MATEMATIKA, 2009, Volume 25, Number 1, 79–85

Asymptotic Distribution of Sample Covriance Determinant

Maman A. Djauhari
Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia
e-mail: maman@utm.my

Abstract Under normality, an asymptotic distribution of sample covariance determinant will be derived. We show that this asymptotic distribution is more applicable in practice than the classical one. This will justify its usefulness in inferential study and other applications in a more comprehensive and accurate manner.

Keywords Keywords: Convergence in distribution, generalized variance, multivariate dispersion, parameter estimation, unbiased estimate.

1 Introduction

In the literature on multivariate statistical theory, e.g., [2] and [15], we can find the mathematical derivation showing that the sample covariance determinant or generalized variance converges in distribution to a normal distribution. However, this classical result is not suitable for practical purposes because the distribution parameters are the limits of the of the true parameters when the sample size tends to infinity and in addition, the convergence to this limit is slow. On the other hand, in the literature on multivariate statistical applications, e.g., Alt and Smith [1] and Montgomery [14], the true parameters are considered but not the distribution. In the construction of control region, for example, Montgomery [14] points out that "most of the probability distribution of sample covariance determinant is contained in the interval ± 3 times the standard deviation of sample covariance determinant from its mean" without specifying the distribution. Such a statement is ambiguous and needs formal explanation. This is the main topic of this paper.

The importance of covariance determinant as a measure of multvariate dispersion lies in its important role in scientific investigation based on multivariate data sets. Although it is a scalar simplication of a complex structure of multivariate dispersion (Montgomery [14]), it is the most common and widely used measure in practice. This is due to the fact that covariance determinant is very simple in its geometric interpretation and its computation.

Mardia et al. [12] says that besides covariance determinant there is another common measure, i.e., total variance. But we do not recommend it for general purposes. It is only the sum of all diagonal elements of covariance matrix and does not involve all covariance structure. Hence, its use as multivariate dispersion measure is very limited, i.e., to the case where the variables are independent to each other. Therefore, in what follows we focus on covariance determinant.

We can find the role of covariance determinant in controlling the stability of covariance structure in a wide spectrum of scientific investigations; from hard sciences like astronomy and theoretical physics to soft sciences like supply chain management, and from manufacturing industry to service industry. See Hubert et al. [9] for an application in astronomy, Edelman and Rao [6] in theoretical physics, and Beamon and Ware [3] in supply chain

management. We can also see, for example, its application in service industry in Roes and Dorr [17], Sulek [21] and Wood [24]; Hanslik et al. [8], Sellick [18], Shahian et al. [20] and Woodall [25] in healthcare industry; Kruegel et al. [11] and Ye et al. [27] in information industry; Florac et al. [7] and Jakolte and Saxena [10] in software industry; Da Costa et al. [4], Ragea [16] and Tang [23] in financial industry; and Djauhari [5], Mason et al. [13], Sullivan et al. [22] and Woodall and Montgomery [26] in manufacturing industry.

Due to that important role, in the present paper the asymptotic distribution of sample covariance determinant with true parameters will be derived. The discussion will begin in Section 2 with a brief review of the classical asymptotic distribution. In Section 3 we introduce a theorem on an asymptotic distribution with true parameters. Later on, parameter estimates based on several independent random samples will be derived in Section 4. Additional remarks will close this presentation.

2 Classical Asymptotic Distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a $p-variate$ normal distribution $N_p(\mu, \Sigma)$ will be assumed positive definite throughout the paper. The sample mean vector and sample covariance matrix are, respectively,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^t,$$

and sample covariance determinant, also called generalized variance, is $|S|$. To investigate the distribution of $|S|$ in the classical approach, the following procedure which consists of two steps, is standard. First, consider the function $f(S) = |S|$. This function is a real valued and continuous function of S where the first and second derivatives exist. In the second step, we apply Theorem 4.2.5 in Anderson ([2], p. 76) to $f(S)$. For clarity, that theorem is rewritten as Theorem 1.

Theorem 1 *Let Y be a random vector of p dimension, c be a constant vector in R^p, $\sqrt{n-1}(Y-c) \xrightarrow{d} N_p(0,\Gamma)$ and $f(Y)$ be a real valued function of Y where the first two derivatives at $Y = c$ exist. Let also Y_i and $\left.\frac{\partial f(Y)}{\partial Y_i}\right|_{Y=c}$ be the i-th component of Y and ϕ, respectively. Then*

$$\sqrt{n-1}(f(Y) - f(c)) \xrightarrow{d} N_p(0, \phi'\Gamma\phi).$$

Based on this theorem and the central limit theorem, Anderson [2] derives directly the following corollary. (See also Muirhead [15]).

