Distribution Sampling of Vector Variance without Duplications

Erna T. Herdiani, and Maman A. Djauhari

***Abstract*—**In recent years, the use of vector variance as a measure of multivariate variability has received much attention in wide range of statistics. This paper deals with a more economic measure of multivariate variability, defined as vector variance minus all duplication elements. For high dimensional data, this will increase the computational efficiency almost 50 % compared to the original vector variance. Its sampling distribution will be investigated to make its applications possible.

***Keywords*—**Asymptotic distribution, covariance matrix, likelihood ratio test, vector variance.

I. Introduction

HYPOTHESIS testing about the stability of covariance structure is one of the fundamental issues in multivariate analysis. It is usually realized based on likelihood ratio test (LRT). See, for example, [2], [12], [18], [20], and [21] for the details of simultaneous test and [1], [7], [19], and the references therein for repeated test. Its wide range of applications can be easily found in literature. To mention some, see [3] for an early development; or [29] and [18] for its application in *Manova*; or [1] , [35], [31], [36], [19], [15], [7], and [32], and the references there in, for historical background and its development in manufacturing industry; or [26] and [2] in biological research.

Under Normality, LRT means that one has to use covariance determinant (CD) as the measure of multivariate variability. This implies that LRT can only be used when the number of variables p is limited. In practice, It is not rare that the number of variables p is large. See, for example, [34], [26], and [4], for the discussion when the sample size n > p and [17] and [15]-[16] for the case n < p. This is a serious problem because, when p is large, the computation of CD is quite cumbersome and tedious. Its computational complexity is of order $O(p^3)$. Due to that limitation of CD, very recently, in [13] we propose to use vector variance (VV) as an alternative measure of multivariate variability. It is derived from the notion of vector covariance presented and used in [5], and originally introduced by [11] to measure the linear relationship between two random vectors. Although our approach in [13] is more heuristics than analytical, VV was successfully used as the stopping rule in fast minimum covariance determinant (FMCD) algorithm proposed by [25]. It reduces significantly the computational complexity of data concentration step. See [22], [23], [24], and [14] for in depth presentation and discussion on MCD. A more comprehensive and analytical discussion on VV is presented in our recent paper [8]. In that paper, we show analytical the properties of VV and its advantage relative to CD. Most recently, [9], we show the advantage of VV in monitoring multivariate process variability.

Erna T. Herdiani, Hasanuddin University , Makassar, Indonesia (phone: +62-411-585643; e-mail: herdiani.erna@ gmail.com).

M.A. Djauhari, was with Institut Teknologi Bandung, Indonesia. He is now with Universiti Teknologi Malaysia, Johor Bahru, Malaysia (e-mail: maman@utm.my).

Let Σ be the covariance matrix of the population under study. We assume that it is definite positive. Vector variance is the trace of the squared covariance matrix, i.e. $VV = Tr(\Sigma^2)$. It is the sum of square of all elements of Σ. The fact that Σ is symmetric, it is no need to involve all elements of Σ. The element of its upper (lower) triangular matrix are sufficient. This is what we want to discuss in this present paper. The rest of the paper is organized as follows. In section II, the problem formulation will be presented. Later on, in section III, we discuss the asymptotic distributional properties of modified vector variance (MVV), i.e. VV without all duplicated elements. Our approach will be based on the notions of vec operator and commutation matrix.

