JIRSS (2014)
Vol. 13, No. 1, pp 57-67

Inferences on the Generalized Variance under Normality

A. A. Jafari[1], **M. R. Kazemi**[2]

[1]Department of Statistics, Yazd University, Yazd, Iran.
[2]Department of Statistics, Fasa University, Fasa, Iran.

Abstract. Generalized variance is applied for determination of dispersion in a multivariate population and is a successful measure for concentration of multivariate data. In this article, we consider constructing confidence interval and testing the hypotheses about generalized variance in a multivariate normal distribution and give a computational approach. Simulation studies are performed to compare this approach and three approximate methods; the simulations show that our approach is satisfactory. At the end, two practical examples are given.

Keywords. Actual size, coverage probability, generalized variances, Monte Carlo simulation.

MSC: 62H99, 62F03.

1 Introduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an independent random sample from a p-variate normal population with positive definite covariance matrix Σ. Wilks (1932) introduced generalized variance, which is defined as the determinant of covariance matrix, $|\Sigma|$, for dispersion of a multivariate population. This measure is used for overall multivariate scatter, and one can rank distinct groups and populations based on the order of their spread. Generalized variance is a successful measure for concentration

A. A. Jafari(✉)(aajafari@yazd.ac.ir), M. R. Kazemi(kazemi@fasau.ac.ir)
Received: February 2013; Accepted: November 2013

of multivariate data (Djauhari, 2007) and is applicable in monitoring process variability (Djauhari, 2005; Djauhari, et al., 2008). Therefore, many researchers have studied the estimation of generalized variance. Sarkar (1989) provided the shortest confidence interval for generalized variance. Sarkar (1991), and Iliopoulos and Kourouklis (1998) obtained a stein-type interval and an improvement confidence interval for generalized variances, respectively. For natural exponential family, Kokonendji (2003) proposed uniformly minimum variance unbiased estimator (UMVUE) of the generalizad variance, and Kokonendji and Pommeret (2007) compared this estimator and the maximum likelihood estimator (MLE). Djauhari (2009) derived an asymptotic distribution for the sample generalized variance.

Our focus is to find a confidence interval for $|\Sigma|$ and one-sided test of hypothesis

$$H_0 : |\Sigma| \leq d_0 \qquad vs. \qquad H_1 : |\Sigma| > d_0, \tag{1}$$

and two-sided test of hypothesis

$$H_0 : |\Sigma| = d_0 \qquad vs. \qquad H_1 : |\Sigma| \neq d_0. \tag{2}$$

For $p = 1, 2$, the distribution of sample generalized variance has a simple form. But for $p > 2$, its exact distribution has a complicated nature, and it would as well be difficult to obtain the interval estimation and test of $|\Sigma|$; therefore, the authors gave some approximate methods. In this paper, the state $p > 2$ is our interest. However, our paper is organized as follows: In Section 2, for constructing confidence interval and testing the hypotheses in (1) and (2), we give a computational approach and review three approximate methods. In Section 3, the methods are compared using the Monte Carlo simulation, and we illustrate our method using two real examples. Some conclusions are provided in Section 4.

2 Inferences

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from a p-variate normal distribution with positive definite covariance matrix Σ. Let

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})', \tag{3}$$

be the sample covariance matrix. Our interest is to find a confidence interval and hypothesis test for the generalized variance, $|\Sigma|$. For this purpose, we can use the exact distribution of the determinant of S. It is well known that $|S|$ is distributed as a product of chi-square distributions (See Anderson, 2003), i.e.

$$U = \frac{(n-1)^p}{|\Sigma|}|S| \sim \prod_{j=1}^{p} \chi^2_{(n-j)}, \tag{4}$$

where $\chi^2_{(n-j)}$, $j = 1, ..., p$ are independently chi-square random variables with $n - j$ degrees of freedom. Therefore, distribution of $|S|$ is stochastically increasing in $|\Sigma|$, i.e. $P(|S| > a)$ is an increasing function of $|\Sigma|$ and the p-value for the one-sided hypothesis test in (1) is given by

$$p = P\left(|S| \geq |s| \mid H_0\right) = P\left(U \geq \frac{(n-1)^p}{d_0}|s|\right), \tag{5}$$