Corollary $\sqrt{n-1}(|S| - |\Sigma|) \xrightarrow{d} N(0, 2p|\Sigma|^2)$.

Theoretically, the asymptotic distribution of $|S|$ in this corollary is very important. For example, it is used in exploring another asymptotic distribution as can be seen in the next section. However, it is not useful in practice because the parameters $|\Sigma|$ and $\frac{2p}{n-1}|\Sigma|^2$ are not the true mean and variance. We show that those parameters are the limit from left of the true ones when n tends to infinity. In the next section we derive the true parameters of $|S|$

and another version of its asymptotic distribution. This version will ensure the applicability of $|S|$ in practice. Later on, we show that the convergence of the true parameters to the parameters in the above corollary is very slow.

3 A Non-classical Asymptotic Distribution

Let us start with considering the exact distribution of $|S|$. In the theoretical literature on multivariate analysis, we can easily find that the distribution of $|S|$ is equal to that of

$$\frac{|\Sigma|}{(n-1)^p} U_1 U_2, \cdots, U_p$$

where $U_1 U_2, \cdots, U_p$ are independent and U_k is distributed as $\chi^2_{n-k}; k = 1, 2, \ldots, p$. See, for example, Anderson [2] and Muirhead [15]. From this distribution, it can be shown that the mean and variance of $|S|$ are

$$E(|S|) = b_1|\Sigma| \quad \text{and} \quad \text{Var}(|S|) = b_2|\Sigma|^2, \tag{1}$$

respectively, where

$$b_1 = \frac{1}{(n-1)^p} \prod_{k=1}^{p} (n-k) \quad \text{and} \quad b_2 = b_1 \Big\{ \frac{1}{(n-1)^p} \prod_{k=1}^{p} (n-k+2) - b_1 \Big\}.$$

See also Montgomery [14].

In (1), the true mean of $|S|$ is $b_1|\Sigma|$ and the true variance is $b_2|\Sigma|^2$. It is clear that $\lim_{n\to\infty} b_1 = 1$. Thus, the true mean tends to the mean in the corollary of Theorem 1. It can also be shown that

$$\lim_{n\to\infty} \frac{2p/(n-1)}{b_2} = 1.$$

Furthermore, based on the result in (1), a numerical computation shows that for $p = 2$ needs n to be larger than 10,000 in order for the parameter mean in the corollary of Theorem 1 to differ no more than 10^{-5} from the true mean. For larger value of p, it needs even larger sample size n. This shows that the convergence of the true parameters to the parameters in the corollary of Theorem 1 is very slow.

In order to have a more useful asymptotic distribution for practical purposes, where the parameters mean and variance are sensitive to the change of sample size n and the number of variables p, we introduce the following theorem.

Theorem 2 $|S| \xrightarrow{d} N(b_1|\Sigma|, b_2|\Sigma|^2)$

Proof

It can be shown that,

(i) $\lim_{n\to\infty} \dfrac{b_1|\Sigma| - |\Sigma|}{\sqrt{\dfrac{2p}{n-1}}|\Sigma|^2} = 0$. Here the numerator is the difference betweeen the hypothetical mean and the mean in the corollary of Theorem 1. The denominator is the standard deviation in that corollary;

(ii) $\lim_{n\to\infty} \dfrac{b_2|\Sigma|^2}{\sqrt{\dfrac{2p}{n-1}|\Sigma|^2}} = 1$. The numerator is the hypothetical standard deviation while the denominator is as in above.

Based on these properties, by using Lemma A in Serfling ([19], p. 20), we arrive at the statement of the theorem.□

This theorem appears straightforward. Nevertheless, we did not find any similar theorem in the literature. For this theorem to be more practical, it is left for us to estimate the parameters $|\Sigma|$ and $|\Sigma|^2$ based on random samples. This is discussed in the next section.

4 Estimation

Consider again the random sample discussed in Section 2. In the previous section we have

$$E(|S|) = b_1|\Sigma| \quad \text{and} \quad \text{Var}(|S|) = b_2|\Sigma|^2.$$

The first expression implies that $|S|/b_1$ is an unbiased estimate of $|\Sigma|$. From the second, after an algebraic manipulation, we conclude that $|S|^2/(b_1^2 + b_2)$ is an unbiased estimate of $|\Sigma|^2$. The denominators b_1 and $(b_1^2 + b_2)$ are the bias factors when we estimate $|\Sigma|$ and $|\Sigma|^2$ based on single random sample.

