Communications for Statistical Applications and Methods
2017, Vol. 24, No. 6, 641–649

https://doi.org/10.29220/CSAM.2017.24.6.641
Print ISSN 2287-7843 / Online ISSN 2383-4757

Multivariate confidence region using quantile vectors

Chong Sun Hong[1,a], Hong Il Kim[a]

[a]Department of Statistics, Sungkyunkwan University, Korea

Abstract

Multivariate confidence regions were defined using a chi-square distribution function under a normal assumption and were represented with ellipse and ellipsoid types of bivariate and trivariate normal distribution functions. In this work, an alternative confidence region using the multivariate quantile vectors is proposed to define the normal distribution as well as any other distributions. These lower and upper bounds could be obtained using quantile vectors, and then the appropriate region between two bounds is referred to as the quantile confidence region. It notes that the upper and lower bounds of the bivariate and trivariate quantile confidence regions are represented as a curve and surface shapes, respectively. The quantile confidence region is obtained for various types of distribution functions that are both symmetric and asymmetric distribution functions. Then, its coverage rate is also calculated and compared. Therefore, we conclude that the quantile confidence region will be useful for the analysis of multivariate data, since it is found to have better coverage rates, even for asymmetric distributions.

Keywords: coverage rate, ellipse, ellipsoid, quantile vector

1. Introduction

For multivariate data with dimensionality equal to or more than two, the multivariate confidence region (MCR) of the population mean vector μ is defined with the theory that the well-known statistic, $(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu)$, follows a chi-squared distribution function with some degree of freedom. This is explained under the assumption that the multivariate date follows the normal distribution function (Chew, 1966; Fan and Zhang, 2000; Frank, 1966; Johnson and Whichern, 2002; Sun and Loader, 1994). The $(1 - \alpha)$ confidence regions for the bivariate distribution function are represented with circular or elliptical shapes with respect to values of the correlation coefficient ρ, and the confidence regions for the trivariate distribution function are expressed with spherical or ellipsoid shapes with respect to the types of variance and covariance matrix.

In the real world, most data does not satisfy the normality assumption. Therefore, it is not easy to define and obtain the MCR of the population mean vector frequently (see Asgharzadeh and Abdi (2011) for more detail). In this work, an alternative confidence region using multivariate quantile vectors is proposed to define the normal distribution as well as any other distribution functions.

When the multivariate random vector $\underline{Z} = (Z_1, \ldots, Z_k)$ is considered, Hong *et al.* (2016) proposed that for a given $\alpha \in (0, 1)$, the multivariate α quantile vector $\underline{z_\alpha} = (z_1, \ldots, z_k)$ is defined as follows:

1. The cumulative distribution function (cdf) of any point in the multivariate α quantile vector $\underline{z_\alpha}$ has the same value. For example, consider the bivariate cdf graph whose horizontal plane is represented

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

Published 30 November 2017 / journal homepage: http://csam.or.kr

as (Z_1, Z_2) coordinates, and vertical axis is for cdf values. Then one can select the $\underline{z_\alpha} = (z_1, z_2)$ coordinates corresponding to the α cdf value, which are parallel to the horizontal plane and represented as the curve. Any point in this curve has an equivalent value of the cumulative distribution function that also notes trivariate quantile vectors represented as the surface shape.

2. The probability for the upper region, R_α^k, of the multivariate α quantile vector $\underline{z_\alpha}$ is equal to $1 - \alpha$. That is,

$$P\left[\underline{Z} = (Z_1, \ldots, Z_k) \in R_\alpha^k\right] = \int \cdots \iint_{(z_1,\ldots,z_k)\in R_\alpha^k} dF(z_1, \ldots, z_k) = 1 - \alpha,$$

   where $F(\,\cdot\,, \ldots, \,\cdot\,)$ is a k-variate distribution function.

These lower and upper bounds of an alternative confidence region could be obtained by using the quantile vectors, and then the appropriate region between two bounds is denoted as the quantile confidence region (QCR). The existing ellipse and ellipsoid type of confidence region is now referred to as MCR.