and for the two-sided hypothesis test in (2) it is given by

$$p = 2\min\left\{P\left(U \geq \frac{(n-1)^p}{d_0}|s|\right), P\left(U \leq \frac{(n-1)^p}{d_0}|s|\right)\right\}, \tag{6}$$

where $|s|$ is the observed value of $|S|$. Also, an $(1-\alpha)$ confidence interval for $|\Sigma|$ based on U is

$$\left(\frac{(n-1)^p}{U_{\alpha/2}}|s|, \frac{(n-1)^p}{U_{1-\alpha/2}}|s|\right), \tag{7}$$

where U_γ is the $(1-\gamma)$th percentile for distribution of U.

The p-values in (5) and (6), and confidence interval in (7) cannot be calculated exactly because density function of U has a complicated form. However, it can be estimated numerically using Monte Carlo simulation. The following algorithm is useful to estimate the confidence interval for $|\Sigma|$ in (7) and p-values in (5) and (6):

Algorithm 1 *Given p, n and $|s|$;*
1- Generate $U_j \sim \chi^2_{(n-j)}$.
2- Calculate $U = \prod_{j=1}^{p} \chi^2_{(n-j)}$.
3- Calculate $V = \frac{(n-1)^p}{U}|s|$.
4- Repeat steps 1-3 for large number of times (say $m = 5000$) and obtain the m values of V.

Let $V_{(\gamma)}$ be the 100γth percentile of V_l's, $i = 1, ..., m$. Then $1-\alpha$ confidence interval for $|\Sigma|$ is $[V_{(\alpha/2)}, V_{(1-\alpha/2)}]$. The p-value in (5) *for testing hypothesis in* (1) *is estimated by proportions of V_l's that are less or equal to d_0.*

The p-values in (5) and (6), and the confidence interval in (7), can also be calculated approximately. In the following sections, we consider three normal approximations.

2.1 Anderson Approximation

Based on central limit theorem, Anderson (2003) showed that

$$\sqrt{n-1}\left(\frac{|S|}{|\Sigma|}-1\right)\xrightarrow{d}N\left(0,2p\right). \tag{8}$$

Using this approximation, the p-value in (5) becomes

$$p = P\left(|S| \geq |s| \mid H_0\right) = 1-\Phi\left(\frac{\sqrt{n-1}}{\sqrt{2p}}\left(\frac{|S|}{d_0}-1\right)\right), \tag{9}$$

where $\Phi(.)$ is the cumulative distribution function of standard normal distribution. In addition, the $(1-\alpha)$ confidence interval based on this approximation is

$$\left(\frac{\sqrt{n-1}\,|S|}{\sqrt{n-1}+\sqrt{2p}Z_{\alpha/2}}, \frac{\sqrt{n-1}\,|S|}{\sqrt{n-1}-\sqrt{2p}Z_{\alpha/2}}\right), \tag{10}$$

where $Z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of standard normal distribution. This formula requires, however, that n be sufficiently large, i.e. $n > 2pZ^2_{\alpha/2}+1$, provided that the upper bound is positive.

2.2 Sarkar Approximation

Hoel (1937) suggested an approximation density function for U as

$$f^*_{p,n}(x) = \gamma p^{-1} f_\tau\left(\gamma x^{\frac{1}{p}}\right) x^{\frac{1}{p}-1},$$

where $f_\tau(x)$ is the density of a chi-square random variable with $\tau = p(n-p)$ degrees of freedom and $\gamma = p(1-\frac{1}{2n}(p-1)(p-2))^{\frac{1}{p}}$, where $f^*_{p,n}(x)$ is exact for $p = 1$ and $p = 2$. Sarkar (1989) showed that this approximation is useful for $p \leq 3$, and suggested the following normal

approximation for $p > 3$. Based on Johnson et al. (1994), we approximately have

$$\log\left(\chi^2_{(v)}\right) \sim N\left(\psi\left(\frac{v}{2}\right) + \log(2), \psi'\left(\frac{v}{2}\right)\right), \tag{11}$$

where $\psi(.)$ is the digamma function and $\psi'(.)$ is its derivative. Applying this approximation for