II. Problem Formulation

Let X is a random vector with mean vector μ and definite positive covariance matrix $\sum$. Consider X as the superposition of two random vectors $X^{(1)}$and $X^{(2)}$of dimensions p and q, respectively,

$$X = \left(X^{(1)} X^{(2)}\right)^t. \tag{1}$$

If

$$\mu^{(i)} = E(X^i); \tag{2}$$

$i = 1, 2$ and

$$\textstyle\sum_{ij} = E\left[\left(X^{(i)} - \mu^{(i)}\right)\left(X^{(j)} - \mu^{(j)}\right)^t\right]; \tag{3}$$

$i, j = 1, 2$. Then $\sum$ can be written in form of partitioned matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{4}$$

[5] uses $Tr(\Sigma_{12}\Sigma_{21})$ to measure the linear relationship between the two random vectors $X^{(1)}$and $X^{(2)}$. He calls this parameter vector covariance. It is the sum of square of all diagonal elements of $\Sigma_{12}\Sigma_{21}$. Thus $Tr\left(\Sigma_{11}^2\right)$ and $Tr\left(\Sigma_{22}^2\right)$are called vector variance (VV) of $X^{(1)}$and $X^{(2)}$ respectively. In a special case, where p = q = 1, vector covariance is the square of the classical covariance.

According to the above point of view, thus, VV of X is simply $Tr\left(\Sigma^2\right)$, i.e., the sum of square of all elements of Σ.

But, by using the vec operator, see [20] and [27], it can also be represented as $\|vec(\Sigma)\|^2$. The vec operator transforms Σ into the vector $vec(\Sigma)$ of p^2 dimension by stacking its column one after another. We see that if VV, $\|vec(\Sigma)\|^2$, is a quadratic form, covariance determinant (CD), $|\Sigma|$, is a multilinear form. Thus, the computational complexity of VV is of order $O(p^2)$ whereas that of CD, as mentioned previously, is of order $O(p^3)$. This advantage of VV is very promising especially when we work with multivariate data of high dimension. However, as Σ is symmetric, there are $\frac{(p-1)p}{2}$ elements of Σ which are doubly counted in $\|vec(\Sigma)\|^2$. This is the first problem that we want to discuss in this present paper. More specifically, instead of using the vec operator, we propose to use further operator which will transform the lower triangular part Σ_L of Σ into the vector $v(\Sigma_L)$of dimension $\frac{p(p+1)}{2}$ by stacking its column one after another. From now on we call the parameter $\|v(\Sigma_L)\|^2$ vector variance without duplication or simply modified vector variance (MVV). It is clear that MVV is more economic than VV. The second problem is to investigate the distributional properties of sample MVV. This will guide us to a more economic hypothesis testing about the stability of covariance structure mentioned in section I.

The solution for the first problem is given by teorema 1.5. in [27]. Let $\overrightarrow{u_{ij}}$ be a vector dimension $\frac{p(p+1)}{2}$ defined as follows. The$\left\{(j-1)p+i-\frac{j(j-1)}{2}\right\}$-th component is equal to 1 and 0 otherwise; i = 1, 2, …, p and j = 1, 2, …, i. Let also H_{ij} be a matrix of size (pxp) its (i,j) −th element is equal to 1 and 0 otherwise. If we define

$$T_{ij}=\begin{cases}H_{ij}+H_{ji};\ \ jika\ i\neq j\\ \quad H_{ii};\ \ \ jika\ i=j\end{cases} \tag{5}$$

Then Theorem 1.5 in [27] gives us the following result.

$$\sum_{i\geq j}\{vec(T_{ij})\}\overrightarrow{u_{ij}^t}.v(\Sigma_L)=vec(\Sigma) \tag{6}$$

Furthermore, if we denote

$$D_p=\sum_{i\geq j}\{vec(T_{ij})\}\overrightarrow{u_{ij}^t} \tag{7}$$

then

$$D_p v(\Sigma_L)=vec(\Sigma). \tag{8}$$

Consequently, if $D_p^+=\left(D_p^tD_p\right)^{-1}D_p^t$ is the generalized inverse of D_p we have the following transformation

$$v(\Sigma_L)=D_p^+vec(\Sigma) \tag{9}$$

This transformation is also valid for all symmetric matrices.