Both the MCR and QCR are obtained for various types of distribution functions that are both symmetric and asymmetric distribution functions. Then, we will compare the coverage rates of the MCR and QCR.

In Section 2, the procedures to obtain the QCR are explained in terms of the multivariate quantile vectors proposed by Hong *et al.* (2016). Several QCRs are demonstrated for bivariate and trivariate standard normal distribution functions with some values of correlation coefficients. To discuss the properties between the QCR and MCR, the coverage rates of the MCR and OCR are obtained and compared for multivariate normal distribution functions with various values of correlation coefficients in Section 3; in addition, the corresponding coverage rates for both the MCR and QCR are calculated in Section 4 and discussed for multivariate symmetric distribution functions and asymmetric distribution functions in terms of many kinds of variance and covariance matrix types. Finally, Section 5 summarizes the results of this study and discusses further research.

2. Quantile confidence region

The $1 - \alpha$ QCR is derived with the following three steps using the multivariate quantile vectors of Hong *et al.* (2016).

1. For a given $\alpha \in (0, 1)$, the lower and upper bounds in k-variate random sample are obtained from $\alpha/2$ and $1-\alpha/2$ quantile vectors, $\underline{z_{\alpha/2}}$ and $\underline{z_{1-\alpha/2}}$, respectively. For a bivariate distribution function, $\underline{z_{\alpha/2}}$ and $\underline{z_{1-\alpha/2}}$ can be obtained like the left one in Figure 1, and the right graph in Figure 1 is the corresponding two sets of vectors drawn on a two dimensional plan.

2. For any positive constant value Δ and point $(z_{1,\alpha/2}, \ldots, z_{k,\alpha/2})$ in multivariate α quantile vector $\underline{z_\alpha}$, we find the smallest $z_{i0,\alpha/2}$ to satisfy

$$\int_{-\infty}^{z_{1,\frac{\alpha}{2}}} \cdots \int_{-\infty}^{z_{i0,\frac{\alpha}{2}}} \cdots \int_{-\infty}^{z_{k,\frac{\alpha}{2}}} dF(z_1, \ldots, z_k) = \int_{-\infty}^{z_{1,\frac{\alpha}{2}}} \cdots \int_{-\infty}^{z_{i0,\frac{\alpha}{2}}+\Delta} \cdots \int_{-\infty}^{z_{k,\frac{\alpha}{2}}} dF(z_1, \ldots, z_k),$$

   for each $i = 1, \ldots, k$. Then one can determine the end point vector $(z_{10,\alpha/2}, \ldots, z_{i0,\alpha/2}, \ldots, z_{k0,\alpha/2})$ in $\underline{z_\alpha}$ for all i. We may determine that the lower bound, $\{\underline{Z_{L,\alpha}}\}$, of the α QCR is defined as $\{(z_{1,\alpha/2}, \ldots, z_{k,\alpha/2})^T\}$, where each value $z_{i,\alpha/2} \le z_{i0,\alpha/2}$ for all i.

