathbf{x} - \bar{\mathbf{x}})' V^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \le c\right\}, \tag{1}$$

CONTACT W. Liu w.liu@maths.soton.ac.uk S3RI and School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK

where c is a critical constant chosen so that

$$P_{\bar{\mathbf{x}},\mathbf{V}}\left\{P_{\mathbf{y}|\bar{\mathbf{x}},\mathbf{V}}\{\mathbf{y} \in R_c(\mathbf{X})\} \geq p\right\} = 1-\alpha, \tag{2}$$

where the probability $P_{\mathbf{y}|\bar{\mathbf{x}},\mathbf{V}}\{\cdot\}$ is calculated with respect to $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$ conditional on $(\bar{\mathbf{x}}, \mathbf{V})$, and $P_{\bar{\mathbf{x}},\mathbf{V}}\{\cdot\}$ is calculated with respect to $(\bar{\mathbf{x}}, \mathbf{V})$ whose distributions are specified in the last paragraph; see Ref. [5], Section 9.2 for more details.

To the best of our knowledge, the only available p-content $(1-\alpha)$-confidence tolerance region specifically for $N(\boldsymbol{\mu}, \Sigma)$ with $q \geq 2$ is of the ellipsoidal form in (1). The central ellipsoidal tolerance region of Dong and Mathew [10] is conservative, i.e. the probability on the left side of Equation (2) is strictly larger than $1-\alpha$.

One key factor in the choice of the $R_c(\mathbf{X})$ in (1) is that the probability expression on the left side of Equation (2) does not depend on the unknown $\boldsymbol{\mu}$ or Σ, which is clear from (3). But even in this case, the computation of c is very challenging; see, for example, Refs. [5,11–13]. All these methods involve approximation; more details are given in Section 2. Without knowing the accurate value of c, it is impossible to assess the accuracy of these approximation methods.

In this paper, a new method is given to compute accurately the critical constant c in (1) for $q = 2$ or 3, i.e. for the bivariate or trivariate normal distributions. Bivariate normal distribution is the gateway to a general multivariate normal distribution and probably the most frequently used multivariate normal distribution in applications; see, e.g. Ref. [[14], p. 251–309]. Due to the importance of bivariate and trivariate normal distributions and a central role played by ellipsoidal tolerance regions for multivariate normal distributions, the new method is an important contribution to the construction of ellipsoidal tolerance regions for bivariate and trivariate normal distributions. The new method can easily be used with the `R` codes provided. It also allows the accuracy of the approximation methods for at least bivariate and trivariate normal distributions to be assessed.

The paper is organized as follows. Section 2 gives the new method for computing the critical constant c. Section 3 studies the computational accuracy of the new method and assesses the accuracy of two approximation methods. An illustrative example is given in Section 4. Finally, the paper closes with a brief discussion in Section 5.

2. Computation of the critical constant c

First note that

$$(\mathbf{y}-\bar{\mathbf{x}})'V^{-1}(\mathbf{y}-\bar{\mathbf{x}}) = (n-1)\left(\mathbf{u}-\frac{\mathbf{z}}{\sqrt{n}}\right)'\Lambda^{-1}\left(\mathbf{u}-\frac{\mathbf{z}}{\sqrt{n}}\right),$$

where $\mathbf{u} = \Sigma^{-(1/2)}(\mathbf{y}-\boldsymbol{\mu}) \sim N(\mathbf{0}, I)$, $\mathbf{z} = \sqrt{n}\Sigma^{-(1/2)}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \sim N(\mathbf{0}, I)$ and $\Lambda = \Sigma^{-(1/2)}(n-1)\mathbf{V}\Sigma^{-(1/2)} \sim W(I, n-1)$ are independent. It follows immediately that Equation (2) can be written as

$$P_{\mathbf{z},\Lambda}\left\{P_{\mathbf{u}|\mathbf{z},\Lambda}\left\{\left(\mathbf{u}-\frac{\mathbf{z}}{\sqrt{n}}\right)'\Lambda^{-1}\left(\mathbf{u}-\frac{\mathbf{z}}{\sqrt{n}}\right) \leq \frac{c}{n-1}\right\} \geq p\right\} = 1-\alpha, \tag{3}$$

where the probability $P_{\mathbf{u}|\mathbf{z},\Lambda}\{\cdot\}$ is calculated with respect to $\mathbf{u} \sim N(\mathbf{0}, I)$ conditional on $(\mathbf{z}, \Lambda)$, and $P_{\mathbf{z},\Lambda}\{\cdot\}$ is calculated with respect to $(\mathbf{z}, \Lambda)$.