III. Distributional Properties of Sample Vector Variance without Duplication

Let $X_1, X_2, \dots, X_n$be a sample random of size n drawn from a p-variate normal distribution $N_p(\mu,\Sigma)$. Its sample mean vector and sample covariance matrix are, respectively,

$$\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i \ \text{ and } S=\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})(X_i-\bar{X})^t \tag{10}$$

Sample VV is defined as $\|vec(S)\|^2$. Accordingly sample MVV is $\|v(S_L)\|^2$. To investigate the asymptotic distribution of sample MVV, our approach here is based on the notions of vec operator and commutation matrix can be found, for example, in [20], [27], [28], and [10]. The vec operator simplifies the study of random matrix by means of random vector and commutation matrix simplifies the investigation of parameters. First, we recall the following result given in [27], about the asymptotic distribution of $vec(S)$and its covariance matrix which is represented by using commutation matrix K. See also [30] for the notation of convergence in distribution.

$$\sqrt{n-1}\{vec(S)-vec(\Sigma)\}\overset{d}{\to}N_{p^2}(0,\Gamma) \tag{11}$$

where

$$\Gamma=\left(I_{p^2}+K\right)(\Sigma\otimes\Sigma), \tag{12}$$

$$K=\sum_{i=1}^p\sum_{j=1}^p\left(H_{ij}\otimes H_{ij}^t\right) \tag{13}$$

and H_{ij} Is defined in the previous section, i.e., a matrix of size (pxp) where its (i,j)-th element is equal to 1 and 0 otherwise. From this result, if the transformation (1) is used on S, by using the result in [20] we have

$$\sqrt{n-1}\{v(S_L)-v(\Sigma_L)\}\overset{d}{\to}N_k(0,\Lambda) \tag{14}$$

where ,

$k=\frac{p(p+1)}{2}$ and

$\Lambda=var\left(v(S_L)\right)=D_p^+var\left(vec(S)\right)\left(D_p^+\right)^t=D_p^+.\Gamma.\left(D_p^+\right)^t$

Further, based on corollary 3.2. and Proposition 3.3. in [28], if we define $u\left(v(S_L)\right)=\|v(S_L)\|^2$ arrive at the following proposition about the asymptotic distribution of sample MVV.

Proposition 1

$$\sqrt{(n-1)}\{\|v(S_L)\|^2-\|v(\Sigma_L)\|^2\}\overset{d}{\to}N(0,\sigma^2) \tag{15}$$

where $\sigma^2=4\left(v(\Sigma_L)\right)^tD_p^+\Gamma\left(D_p^+\right)^t\left(v(\Sigma_L)\right)$

This proposition is seemingly complicated to be used in application because the variance of sample MVV, $\|v(S_L)\|^2$,involves multiplication of large size matrix Γ $(p^2\times p^2)$, size even for moderate value of p. However, the following proposition helps us to simplify the computation of that variance. The proof is only a matter of algebraic manipulation using the properties of vec operator and commutation matrix.

Proposition 2

Let Ω be a matrix of size (p x p) such that

$$vec(\Omega) = \left(D_p^+\right)^t D_p^+ vec(\Sigma) \tag{16}$$

then

$$\sigma^2 = 8\|vec(\Omega\Sigma)\|^2 \tag{17}$$

Proof

Since $\Gamma = \left(I_{p^2} + K\right)(\Sigma \otimes \Sigma)$, then σ^2 can be written in the form

$$\sigma^2 = 4\left(v(\Sigma_L)\right)^t D_p^+ \left(I_{p^2} + K\right)(\Sigma \otimes \Sigma)\left(D_p^+\right)^t \left(v(\Sigma_L)\right) \tag{18}$$

$$\sigma^2 = 8\left(v(\Sigma_L)\right)^t D_p^+ N_p (\Sigma \otimes \Sigma)\left(D_p^+\right)^t \left(v(\Sigma_L)\right) \tag{19}$$

where

$$N_p = \frac{1}{2}\left(I_{p^2} + K\right) \tag{20}$$

But,

$$N_p(\Sigma \otimes \Sigma) = N_p(\Sigma \otimes \Sigma)N_p. \tag{21}$$

Hence,

$$\sigma^2 = 8\left(v(\Sigma_L)\right)^t D_p^+ N_p (\Sigma \otimes \Sigma) N_p \left(D_p^+\right)^t \left(v(\Sigma_L)\right) \tag{22}$$