Now, suppose that m independent random samples of the same size n, drawn from $N_p(\mu, \Sigma)$ are available. This is the common situation in industrial applications of statistical process control. To look for the unbiased estimates of $|\Sigma|$ and $|\Sigma|^2$ based on those samples, we start by first looking at the average of sample covariance matrices which is the best estimate of Σ, i.e., unbiased and has minimum variance based on several independent samples. The following approach can be seen in Djauhari [5].

Let S_k be the covariance matrix of sample $k; k = 1, 2, \ldots, m$ and $\overline{S}$ their average. As $(n-1)S_1, (n-1)S_2, \ldots, (n-1)S_p$ have independent and identical Wishart distributions $W_p(\Sigma, n-1)$, $m(n-1)\overline{S}$ is distributed as $W_p(\Sigma, m(n-1))$, and consequently,

$$|\overline{S}| \text{ is distributed as } \frac{|\Sigma|}{\{m(n-1)\}^p} V_1 V_2 \cdots V_p \tag{2}$$

where $V_1, V_2, \ldots, V_p$ are independent and V_k has $\chi^2_{m(n-1)-k+1}; k = 1, 2, \ldots, p$. Accordingly, in general, the r-th moment of $|\overline{S}|$ is

$$\begin{aligned} E\big(|\overline{S}|^r\big) &= \left(\frac{|\Sigma|}{(m(n-1))^p}\right)^r \prod_{k=1}^{p} E(V_k^r) \\ &= \left(\frac{2}{m(n-1)}\right)^{pr} |\Sigma|^r \prod_{k=1}^{p} \frac{\Gamma\Big(r + \dfrac{m(n-k)-k+1}{2}\Big)}{\Gamma\Big(\dfrac{m(n-k)-k+1}{2}\Big)} \end{aligned}$$

where the function $\Gamma(x)$ on the right hand side is the gamma function of x,

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

From this relation or directly from (2), we obtain the mean and variance of $|\overline{S}|$,

$$E\big(|\overline{S}|\big) = b_3|\Sigma| \quad \text{and} \quad \text{Var}\big(|\overline{S}|\big) = b_4|\Sigma|^2 \tag{3}$$

where the constant

$$b_3 = \frac{1}{\{m(n-p)\}^p} \prod_{k=1}^{p} \{m(n-1) - k + 1\}$$

and, after a simple manipulation,

$$b_4 = b_3 \left[\frac{1}{\{m(n-1)\}^p} \prod_{k=1}^{p} \{m(n-1) - k + 3\} - b_3 \right].$$

From (3) we conclude that $|\overline{S}|/b_3$ is an unbiased estimate of $|\Sigma|$. Furthermore, after an algebraic manipulation, we obtain that $|\overline{S}|^2/(b_3^2 + b_4)$ is an unbiased estimate of $|\Sigma|^2$. The bias factors b_3 and $(b_3^2 + b_4)$ are originally proposed by Djauhari (2005) to estimate $|\Sigma|$ and $|\Sigma|^2$ based on several independent random samples.

5 Conclusion

We conclude that, for industrial applications, the following distribution of $|S_k|; k = 1, 2, \ldots, p$ is far better than the classical one in the corollary of Theorem 1 and even than the one in Montgomery ([14], Example 10.2),

$$|S_k| \xrightarrow{d} N\left(\frac{b_1}{b_2}|\overline{S}|, \frac{b_2}{b_3^2 + b_4}|\overline{S}|^2 \right) \tag{4}$$

This is due to Theorem 2 and the unbiased nature of the estimates.

6 Additional Remarks

We derive the true mean and variance of sample covariance determinant and then show an asymptotic distribution which, for practical purposes, is better than the classical one in the following sense. The former has the true mean and variance which depend on the sample size n and the number of variables p while the mean and variance of the latter are constant and equal to the limit from left of the true mean and variance if n goes to infinity.

For practical purposes, the distribution in (4) is better than that given by classical approach and even than the standard one given in Montgomery ([14], Example 10.2).

Theoretically, a sequence of random variables might converge in distribution to more than one distributions. It is so with the sequence of sample covariance determinants as a function of sample size. Two problems are still open to be explored:

(i) Is there any other asymptotic distribution to which the sequence converges faster in distribution?

(ii) Are the parameter estimates in Section 4 the best?

Acknowledgement

Special thanks go to Universiti Teknologi Malaysia for providing research facilities.