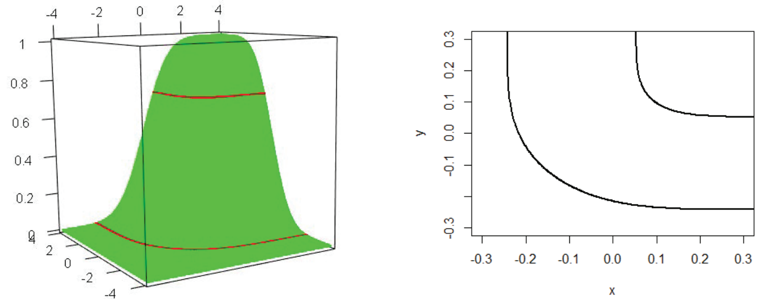

Figure 1: *Two bivariate quantile vectors.*

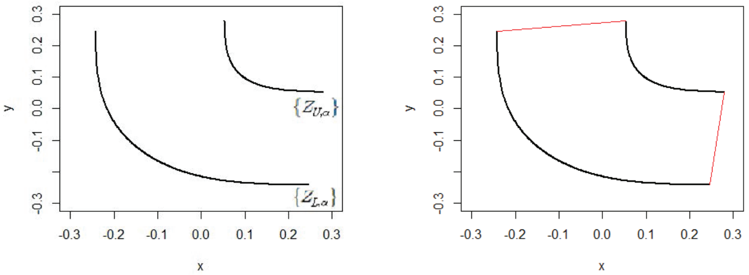

Figure 2: *Second and third steps for bivariate quantile confidence region.*

And for any point $(z_{1,1-\alpha/2},\ldots,z_{k,1-\alpha/2})$ in multivariate $1-\alpha$ quantile vector $z_{1-\alpha}$, we obtain the smallest $z_{i0,1-\alpha/2}$ $(i=1,\ldots,k)$ to satisfy

$$\int_{-\infty}^{z_{1,1-\frac{\alpha}{2}}}\cdots\int_{-\infty}^{z_{i0,1-\frac{\alpha}{2}}}\cdots\int_{-\infty}^{z_{k,1-\frac{\alpha}{2}}} dF(z_1,\ldots,z_k) = \int_{-\infty}^{z_{1,1-\frac{\alpha}{2}}}\cdots\int_{-\infty}^{z_{i0,1-\frac{\alpha}{2}}+\Delta}\cdots\int_{-\infty}^{z_{k,1-\frac{\alpha}{2}}} dF(z_1,\ldots,z_k).$$

For all $i=1,\ldots,k$, one can determine the end point vector $(z_{10,1-\alpha/2},\ldots,z_{i0,1-\alpha/2},\ldots,z_{k0,1-\alpha/2})$ in $z_{1-\alpha}$. Then the upper bound, $\{\underline{Z_{U,\alpha}}\}$, of the α QCR is defined as $\{(z_{1,1-\alpha/2},\ldots,\ z_{k,1-\alpha/2})^T\}$, where each value $z_{i,1-\alpha/2}\le z_{i0,1-\alpha/2}$ for all i.

That is, the lower bound, $\{\underline{Z_{L,\alpha}}\}$, and upper bound, $\{\underline{Z_{U,\alpha}}\}$, of the α QCR can be explained as the truncated bounds based on the end point vector $(z_{10,\alpha/2},\ldots,z_{i0,\alpha/2},\ \ldots,z_{k0,\alpha/2})$, and $(z_{10,1-\alpha/2},\ldots,z_{i0,1-\alpha/2},\ldots,z_{k0,1-\alpha/2})$, respectively. For a bivariate distribution function, two truncated bounds of the bivariate QCR look like the left one in Figure 2.

3. The multivariate $1-\alpha$ QCR are defined as the space between the lower and upper bounds $\{\underline{Z_{L,\alpha}}\}=\{(z_{1,\alpha/2},\ldots,z_{k,\alpha/2})^T\}$, and $\{\underline{Z_{U,\alpha}}\}=\{(z_{1,1-\alpha/2},\ldots,z_{k,1-\alpha/2})^T\}$. For a bivariate distribution function, one can obtain the bivariate QCR like the right graph in Figure 2.

Two bivariate 0.90 QCR and MCR are illustrated in Figure 3 for a bivariate standard normal distribution function with the correlation coefficient $\rho=-0.9, 0.0$, and 0.9 .

In Figure 3, the dotted and solid lines are represented for the MCR and QCR, respectively. The shapes of the change can be seen in both MCR and QCR according to values of ρ. When ρ is 0, the MCR has a circular shape. When ρ is positive (or negative), the MCR has an elliptical shape with a slope of 45 (or -45), and the ellipse becomes tighter as absolute values of ρ increase to 0.9. However,