Finally, since $D_p^+ N_p = D_p^+$ and N_p is symmetric, we get

$$\sigma^2 = 8\left(\left(D_p^+\right)^t v(\Sigma_L)\right)^t (\Sigma \otimes \Sigma)\left(D_p^+\right)^t \left(v(\Sigma_L)\right) \tag{23}$$

$$\sigma^2 = 8\left(vec(\Omega)\right)^t (\Sigma \otimes \Sigma)\, vec(\Omega) \tag{24}$$

Because

$$\left(D_p^+\right)^t v(\Sigma_L) = \left(D_p^+\right)^t D_p^+ vec(\Sigma) \tag{25}$$

$$\sigma^2 = 8\|vec(\Omega\Sigma)\|^2 \tag{26}$$

IV. Conclusion

If vector variance $vec(\Sigma)$ is of dimensionp^2, $v(\Sigma_L)$ is of dimension $k = p(p + 1)/2$. This gain is too good to be neglected. Furthermore, Proposition 1 and 2 have made possible the application of modified vector variance, $v(\Sigma_L)$ where σ^2simply eight is times the sum of square of all elements of ΩΣ.

Acknowledgment

This research is partially supported by the Hasanuddin University 2012 research basis program study the authors thank the hasanuddin university, for that support.

References

[1] F.B. Alt, and N.D. Smith Multivariate process control. In: Krishnaiah, P.R., Rao, C.R., eds. Handbook of Statistics, Vo. 7 Elsevier Sciences Publishers, 1988, pp. 333-351.

[2] M.J. Anderson, Distance based test for homogeneity of multivariate dispersion, Biometrics, 62, 2006. 245 -253.

[3] T.W. Anderson, An Introduction to multivariate statistical analysis. John Wiley & Sons, Inc., New York. 1958.

[4] J. Ng. R Chilson., A. Wagner, and R. Zamar, Parallel Computation of high dimensional robust correlation and covariance matrices, algorithmica, 45(3), 2006 pp. 403 – 431.

[5] R. Cleroux, Multivariate Association and Inference Problems in Data Analysis, proceedings of the fifth international symposium on data analysis and informatics, vol. 1, Versailles, France. 1987.

[6] Jr, N. Da costa,. S. Nunes, P.Ceretta and S. Da Silva, Stock market co-movements revisited, Economics Bulletin, 7(3), 2005, pp. 1-9.

[7] M. A. Djauhari, Improved Monitoring of Multivariate Variability, Journal of Quality Technology, 37(1), 2005, pp. 32-39.

[8] M. A. Djauhari, A Measure of Data Concentration, Journal of Probability and Statistics, 2(2), 2007, pp. 139-155.

[9] M. A. Djauhari, M. Mashuri, and D. E. Herwindiati, Multivariate Process Variability Monitoring, Communications in statistics – Theory and Methods, 37(1), 2008, pp. 1742 – 1754.

[10] H. El Maache, and Y. Lepage, Measures d'Association Vectorielle Basées sur une Matrice de Corrélation. Revue de Statistique Appliquée, 46(4), 1998, pp. 27-43.

[11] Y. Escoufier, Le traitement des variables vectorielles. Biometrics, 29, 1973, pp.751-760.

[12] A. K. Gupta, and J. Tang, Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models, Biometrika, 71(3), 1984, pp. 555-559.

[13] D. E., Herwindiati, M. A., Djauhari, and M., Mashuri, Robust Multivariate Outlier Labeling, Communications in statistics – Computation and Simulation, 36(6), 2007, pp. 1287 – 1294.