We use the following Monte-Carlo simulation method to compute the c from Equation (3).

- Step 1: simulate one $(\mathbf{z}, \Lambda)$ from the distributions given in the last paragraph: $\mathbf{z} \sim N(\mathbf{0}, I)$, $\Lambda \sim W(I, n-1)$, and $\mathbf{z}$ and Λ are independent.
- Step 2: given the simulated $(\mathbf{z}, \Lambda)$, solve c from the equation

$$P_{\mathbf{u}|\mathbf{z},\Lambda}\left\{\left(\mathbf{u}-\frac{\mathbf{z}}{\sqrt{n}}\right)'\Lambda^{-1}\left(\mathbf{u}-\frac{\mathbf{z}}{\sqrt{n}}\right)\le\frac{c}{n-1}\right\}=p. \tag{4}$$

- *Step 3*: repeat Steps 1 and 2 for a large number of L times, say $L = 100,000$, to get the corresponding $c_1, \ldots, c_L$; order these values as $c_{[1]} \le \cdots \le c_{[L]}$ and use $c_{[\langle(1-\alpha)L\rangle]}$ as the c that solves Equation (3). Here, $\langle(1-\alpha)L\rangle$ denotes the integer part of $(1-\alpha)L$.

The key idea of this method is to use the sample quantile $c_{[\langle(1-\alpha)L\rangle]}$ to approximate the population quantile c, and it is well known (cf. Ref. [15]) that $c_{[\langle(1-\alpha)L\rangle]}$ converges almost surely to c as $L \to \infty$. Hence, $c_{[\langle(1-\alpha)L\rangle]}$ can be regarded as accurate so long as the number of simulations L is large enough. The accuracy of c for a finite L can be assessed in different ways; see, e.g. Ref. [[16], p. 243–245] and Section 3.

If one uses, in *Step 2*, an inner loop of Monte-Carlo simulation to compute an approximation c_i to the c that solves Equation (4), which is used in Ref. [11], then the computation is very time-consuming and the resultant critical constant has a large variance and so is not accurate; see Ref. [[5], Section 9.4, p. 232–233]. Note from the expression (5) that the probability in (4) is related to the cumulative distribution function of a linear combination of independent non-central chi-square random variables. Krishnamoorthy and Mondal [12] have proposed to compute an approximation c_i to the value of c in *Step 2* using the analytical approximation of Imhof [17] to the cumulative distribution function of a linear combination of independent non-central chi-square random variables. This method will be denoted as KM hereafter. Various other analytical approximations have also been proposed, but the KM method is recommended in the `R` package `tolerance` of Ref. [7]. Mbodj and Mathew [13] have proposed to compute another approximation c_i to the value of c in *Step 2* using the analytical approximation of Liu et al.[18] instead of the analytical approximation of Imhof [17]. This method will be denoted as MM henceforth. While the computation times of both KM and MM are reduced substantially due to the analytical approximations used, the resultant critical constant c is not guaranteed to converge to the exact critical constant even when the number of simulations $L \to \infty$. There is no method to compute the exact critical value c for $q \ge 2$ and so little is known about the accuracy of the KM and MM methods even when $q = 2$ and L is very large, say $L = 1,000,000$. Our new approach given below is to compute the c in *Step 2* by solving Equation (4) numerically.

For *Step 1*, one can easily simulate one $\mathbf{z} \sim N(\mathbf{0}, I)$. To simulate one Λ from $W(I, n-1)$, we use the Bartlett decomposition (cf. Ref. [19] and the references therein) in the following way. *Step (1a)*: form a $q \times q$ matrix G with all its lower triangle elements $g_{ij} = 0$ $(i>j)$, all its upper triangle elements g_{ij} $(i<j)$ being independent $N(0,1)$ random variables and its ith diagonal element $g_{ii} = \sqrt{\chi^2_{n-i}}$ $(i = 1, \ldots, q)$ where $\chi^2_{n-1}, \ldots, \chi^2_{n-q}$ are independent chi-square random variables independent of the $N(0,1)$ random variables g_{ij}. *Step (1b)*: set $\Lambda = G'G$ which has the required Wishart distribution $W(I, n-1)$.