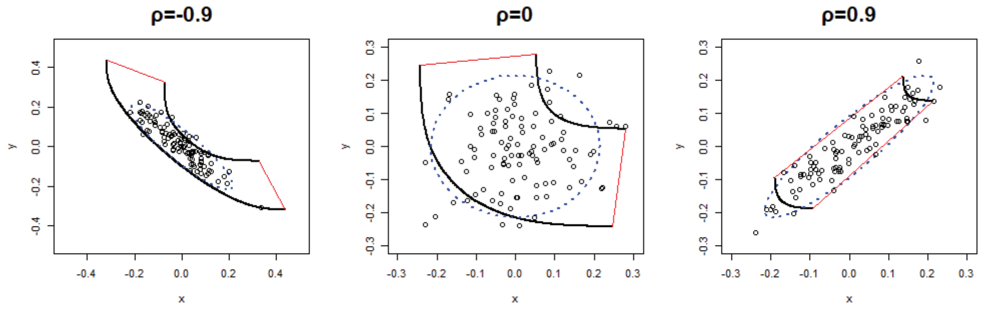

Figure 3: *The 90% bivariate quantile confidence region and multivariate confidence region.*

Table 1: Coverage rates for MCR and QCR

ρ	90% MCR		90% QCR		95% MCR		95% QCR	
	Coverage	SE	Coverage	SE	Coverage	SE	Coverage	SE
−0.90	900.0084	9.498	900.0365	9.421	949.9194	6.812	950.0056	6.885
−0.45	900.0373	9.446	900.2978	9.401	949.9356	6.826	949.7783	6.891
0.00	900.0440	9.637	900.1554	9.527	950.0254	6.887	950.0361	6.896
0.45	900.1064	9.496	900.0470	9.482	949.8844	6.872	949.7975	6.949
0.90	900.0331	9.630	899.8694	9.630	949.9895	6.903	950.0171	6.914

MCR = multivariate confidence region; QCR = quantile confidence region; SE = standard error.

the distance between the two upper and lower bounds of the QCR gradually increases, and the curve becomes shorter as ρ increases -0.9 to 0.9. When ρ is negative, the distance between these bounds is shorter, and the slope of the curve becomes lower compared to those for $\rho = 0$. When ρ is positive, the distance between these bounds is longer, and the slope of the curve becomes steeper compared to those for $\rho = 0$.

3. Coverage rate comparison in the normal case

In order to compare the QCR proposed in this work with MCR, the corresponding coverage rates are obtained for various kinds of distribution functions and some significant levels. First, under the bivariate and trivariate normal distribution assumption, the coverage rates are calculated under the simulation situations.

3.1. Bivariate normal

Consider the bivariate standard normal distributions with zero mean vector, unit variance, and with correlation coefficient ρ. Both the MCR and QCR are derived for these distribution functions with $\rho = -0.9$ to 0.9 in increments of 0.45 and $\alpha = 0.10, 0.05$. The sample means of size $n = 1{,}000$ are generated from the bivariate standard normal distributions with the variance-covariance matrix times $1/n$. With these 1,000 sample means, the corresponding coverage rates are calculated as the number belonging to each confidence region with 10,000 iterations.

The coverage rates and standard errors are obtained for $\rho = -0.9$ to 0.9 in increments of 0.45 with $\alpha = 0.10$ and 0.05 summarized in Table 1. From Table 1, it can be found that the coverage rates of the 90 and 95% MCR and QCR have similar values of 0.90 and 0.95, respectively. Each standard errors of the MCR and OCR also have very close values. Therefore, we can explore how the QCR has a very similar performance with MCR.

Table 2: Coverage rates for 90% MCR and QCR

ρ	90% MCR		90% QCR	
	Coverage	SE	Coverage	SE
−0.9	900.1214	9.348	900.0037	9.315
0.0	899.9087	9.488	900.6709	9.417
0.9	900.0452	9.630	900.1320	9.634

MCR = multivariate confidence region; QCR = quantile confidence region; SE = standard error.

3.2. Trivariate normal

Consider the standard trivariate normal distribution function with zero mean vector and simple variance-covariance matrix, such as $\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$.