References

[1] F.B. Alt & N.D. Smith, *Multivariate Process Control* Handbook of Statistics, Vol. 7(1988), (Krishnaiah P.R. and Rao C.R. Eds), Elsevier Science Publishers, p. 333-351.

[2] T.W. Anderson, *An Introduction to Multivariate Analysis,* John Wiley and Sons, Inc., New York, 1966.

[3] B.M. Beamon & T.M. Ware, *A process quality model for the analysis, improvement and control of supply chain systems,* International Journal of Physical Distribution & Logistics Management, 28(1998), 704-715.

[4] N. Da Costa Jr., S. Nunes, P. Ceretta & S. Da Silva, *Stock-market co-movements revisited,* Economics Bulletin, 7(2005), 1-9.

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

[6] A. Edelman & N.R. Rao, *Random Matrix Theory,* Acta Numerica, (2005), 1-65.

[7] A.W. Florac, A.D. Carleton & J.R. Barnard, *Statistical Process Control: Analyzing a Space Shuttle Onboard Software Process,* IEEE Software, July/August(2000), 97-106.

[8] T. Hanslik, P. Boelle & A. Flahault, *The control chart: an epidemiological tool for public health monitoring,* Public Health, 115(2001), 277-281.

[9] M. Hubert, P.J. Rousseeuw & S. van Aelst, *High-Breakdown Robust Multivariate Methods,* Statistical Science, 23(1)(2008), 92-119.

[10] P. Jakolte & A. Saxena, *Optimum Control Limits for Employing Statistical Process Control in Software Process,* IEEE Transactions on Software Engineering, 28(2002), 1126-1134.

[11] C. Kruegel, F. Valuer & G. Vigna, *Intrusion Detection and Correlation,* Challenges and Solutions, Springer Science + Business Media, Inc. Boston, 2005.

[12] K.V. Mardia, J.T. Kent & J.M. Bibby, *Multivariate Analysis,* Academic Press, London, (2000).

[13] R.L. Mason, Y.M. Chou & J.C. Young, *Monitoring Variation in a Multivariate Process when the Dimension is Large Relative to the Sample Size,* Communications in Statistics-Theory and Methods, 38(6)(2009), 939-951.

[14] D.C. Montgomery, *Introduction to Statistical Quality Control,* Fifth Edition, John Wiley and Sons, Inc., New York, 2005.

[15] R.J. Muirhead, *Aspects of Multivariate Statistical Theory,* John Wiley & Sons, Inc., New York, 1982.

[16] V. Ragea, *Testing Correlation Stability During Hectic Financial Markets,* Financial Market and Portfolio Management, 17(2003), 289-308.

[17] K.C.B. Roes & D. Dorr, *Implementing statistical process control in service processes,* International Journal of Quality Science, 2(1997), 149-66.

[18] J.A. Sellick Jr. *The use of statistical process control charts in hospital epidemiology,* Infection Control and Hospital Epidemiology, 14(1993), 649-56.

[19] R.J. Serfling, *Approximation Theorems of Mathematical Statistics,* John Wiley , New York,(1980).

[20] D.M. Shahian, W.A. Williamson, L.G. Svensson, J.D. Restuccia & R.S. dAgostino, *Applications of statistical quality control to cardiac surgery,* Annals of Thoracic Surgery, 62(1996), 1351-1359.

[21] J.M. Sulek, *Statistical quality control in services,* International Journal of Services Technology and Management, 5(2004), 522-531.

[22] J.H. Sullivan, Z.G. Stoumbos, R.L. Mason & J.C. Young, *Step-Down Analysis for Changes in the Covariance Matrix and Other Parameters,* Journal of Quality Technology, 39(2007), 66-84.

[23] G.Y.N. Tang, *The Intertemporal Stability of the Covariance and Correlation Matrices of Hong Kong Stock Returns,* Applied Financial Economics, 8(1998), 359-365.

[24] M. Wood, *Statistical methods for monitoring service process,* International Journal of Service Industry Management, 5(1994), 53-68.

[25] W.H. Woodall, *Use of Control Charts in Health Care Monitoring and Public Health Surveillance,* Journal of Quality Technology, 38 (2006), 89-104.

[26] W.H. Woodall & D.C. Montgomery, *Research Issues and Ideas in Statistical Process Control,* Journal of Quality Technology, 31(1999), 376-386.

[27] N. Ye, J. Giardano & J. Feldman, *A Process Control Approach to Cyber Attack Detection,* Communication of the ACM, 44(2001), 76-82.