For *Step 2*, we need to compute the probability in (4) for a given $c>0$ and the simulated $(\mathbf{z}, \Lambda)$. Let Q be an orthogonal matrix such that $\Lambda^{-1} = Q'\text{diag}\{\lambda_1, \ldots, \lambda_q\}Q$. Then the probability in (4) can be written as

$$P_{\mathbf{v}}\left\{(\mathbf{v}-\mathbf{w})'\ \text{diag}\{\lambda_1, \ldots, \lambda_q\}\,(\mathbf{v}-\mathbf{w}) \le \frac{c}{n-1}\right\}, \tag{5}$$

where $\mathbf{v} = Q\mathbf{u} \sim N(\mathbf{0}, I)$ and $\mathbf{w} = Q\mathbf{z}/\sqrt{n} = (w_1, \ldots, w_q)'$.

For $q = 2$, direct algebraic manipulation shows that the probability in (5) can be written as

$$\int_{w_1-\sqrt{c/[(n-1)\lambda_1]}}^{w_1+\sqrt{c/[(n-1)\lambda_1]}} \phi(v_1)\Phi\left(w_2 \pm \sqrt{\left[c/(n-1) - \lambda_1(v_1-w_1)^2\right]/\lambda_2}\right) \mathrm{d}v_1, \tag{6}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the pdf and cdf of $N(0,1)$, and $\Phi(a \pm b) = \Phi(a+b) - \Phi(a-b)$. For a given c, this probability can easily be computed from expression (6) using one-dimensional numerical quadrature, for example, the `R` function `integrate` is used in our `R` codes. Furthermore, note this probability is clearly monotone increasing in c. Hence the unique solution c of Equation (4) can easily be computed using a numerical searching algorithm, for example, the bi-section method is used in our codes. From our experience, the computation of one c in *Step 2* takes only a small fraction of a second on an ordinary PC; see more details on computation times in the next section.

Similarly, for $q = 3$, the probability in (5) can be written as

$$\int_{w_1-\sqrt{c/[(n-1)\lambda_1]}}^{w_1+\sqrt{c/[(n-1)\lambda_1]}} \phi(v_1) \int_{w_2-\sqrt{\frac{1}{\lambda_2}[c/(n-1)-\lambda_1(v_1-w_1)^2]}}^{w_2+\sqrt{\frac{1}{\lambda_2}[c/(n-1)-\lambda_1(v_1-w_1)^2]}} \phi(v_2)$$
$$\Phi\left(w_3 \pm \sqrt{\frac{1}{\lambda_3}\left[c/(n-1) - \lambda_1(v_1-w_1)^2 - \lambda_2(v_2-w_2)^2\right]}\right) \mathrm{d}v_2\,\mathrm{d}v_1. \tag{7}$$

For a given c, this probability can easily be computed from expression (7) using two-dimensional numerical quadrature. Hence, the unique solution c of Equation (4) can be computed in a similar way as for $q = 2$ above. From our experience, the computation of one c in *Step 2* still takes only a fraction of a second on the same PC, but about 5 times of the computation time for $q = 2$ due to the two-dimensional numerical quadrature involved; see more details in the next section.

For $q = 4$, it is clear that the probability in (5) can be expressed as a three-dimensional integral. As we need to solve c from Equation (4) involving three-dimensional numerical quadrature, for each of the L simulation replications, this takes too long time on an ordinary PC. Hence, we focus on $q = 2$ and 3 in this paper.

3. Assessment of the methods

It is clear from (3) that the critical constant c depends only on $1-\alpha$, p, n and q.

3.1. For bivariate normal distribution with $q = 2$

For $q = 2$, the computation of one c based on $L = 1,000,000$ simulations takes about 3754 s (or 63 min) on an ordinary Window's PC (Intel(R) Core(TM) i5-8265U CPU

with 1.60 GHz, 1.80 GHz, RAM 8.0 GB). The computation based on $L = 100,000$ simulations takes about 375 s. This agrees with the intuition that the computation time of one c is proportional to the number of simulations L. One can download the R code `CritConst2F.R` from http://www.personal.soton.ac.uk/wl/Bi-Tri-Norm-Toler-Ellip/ for computing the c for given $q = 2, 1-\alpha, p$ and n.