Both the MCR and QCR are derived for these trivariate standard normal distribution functions with various ρ and α. The sample means of size $n = 1{,}000$ are generated from the trivariate standard normal distributions with the variance-covariance matrix Σ/n. With these 1,000 sample means, the corresponding coverage rates are calculated as the number belonging to each confidence regions with 10,000 iterations. Coverage rates and standard errors are obtained for $\rho = -0.9$, 0, 0.9 with $\alpha = 0.1$ and are summarized in Table 2.

Table 2 indicates that the coverage rates for the 90% MCR and QCR have almost the same values as 0.90, respectively. The standard errors of the MCR and OCR are also very close. These behaviors are similar to those of the bivariate normal distribution functions discussed in Section 3.1. Therefore, we can conclude that the QCR has very similar performance with the MCR for multivariate normal distribution functions.

4. Coverage rate comparison in non-normal case

The two confidence regions have similar coverage rates for normal distribution functions with various kinds of variance and covariance matrix and significant levels. Moreover, it can be assumed that the QCR has better advantages than the MCR if the QCR has better coverage rates than the MCR in non-normal distribution functions. In order to show that, we need to consider various kinds of the symmetric and asymmetric distribution functions. A large set of the sample means are generated and the coverage rates are calculated with analogous arguments in Section 3.

4.1. Symmetric mixture normal case

The following mixed normal distribution function is set as one multivariate distribution function that is not normal but symmetric with respect to the origin:

$$f(x) = 0.5f_1(x) + 0.5f_2(x),$$

where $f_1(x) = N\Big(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\Big)$, $f_2(x) = N\Big(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\Big)$.

Both MCR and QCR are derived for these normal mixtures with $\rho = -0.9$ to 0.9 in increments of 0.45 and $\alpha = 0.10, 0.05$. With these samples, the corresponding coverage rates are calculated as the number belonging to each confidence region with 10,000 iterations.

Figure 4 illustrates the bivariate 0.90 QCR and MCR for this symmetric normal mixture with the correlation coefficient $\rho = -0.9, 0.0$, and 0.9. In this normal mixture, the shapes of MCR and QCR in Figure 4 are different compared to the values of ρ. The shapes of the MCR become more slim as ρ increases to 0.9, and the pattern of this QCR becomes similar compared to QCR with normal

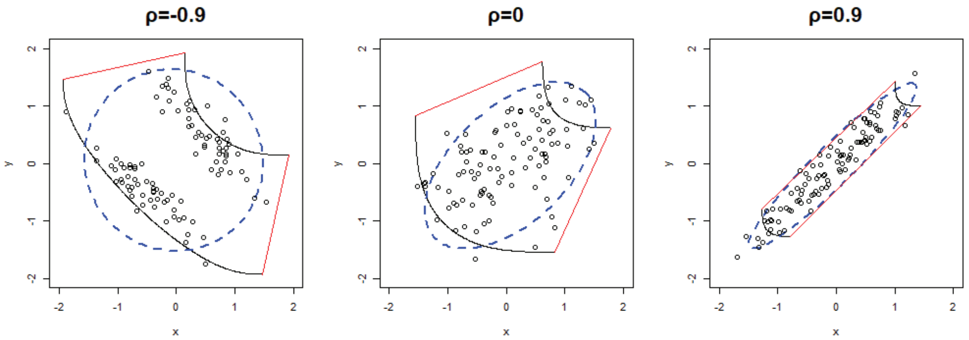

Figure 4: *The bivariate 0.90 multivariate confidence region and quantile confidence region for symmetric normal mixture.*

Table 3: Coverage rates for MCR and QCR

ρ	90% MCR		90% QCR		95% MCR		95% QCR	
	Coverage	SE	Coverage	SE	Coverage	SE	Coverage	SE
−0.90	938.1345	7.416	902.0384	9.613	973.7676	5.078	951.6531	6.712
−0.45	928.7346	8.289	901.1082	9.316	969.1177	5.435	951.3128	6.652
0.00	919.4502	8.551	900.7678	9.421	964.4892	5.734	950.7263	6.422
0.45	913.2595	8.469	900.1465	9.472	960.8215	6.129	950.2296	6.794
0.90	908.4519	9.303	900.0415	9.197	958.2762	6.292	949.8492	6.767

MCR = multivariate confidence region; QCR = quantile confidence region; SE = standard error.