$$Y = \log(U) = \sum_{j=1}^{p} \log\left(\chi^2_{(n-j)}\right), \tag{12}$$

we can conclude that Y has a normal distribution with mean $\mu_Y = \sum_{j=1}^{p} \psi\left(\frac{n-j}{2}\right) + p\log(2)$ and variance $\sigma^2_Y = \sum_{j=1}^{p} \psi'\left(\frac{n-j}{2}\right)$. Therefore, we approximately have

$$P\left(U \geq \frac{(n-1)^p}{d_0}|s|\right) = 1 - \Phi\left(\frac{p\log(n-1) + \log(|s|) - \log(d_0) - \mu_Y}{\sigma_Y}\right),$$

where $\Phi(.)$ is the cumulative distribution function of standard normal distribution. From approximation for Y in (12), we can also obtain an approximate confidence interval for $|\Sigma|$ as

$$\left((n-1)^p |s| \exp\left(-\mu_Y - \sigma_Y Z_{\alpha/2}\right), (n-1)^p |s| \exp\left(-\mu_Y + \sigma_Y Z_{\alpha/2}\right)\right). \tag{13}$$

2.3 Djauhari Approximation

Djauhari (2009) derived an asymptotic distribution for $|S|$. He showed that

$$\frac{|S|}{|\Sigma|} \xrightarrow{d} N(b_1, b_2), \tag{14}$$

where $b_1 = \frac{1}{(n-1)^p} \prod_{j=1}^{p} (n-j)$ and $b_2 = \frac{b_1}{(n-1)^p} \prod_{j=1}^{p} (n-j+2) - b_1^2$. Using this approximation, the p-value in (5) becomes

$$p = P(|S| \geq |s| \mid H_0) = 1 - \Phi\left(\frac{|s| - b_1 d_0}{d_0\sqrt{b_2}}\right). \tag{15}$$

In addition, the $(1-\alpha)$ confidence interval based on this approximation is

$$\left(\frac{|S|}{b_1 + \sqrt{b_2} Z_{\alpha/2}}, \frac{|S|}{b_1 - \sqrt{b_2} Z_{\alpha/2}}\right). \tag{16}$$

This formula requires, that $b_1^2 > b_2 Z^2_{\alpha/2}$, provided that the upper bound is positive.

3 Numerical Studies

In this section, the given approaches in Section 2 are compared using the Monte Carlo simulation. In addition, these approaches are illustrated by using two real examples.

3.1 Simulation

Using the Monte Carlo simulation, with 10000 replications, we compare the coverage probabilities and expected lengths of the confidence intervals for generalized variance, $|\Sigma|$. We consider the following methods: i) computational method (CM) using algorithm 1 with $m = 5000$, ii) the Anderson method (AM), iii) the Sarkar method (SM), and iv) the Djauhari method (DM). Note that the Andeson method and Djauhari method are not applicable in some cases.

For this purpose, we generate a random sample of size n from multivariate normal with dimension p, mean vector $\mathbf{0}$, and covariance matrix Σ, and obtain the 95% two sided confidence intervals for $|\Sigma|$ through the above approaches. The coverage probabilities and expected lengths of the methods are given in Table 1, and we can conclude that

1. The coverage probabilities of all methods are close to the confidence coefficient 0.95.
2. The expected length of Anderson method and Sarkar method are greater than other methods.
3. The expected length of all methods increase when the dimension p, of multivariate normal increases.
4. The expected length of all methods decrease when the sample size n, increases.
5. The expected length of all methods increase when the generalized variance $|\Sigma|$, increases.

We also performed a simulation study for comparing the actual sizes and powers of the methods for testing the one-sided test of hypothesis

$$H_0 : |\Sigma| \leq d_0 \qquad vs. \qquad H_1 : |\Sigma| > d_0,$$

with $d_0 = 0.2$. We generate a random sample of size n from multivariate normal with dimension p, mean vector $\mathbf{0}$, and covariance matrix Σ, and

Table 1: Simulated coverage probabilities and expected lengths of 95% two sided confidence intervals for $|\Sigma|$.