To assess the Monte-Carlo simulation accuracy of the c based on L simulations for given $(p, 1-\alpha, n)$, we have computed 20 values of c based on $L = 1,000,000$ simulations for given $(p, 1-\alpha, n) = (0.90, 0.95, 30)$ using 20 different random seeds, which are given by 7.43677, 7.43439, 7.42948, 7.43499, 7.43038, 7.43930, 7.43529, 7.43529, 7.43339, 7.43231, 7.43648, 7.42651, 7.43350, 7.43960, 7.43365, 7.43231, 7.43127, 7.43514, 7.43871, and 7.43231. These 20 values can be thought as a random sample from the population of possible c values based on $L = 1,000,000$ simulations for given $(p, 1-\alpha, n) = (0.90, 0.95, 30)$. The sample mean and standard deviation are given by 7.434 and 0.0033, respectively. Hence, the possible value of c is most likely within the range $7.434 \pm 3 \times 0.0033$ from the three-sigma rule. This indicates that the c value computed using $L = 1,000,000$ in this case is most likely to be accurate to ± 0.0099 in different simulation runs.

We have also tried $L = 100,000$ and the corresponding sample standard deviation based on 20 c-values (using 20 different random seeds) is close to $\sqrt{10} \times 0.0033 = 0.0104$. This agrees with the known result (cf. Ref. [15] or [20]) that the standard deviation of the sample quantile is proportional to $1/\sqrt{L}$ as $L \to \infty$.

It must be emphasized that the Monte-Carlo simulation accuracy can be improved by increasing the value of L. When the value of L becomes large enough, then the c value of our method approaches the exact critical constant required. In this sense, our method is exact.

From the various other methods proposed in the statistical literature for computing c, the KM method of Krishnamoorthy and Mondal [12] is the method that is recommended by Krishnamoorthy and Mathew [[5], Chapter 9, Sections 9.3 and 9.4] and Young's `R` package `tolerance`. The c-values of the KM method are tabulated in Ref. [5], Appendix B, Table B16, using $L = 100,000$. The KM method is implemented in the function `mvtol.region` of Young's `R` package `tolerance` which allows a user to set the value of L. As pointed out in Section 2, the KM method computes an analytical approximation, rather than an exact numerical solution as our method, to the value of c from Equation (4). Mbodj and Mathew [13] have observed that a method that involves an analytical approximation such as the KM method is unlikely to work for all values of $(p, 1-\alpha, n)$. The MM method of Mbodj and Mathew [13] is similar to the KM method but using a different analytical approximation. Since our method is accurate for a sufficiently large L, we can assess the accuracy of KM and MM methods for $q = 2$.

To get some idea about the Monte-Carlo simulation accuracy of the c's of KM and MM methods based on L simulations, we have also computed 20 values of c using KM and MM methods based on $L = 1,000,000$ simulations for given $(p, 1-\alpha, n) = (0.90, 0.95, 30)$ using 20 different random seeds. Based on these 20 values, we have the sample (mean, standard deviation) given by (7.485, 0.0033) and (7.483, 0.0033) for KM and MM methods, respectively. Hence, the simulation accuracies of the c's of the new, KM and MM methods based on $L = 1,000,000$ simulations are approximately the same. It is also noteworthy that the sample means of KM and MM are very close to each other, and only slightly larger than that of the new method.

Table 1. The c of the new (top), KM (middle) and MM (bottom) methods for $q = 2$.