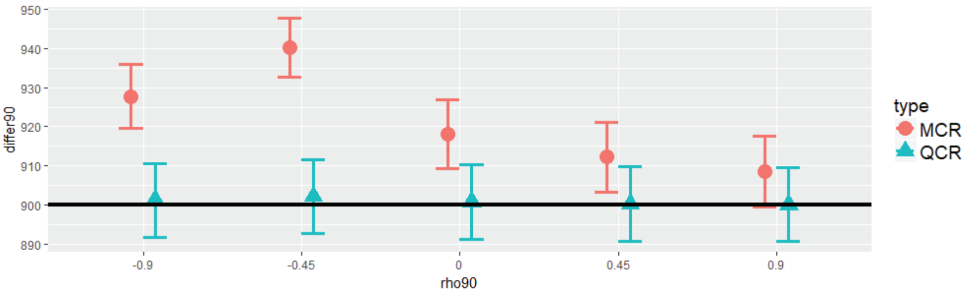

Figure 5: *Coverage rates for 90% MCR and QCR. MCR = multivariate confidence region; QCR = quantile confidence region.*

distribution functions. However, the distance between the upper and lower bounds in this normal mixture is slightly longer than those for normal distribution functions.

Table 3 and Figures 5, 6 show that the coverage rates of the 90% and 95% QCR have similar values as 0.90 and 0.95, respectively; however, the coverage rates of the 90% and 95% MCR have bigger values than 0.90 and 0.95. The standard errors of the MCR are slightly less than the QCR. With these phenomena, we could say that the QCR has better performance than the MCR for symmetric distribution functions which is non-normal distribution functions.

4.2. Asymmetric mixture normal case

Another mixture distribution function is considered as an asymmetric bivariate distribution function

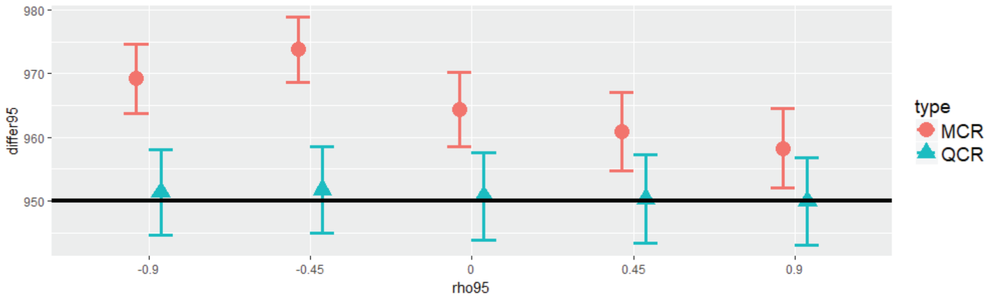

Figure 6: *Coverage rates for 95% MCR and QCR. MCR = multivariate confidence region; QCR = quantile confidence region.*

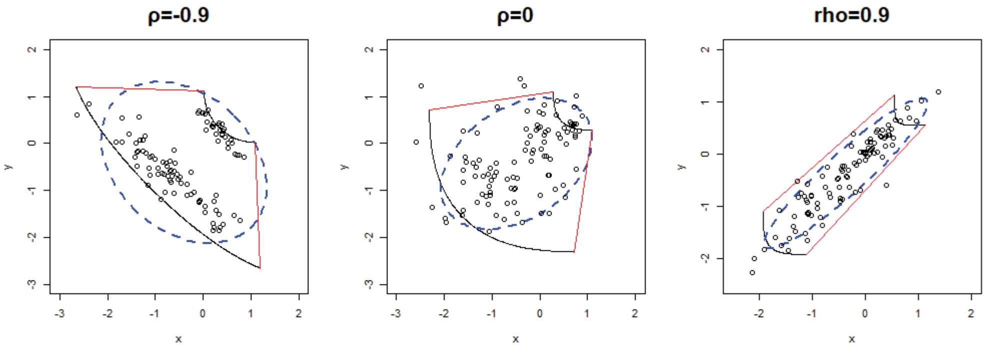

Figure 7: *The bivariate 0.90 multivariate confidence region and quantile confidence region for asymmetric normal mixture.*

which is normal as well as asymmetric about the mean:

$$f(x) = 0.7\, f_1(x) + 0.3\, f_2(x),$$

where $f_1(x) = N\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, $f_2(x) = N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.