[14] M., Hubert, P.J., Rouddeeuw, and S. van Aelst, Multivariate outlier detection and robustness, in handbook of statistics, vol. 24, Elsevier B.V., 2005, pp. 263-302.

[15] M. B. C., Khoo, and S. H. Quah, Multivariate control chart for process dispersion based on individual observations. Quality Engineering, 15(4), 2003, pp. 639-643.

[16] M. B. C., Khoo, and S. H. Quah, Alternatives to the multivariate control chart for process dispersion. Quality Engineering, 16(3), 2004, pp. 423-435.

[17] O., Ledoit and M. Wolf, Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. The Annals of Statistics, 30(4), 2002, pp. 1081-1102.

[18] K.V., Mardia, J.M. Kent, Multivariate analysis, seventh printing, Academic press, London, 2000.

[19] D.C. Montgomery, Introduction to statistical quality control, fourth edition, John Willey & Sons, Inc., New York, 2001.

[20] R. J. Muirhead, Aspects of multivariate statistical theory, John Willey & Sons, Inc., New York, 1982.

[21] V. Ragea, Testing correlation stability during hectic financial markets, financial market and Portfolio Management, 17(3), 2003, pp. 289-308.

[22] P.J. Rosseeuw, Multivariate estimation with high breakdown point. In mathematical statistics and applications, B, Grossman W., Pflug G., Vincze I and Wertz, W., editors, D. Reidel Publishing Company, 1985, pp. 283-297.

[23] P.J., Rosseeuw, and A.M. Leroy, Robust regression and outlier detection, John Willey & Sons, Inc., New York, 1987.

[24] P.J., Rosseeuw, and M. Hubert, Regression Depth, Journal of the American Statistical Association, 94, 1999, pp.388-402.

[25] P.J., Rosseeuw, and K. van Driessen, A fast algorithm for the minimum covariance determinant estimatior, Technometrics, 41, 1999, pp. 212 – 223.

[26] J., Schafer, and K.,Strimmer, A Shrinkage Approach to large scale covariance matrix estimation and implications for functional genomics.

Statistical Applications in genetics and molecular biology, 4, 2005, pp 1-30.

[27] J. R., Schott, Matrix analysis for statistics, John Wiley & Sons, New York, 1997.

[28] J. R., Schott, Some tests for the equality of covariance matrices, journal of statistical planning and inference, 94, 2001, pp 25 – 36.

[29] G.A.F., Seber, Multivariate Observations, John Wiley & Sons, New York, 1984.

[30] R.J. Serfling, Approximation Theorems of mathematical statistics, John Wiley & Sons, New York, 1980.

[31] J.H. Sullivan, and W.H. Woodall, A Comparison of multivariate control charts for individual observations, Journal of Quality Technology, 28(4), 1996, pp 398-408.

[32] J.H., Sullivan, Z.G., Stoumbos, R.L. Mason, and J.C. Young, Step down analysis for changes in the covariance matrix and other parameters, Journal of Quality Technology, 39(1), 2007, pp 66-84.

[33] G.Y.N. Tang, The Intertemporal stability of the covariance and correlation matrices of Hongkong Stock Returns, Applied financial economics, 8, 1998, pp 359-365.

[34] M. Werner, Identification of multivariate outliers in large data sets, PhD dissertation, University of Colorado at Denver, 2003

[35] S.J., Wierda, Multivariate statistical process control, Recent results and directions for future research, statistica neerlandica, 48(2), 1994 pp. 147-168.

[36] W.H. Woodall, and D.C. Montgomery, (1999). Research issues and ideas in statistical process control, Journal of Quality Technology, 31(4), 1999, pp 376-386.

Erna T. Herdiani Bandung, Indonesia, 29 April 1975. The educational background is Institut Teknologi Bandung (ITB), statistics algebra, Bandung, Indonesia, 2004. The current job is lecture of department mathematics, Hasanuddin University, Makassar, Indonesia. 2000-now.