		$p = 0.90$			$p = 0.95$			$p = 0.99$		
		$1-\alpha$			$1-\alpha$			$1-\alpha$		
n		0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
5	New	41.131	67.490	203.831	57.003	93.896	284.790	95.387	157.704	480.793
	KM	41.195	67.333	204.042	56.991	93.620	284.921	95.320	157.610	481.706
	MM	41.221	67.581	203.847	57.063	93.945	284.804	95.337	157.664	480.781
7	New	19.783	27.039	53.962	27.190	37.377	75.311	45.197	62.507	126.896
	KM	19.856	27.171	54.162	27.227	37.510	75.470	45.107	62.599	127.121
	MM	19.878	27.136	54.030	27.287	37.467	75.348	45.171	62.469	126.808
10	New	12.586	15.594	24.723	17.109	21.334	34.245	28.157	35.384	57.378
	KM	12.701	15.711	24.845	17.249	21.469	34.343	28.207	35.420	57.467
	MM	12.679	15.692	24.832	17.214	21.443	34.344	28.154	35.367	57.346
15	New	9.194	10.630	14.422	12.345	14.352	19.711	20.051	23.489	32.719
	KM	9.274	10.746	14.582	12.463	14.512	19.899	20.101	23.579	32.843
	MM	9.266	10.716	14.528	12.445	14.461	19.834	20.066	23.502	32.712
30	New	6.832	7.433	8.786	9.036	9.858	11.728	14.350	15.738	18.943
	KM	6.872	7.482	8.861	9.100	9.937	11.849	14.378	15.776	18.991
	MM	6.873	7.486	8.854	9.103	9.939	11.834	14.378	15.766	18.969
50	New	6.046	6.419	7.212	7.941	8.442	9.520	12.457	13.283	15.103
	KM	6.068	6.446	7.246	7.981	8.491	9.580	12.479	13.312	15.109
	MM	6.071	6.448	7.253	7.985	8.493	9.592	12.482	13.311	15.135

It must be pointed out that even if there is no Monte-Carlo simulation error (i.e. when $L \to \infty$), the c's of KM and MM methods may not be equal to the exact critical constant that solves Equation (3) due to the analytical approximations used. This is a fundamental difference between the new method and KM and MM methods.

Next, we compare the c-values of KM, MM and new methods for various configurations of $(p, 1-\alpha, n)$ to assess to what extend they differ by using $L = 1,000,000$, so that the simulation errors of all three methods are quite small and so negligible as observed above. Table 1 provides the c-values for small and moderate sample sizes n, the situations identified by Mbodj and Mathew [13] that KM and MM may give less accurate critical constants due to the approximations used.

It is clear from Table 1 (we have also tried larger sample sizes, e.g. $n = 150$) that the c-values of KM and MM methods are very close to each other, and only slightly larger than that of the new method. This shows that both KM and MM methods produce quite accurate approximations to the exact critical constant c and there is little difference between KM and MM methods for the bivariate normal distribution.

3.2. For trivariate normal distribution with $q = 3$

The `R` code `CritConst3.R` for computing the critical constant c for given $1-\alpha$, p and n (with $q = 3$) is available at http://www.personal.soton.ac.uk/wl/Bi-Tri-Norm-Toler-Ellip/. The computation of one c based on $L = 100,000$ simulations takes about 1750 s on the same PC, which is about five times of the computation time for $q = 2$. This is because double numerical quadrature is required for $q = 3$ while only single numerical quadrature is used for $q = 2$.

To get some idea about the simulation accuracy of the c based on L simulations for given $(p, 1-\alpha, n)$, we have computed 20 values of c based on $L = 100,000$ simulations for $(p, 1-\alpha, n) = (0.90, 0.95, 30)$ using 20 different random seeds, which have the sample

mean and standard deviation given by 10.182 and 0.0125, respectively. Hence, the possible values of c are most likely within the range $10.182 \pm 3 \times 0.0125$ from the three-sigma rule. If $L = 1,000,000$ is used, then the sample standard deviation is expected to be about $0.0125/\sqrt{10} = 0.0040$.

For the simulation accuracy of the c's of KM and MM methods, we have also computed 20 values of c using KM and MM methods based on $L = 100,000$ simulations for $(p, 1-\alpha, n) = (0.90, 0.95, 30)$ using 20 different random seeds. Based on these 20 values, we have the sample (mean, standard deviation) given by $(10.280, 0.0157)$ and $(10.282, 0.0118)$ for KM and MM methods, respectively. Hence, the simulation accuracies of the c's of the new, KM and MM methods based on $L = 100,000$ simulations are approximately the same.

These results also indicate that the c-values of KM and MM methods are very close to each other and only slightly larger than the c value of the new method. This observation is also borne out by the other configurations of $(p, 1-\alpha, n)$ we have tried using $L = 100,000$.