			p			
n	$\|\Sigma\|$	method	2	3	5	10
15	0.2	CM	0.946(0.653)	0.951(1.022)	0.946(2.110)	0.948(20.543)
		AM	—	—	—	—
		SM	0.942(0.610)	0.948(0.954)	0.943(1.958)	0.951(17.681)
		DM	—	—	—	—
	1	CM	0.953(3.347)	0.956(5.089)	0.949(11.135)	0.947(33.401)
		AM	—	—	—	—
		SM	0.947(3.133)	0.952(4.713)	0.943(10.260)	0.948(31.059)
		DM	—	—	—	—
30	0.2	CM	0.948(0.361)	0.946(0.483)	0.946(0.781)	0.967(1.891)
		AM	0.952(0.595)	0.976(1.545)	—	—
		SM	0.949(0.350)	0.950(0.468)	0.945(0.755)	0.963(1.814)
		DM	0.952(0.673)	0.961(3.212)	—	—
	1	CM	0.947(1.801)	0.953(2.411)	0.951(3.856)	0.942(9.393)
		AM	0.966(2.972)	0.971(7.705)	—	—
		SM	0.949(1.748)	0.947(2.334)	0.955(3.749)	0.944(9.007)
		DM	0.962(3.362)	0.958(16.013)	—	—
50	0.2	CM	0.946(0.253)	0.955(0.325)	0.945(0.466)	0.938(0.879)
		AM	0.960(0.318)	0.966(0.483)	0.979(1.307)	—
		SM	0.952(0.249)	0.955(0.321)	0.948(0.458)	0.942(0.862)
		DM	0.961(0.334)	0.962(0.562)	0.958(3.401)	—
	1	CM	0.951(1.274)	0.950(1.605)	0.948(2.361)	0.944(4.5702)
		AM	0.959(1.601)	0.968(2.382)	0.984(6.596)	—
		SM	0.955(1.253)	0.949(1.582)	0.953(2.315)	0.945(4.497)
		DM	0.958(1.684)	0.960(2.774)	0.955(17.166)	—

consider the cases that each approach reject the null hypothesis at the nominal level, $\alpha = 0.05$. The sizes and powers of the methods are given in Table 2. In this table, the actual sizes are determined when $|\Sigma| = 0.2$, and the powers are determined when $|\Sigma| = 1$. We can conclude that

1. The actual sizes of our method and Sarkar method are close to the nominal level, $\alpha = 0.05$.

2. The actual size of Djauhari method is greater than the nominal level, $\alpha = 0.05$.

3. The actual size of Anderson method is satisfactory when p is small, and is very conservative when p is large.

Table 2: Actual sizes and powers of the tests for $|\Sigma|$ at 5% significance level.

	$\|\Sigma\|$	0.2				1.0			
		p				p			
n	method	2	3	5	10	2	3	5	10
15	CM	0.045	0.050	0.052	0.056	0.885	0.747	0.535	0.288
	AM	0.047	0.036	0.010	0.001	0.892	0.717	0.321	0.001
	SM	0.033	0.040	0.050	0.052	0.871	0.726	0.518	0.275
	DM	0.063	0.064	0.062	0.033	0.906	0.784	0.565	0.214
30	CM	0.049	0.045	0.060	0.042	0.998	0.953	0.834	0.524
	AM	0.058	0.039	0.029	0.001	0.998	0.945	0.739	0.057
	SM	0.045	0.039	0.059	0.038	0.998	0.942	0.830	0.513
	DM	0.073	0.063	0.077	0.045	0.998	0.966	0.861	0.549
50	CM	0.039	0.038	0.046	0.051	1.000	0.999	0.957	0.775
	AM	0.051	0.041	0.028	0.002	1.000	0.999	0.931	0.336
	SM	0.036	0.038	0.043	0.047	1.000	0.999	0.955	0.768
	DM	0.064	0.053	0.061	0.060	1.000	0.999	0.972	0.800

4. The power of Djauhari method is larger than other methods.

5. The power of all methods decrease when p increases.

6. The power of all methods increase when n increases.

3.2 Numerical Examples

Example 1. Six hematology variables are measured on 51 workers (Royston 1983). In addition, the data are given by Rencher (2002, page 109). The sample determinant of covariance matrix, $|s|$, is equal to 6.2453. The %95 confidence intervals for $|\Sigma|$ by different methods are given in Table 3. In addition, for the hypothesis test $H_0 : |\Sigma| \leq 6.0$ vs. $H_1 : |\Sigma| > 6.0$, the p-values are given in Table 3. Therefore, we cannot reject H_0 at 5% significance level. We note that Algorithm 1 has been used by $m = 10000$.