Figure 7 illustrates the bivariate 0.90 QCR and MCR for this asymmetric normal mixture with the correlation coefficient $\rho = -0.9, 0.0$, and 0.9.

The shapes of the MCR in Figure 7 become much slimmer compared to those in Figure 4, and the patterns of the QCR in Figure 7 are a little different from those in Figure 4 for this asymmetric mixture as ρ increases to 0.9. It can be seen that each upper and lower bounds in this asymmetric mixture can have different curves. Table 3 shows that the coverage rates and standard errors are obtained for $\rho = -0.9$ to 0.9 in increments of 0.45 with $\alpha = 0.10$ and 0.05. Their means and confidence intervals of the coverage rates are represented in Figures 8 and 9.

Table 4 and Figures 8, 9 tell us that the coverage rates of the 90% and 95% QCR have similar values as 0.90 and 0.95, respectively, but their coverage rates of the MCR have smaller values than the corresponding values. The standard errors of the MCR are larger than those of the QCR. With these phenomena obtained in Section 4.1 and 4.2, we can conclude that the QCR has better performance than the MCR for symmetric and asymmetric distribution functions that include non-normal distribution functions.

Table 4: Coverage rates for MCR and QCR

ρ	90% MCR		90% QCR		95% MCR		95% QCR	
	Coverage	SE	Coverage	SE	Coverage	SE	Coverage	SE
−0.90	846.9411	10.741	902.1314	9.221	925.0911	8.047	954.5249	6.253
−0.45	828.8471	11.379	904.7747	9.378	905.0871	8.631	954.7135	6.124
0.00	835.3633	11.153	904.6439	9.311	903.8128	9.203	954.4201	6.143
0.45	841.2636	11.214	902.4423	9.218	905.3311	9.192	953.1651	6.754
0.90	846.1028	11.322	900.1011	9.139	907.4139	9.205	950.1371	6.398

MCR = multivariate confidence region; QCR = quantile confidence region; SE = standard error.

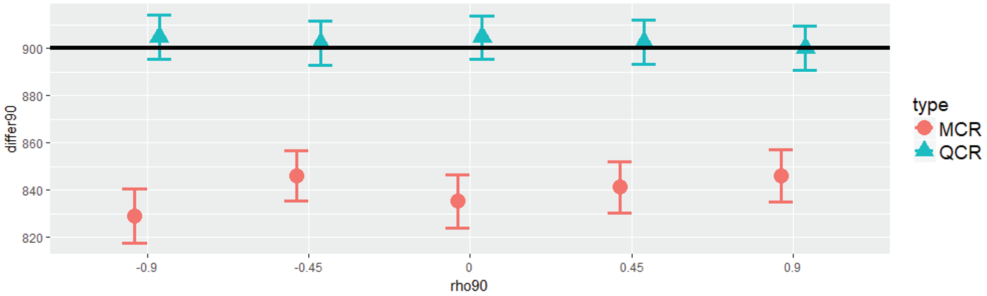

Figure 8: *Coverage rates for 90% MCR and QCR. MCR = multivariate confidence region; QCR = quantile confidence region.*