4. Examples

In this section, we first use the Lumber data, originally from Ref. [[21], Table 5.6] and used also in Ref. [[5], Example 9.1, Table 9.2], to illustrate the tolerance ellipse. The data contain a sample of 30 observations on stiffness (X_1) and bending strength (X_2), in units of pounds per square inch, of a particular grade of lumber.

Some exploratory check of the bivariate normality assumption indicates that the normality assumption is tenable; see Ref. [[5], Example 9.1]. The summary statistics are

$$\bar{\mathbf{x}} = \begin{pmatrix} 1860 \\ 8354 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 124049.8 & 361673.4 \\ 361673.4 & 3486334.0 \end{pmatrix}.$$

For $(p, 1-\alpha, n) = (0.90, 0.95, 30)$ and $q = 2$, the critical constant c computed by our `R` code is 7.433 and so the tolerance ellipse is given by

$$R_c(\mathbf{X}) = \left\{ \mathbf{x} = (x_1, x_2)' : (\mathbf{x} - \bar{\mathbf{x}})' V^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \leq 7.433 \right\}. \tag{8}$$

This tolerance ellipse is plotted in Figure 1, with the $n = 30$ sample observations represented by the crosses, and the centre of the ellipse marked by a diamond and letter S. Hence, with confidence level $1 - \alpha = 0.95$ about the randomness in the sample of 30 observations used to construct this ellipse, at least $100p\% = 90\%$ of the $N(\boldsymbol{\mu}, \Sigma)$ population is included in this ellipse.

For Example 1 given in Ref. [[13], p.43] with $(p, 1-\alpha, n) = (0.95, 0.95, 30)$ and $q = 2$, the critical constant c is computed to be 9.858, 9.937 and 9.939 in Table 1 by the new, KM and MM methods, respectively, based on $L = 1,000,000$ simulations. There seems to be a typo with the c-value 8.94 given in that paper.

For Example 2 given in Ref. [[13], p. 44] with $(p, 1-\alpha, n) = (0.95, 0.95, 284)$ and $q = 3$, the critical constant c is computed to be 8.657, 8.669 and 8.670 by the new, KM and MM methods, respectively, based on $L = 100,000$ simulations. There seems to be a typo with the c-value 7.69 given in that paper.

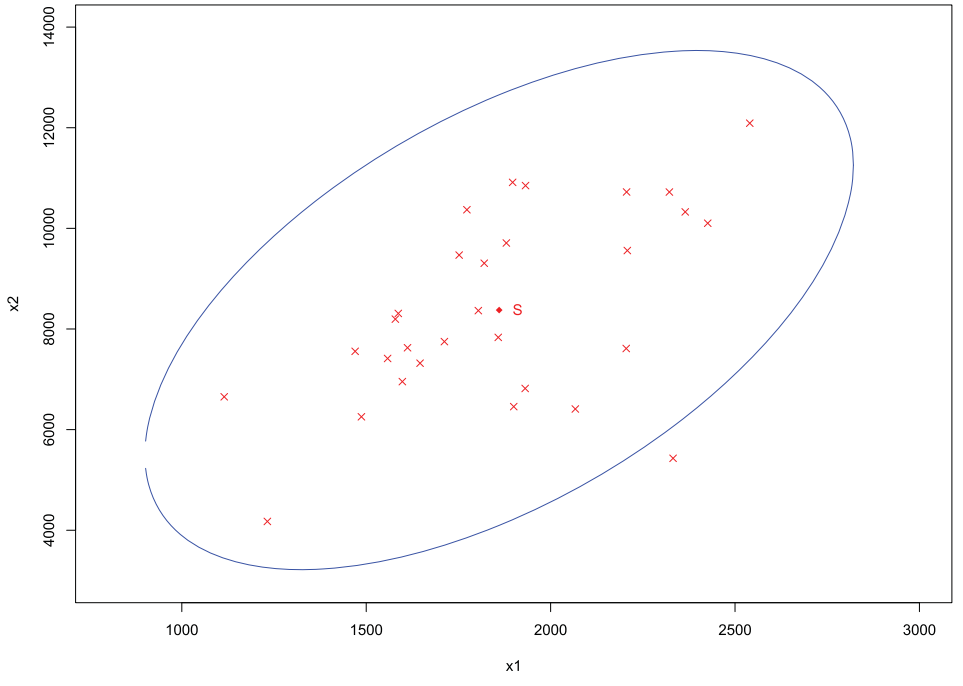

Figure 1. The tolerance ellipse and the observed sample data.