Example 2. Timm (1975, page 233) reported the results of an experiment where $n = 11$ subjects responded to "probe words" at $p = 5$ positions in a sentence. In addition, the data are given by Rencher (2002, page 70). The sample determinant of covariance matrix, $|s|$, is equal to

Table 3: p-values and 95% confidence intervals for generalized variance, $|\Sigma|$.

method	p-value	Confidence interval	length
Computational	0.3356	(3.6022, 14.4061)	10.8039
Anderson	0.4617	(3.6928, 19.3117)	15.6189
Sarkar	0.3301	(3.5051, 14.0165)	10.5113
Djauhari	0.3915	(3.8766, 21.9662)	18.0896

2.7231. The %95 confidence intervals for $|\Sigma|$ by different methods are given in Table 4. Note that we cannot compute the confidence intervals based on Andesron approximation and Djauhari approximation. In addition, for the hypothesis test $H_0 : |\Sigma| \leq 2.7$ vs. $H_1 : |\Sigma| > 2.7$, the p-values are given in Table 4. Therefore, we can reject H_0 at 5% significance level. We note that Algorithm 1 has been used by $m = 10000$.

Table 4: p-values and 95% confidence intervals for generalized variance, $|\Sigma|$.

method	p-value	Confidence interval	length
Computational	0.0537	(1.8612, 226.1532)	224.2919
Anderson	0.0866	—	—
Sarkar	0.0612	(1.6293, 191.6412)	190.0119
Djauhari	0.0553	—	—

4 Conclusion

For testing the hypothesis and constructing confidence interval for the generalized variance in a multivariate normal distribution, we proposed a computational method using the distribution of the sample generalized variance. We compare this computational approach with three other approximate approaches based on the coverage probability and expected length for constructing confidence interval, and based on the actual size and power for testing the hypothesis. For a range of choices of the sample size and parameter configurations, we have investigated the performance of the above approaches using Monte Carlo simulation, and we concluded that the computational method is satisfactory and better than other methods.

Acknowledgments

The authors are thankful to the Editor and two referees for helpful comments and suggestions.

References

Anderson, T. W. (2003), An Introduction to Multivariate Statistical Analysis. 3rd Edition, New York: John Wiley and Sons.

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

Djauhari, M. A. (2007), A measure of multivariate data concentration. Journal of Applied Probability & Statistics, **2**(2), 139-155.

Djauhari, M. A. (2009), Asymptotic distribution of sample covariance determinant. Matematika, **25**, Number 1, 79-85.

Djauhari, M. A., Herwindiati, D. E., and Mashuri, M. (2008), Multivariate Process Variability Monitoring. Communications in Statistics-Theory and Methods, **37**(11), 1742-1754.

Hoel, P. G. (1937), A significance test for component analysis. The Annals of Mathematical Statistics, **8**, 149-158.

Iliopoulos, G. and Kourouklis, S. (1998), On improved interval estimation for the generalized variance. Journal of Statistical Planning and Inference, **66**, 305-320.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994), Continuous Univariate Distributions - 1. 2nd ed. New York: Wiley.

Kokonendji, C. C. (2003), On UMVU estimator of the generalized variance for natural exponential families. Monograf Seminario Mat. Garcia Galdeano, **27**, 353-360.

Kokonendji, C. C. and Pommeret, D. (2007), Comparing UMVU and ML estimators of the generalized variance for natural exponential families. Statistics, **41**, 547-558.

Rencher, A. C. (2002), Methods of Multivariate Analysis. 2nd Editon, New York: John Wiley and Sons.

Royston, J. P. (1983), Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. Applied Statistics, **32**, 121-133.

Sarkar, S. K. (1989), On improving the shortest length confidence interval for the generalized variance. Journal of Multivariate Analysis, **31**, 136-147.

Sarkar, S. K. (1991), Stein type improvements of confidence intervals for the generalized variance. The Annals of Mathematical Statistics, **43**, 369-375.

Timm, N. H. (1975), Multivariate Analysis: With Applications in Education and Psychology. Monterey, Calif.: Brooks/Cole.

Wilks, S. S. (1932), Certain generalizations in the analysis of variance. Biometrika, **24**, 471-494.