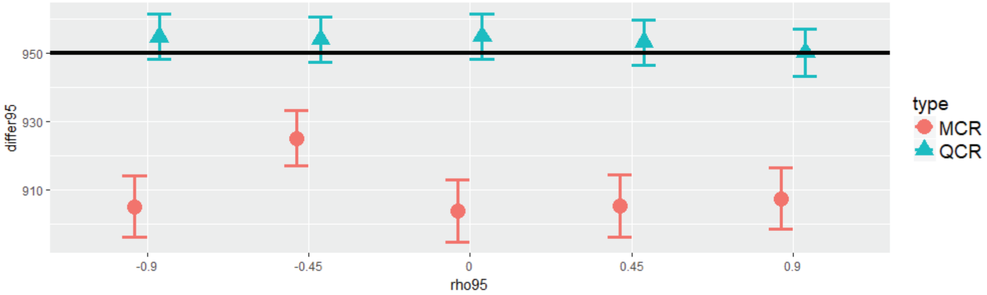

Figure 9: *Coverage rates for 95% MCR and QCR. MCR = multivariate confidence region; QCR = quantile confidence region.*

5. Discussion and further study

The QCR proposed in this study has four features. First, due to the curve and surface shapes of the bivariate and trivariate quantile vectors, the QCR has a non-ellipsoid shape regardless of its distribution. Second, the width of the QCR is longer than that of the MCR which has the ellipsoid type of confidence region in multivariate normal distribution. Third, the QCR does not satisfy the equivariance property as shown in the simulation studies of the normal mixture cases. Finally, if the distribution of the random sample is known, it is possible to obtain the QCR using its cdf and also easily obtain the MCR regardless of its distribution.

In this section, we will discuss further studies based on the features of the QCR explained above. One needs to develop other approaches to solve the shape of the non-ellipsoid QCR in multivariate normal distribution. Then we may ultimate the problem of the QCR which has a wider confidence region than the MCR. We have used the coverage rate as a criteria for the comparison of confidence region, since this is generally used in the simulation studies for the confidence region. Nonetheless, the QCR has non-ellipsoid form and does not satisfy the equivariance property; therefore, one needs to explore some advantages by using other criteria such as the power function of the confidence region.

The QCR needs to compensate for these weaknesses; however, we can say that this alternative method is valuable since the proposed confidence regions can easily be obtained and have good coverage rates for any kind of multivariate distribution.

6. Conclusion

Lots of multivariate data in the real world do not satisfy the normality assumption in many cases. In this situation, it is very hard to derive the MCR of the population mean vector. In this work, an alternative confidence region is proposed using multivariate quantile vectors. It has a good advantage that this confidence region could be defined for normal distribution as well as for any other distribution functions.

The confidence region using multivariate quantile vectors is obtained for various types of distribution functions that are both a symmetric and asymmetric distribution. Then, the coverage rates are also calculated and discussed. The QCR is found to have better coverage rates even for asymmetric distributions.

When we need to obtain the confidence region of the population mean vector using the multivariate quantile vectors for the real multivariate data, one must decide either normal or non-normal distribution functions corresponding to the data. If it turns out to be non-normal, then one can estimate the parameters of an appropriate mixture distribution function. These procedures can easily be performed using the R package 'mixtool'. Therefore, the QCR proposed in this work is found to have better properties than the MCR; therefore, we conclude that the QCR would be very useful for multivariate data analysis.

References

Asgharzadeh A and Abdi M (2011). Confidence intervals and joint confidence regions for the two-parameter exponential distribution based on records, *Communications for Statistical Applications and Methods*, **18**, 103–110.

Chew V (1966). Confidence, prediction, and tolerance regions for the multivariate normal distribution, *Journal of the American Statistical Association*, **61**, 605–617.

Fan J and Zhang W (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models, *Scandinavian Journal of Statistics*, **27**, 715–731.

Frank O (1966). Simultaneous confidence intervals, *Scandinavian Actuarial Journal*, **1966**, 78–89.

Hong CS, Han SJ, and Lee GP (2016). Vector at risk and alternative value at risk, *The Korean Journal of Applied Statistics*, **29**, 689–697.

Johnson RA and Wichern DW (2002). *Applied Multivariate Statistical Analysis* (5th ed), Prentice Hall, Upper Saddle River.

Sun J and Loader CR (1994). Simultaneous confidence bands for linear regression and smoothing, *The Annals of Statistics*, **22**, 1328–1345.

Received July 12, 2017; Revised September 20, 2017; Accepted October 16, 2017