5. Discussion

Bivariate and trivariate normal distributions are probably the most frequently used multivariate distributions, and tolerance regions have wide applications (cf. Refs. [5,6]). Ellipsoidal tolerance regions play a special role for multivariate normal distributions since they do not depend on the unknown population mean or population covariance matrix. But even for the tolerance ellipses of bivariate or trivariate normal populations, only approximate methods are available in the literature for computing the critical constant c. Furthermore, without knowing the accurate value of c, it is impossible to judge the accuracy of the available approximation methods.

In this paper, a new method of computing c is proposed. This method is accurate in the sense that the critical constant c can be computed as close to the exact value as one requires if the number of simulations L is set to be sufficiently large. The available `R` code allows easy implementation of the new method.

This new method also allows one to assess the accuracy of the available approximation methods for bivariate and trivariate normal distributions. The conclusion from our numerical investigation is that the c-values of KM and MM methods are only slightly conservative and there is little difference between the c-values of KM and MM methods.

Computation of exact tolerance ellipses for multivariate normal distribution of dimension four or larger is more complicated but warrants further research.

Acknowledgments

We thank the two anonymous referees for very helpful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Wilks SS. Determination of sample sizes for setting tolerance limits. Ann Math Stat. 1941;12:91–96.
[2] Guttman I. Statistical tolerance regions: classical and Bayesian. London: Griffin; 1970.
[3] Guttman I. Tolerance regions. In: Kotz S, et al., editors. Encyclopedia of statistical sciences. 2nd ed. New York: Wiley; 2006. p. 8644–8659.
[4] Hahn G, Meeker WQ. Statistical intervals: a guide to practitioners. New York: Wiley; 1991.
[5] Krishnamoorthy K, Mathew T. Statistical tolerance regions – theory, applications, and computation. New York: Wiley; 2009.
[6] Meeker WQ, Hahn GJ, Escobar LA. Statistical intervals: a guide for practitioners and researchers. 2nd ed. New York: Wiley; 2017.
[7] Young DS. Tolerance: an r package for estimating tolerance intervals. J Stat Softw. 2010;36:1–39.
[8] Young DS. Normal tolerance interval procedures in the tolerance package. R J. 2016;8(2): 200–212.
[9] Anderson TW. An introduction to multivariate statistical analysis. 3rd ed. New York: Wiley; 2003.
[10] Dong X, Mathew T. Central tolerance regions and reference regions for multivariate normal population. J Multivar Anal. 2015;134:50–60.
[11] Krishnamoorthy K, Mathew T. Comparison of approximation methods for computing tolerance factors for a multivariate normal population. Technometrics. 1999;41:234–249.
[12] Krishnamoorthy K, Mondal S. Improved tolerance factors for multivariate normal distributions. Commun Stat Simul Comput. 2006;35:461–478.
[13] Mbodj M, Mathew T. Approximate ellipsoidal tolerance regions for multivariate normal populations. Stat Probab Lett. 2015;97:41–45.
[14] Kotz S, Balakrishnan N, Johnson N. Continuous multivariate distributions, volume 1: models and applications. 2nd ed. New York: Wiley; 2000.
[15] Serfling RJ. Approximation theorems of mathematical statistics. New York: Wiley; 1980.
[16] Liu W. Simultaneous inference in regression. New York: CRC Press; 2010.
[17] Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961;48:419–426.
[18] Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-definite quadratic forms in non central normal variables. Comput Stat Data Anal. 2009;53:853–856.
[19] Smith WB, Hocking RR. Algorithm AS 53: Wishart variate generator. J R Stat Soc Ser C. 1972;21(3):341–345.
[20] Liu W, Jamshidian M, Zhang Y, et al. Simulation-based simultaneous confidence bands in multiple linear regression with predictor variables constrained in intervals. J Comput Graph Stat. 2005;14(2):459–484.
[21] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 3rd ed. New Jersey: Prentice Hall; 1992.