The Sphericity Test For Samples with Monotone Missing Data

Evelina Veleva

Abstract: *We consider samples with monotone missing data, drawn from a normal populaton to test if the covariance matrix is proportional to a given positive definite matrix. We propose an imputation procedure for the missing data and give the exact distribution of the corresponding likelihood ratio test statistic from the classical complete case.*

Key words: *monotone missing data, Bellman gamma distribution, covariance matrix, hypotheses testing*

INTRODUCTION

The problem of missing data is an important applied problem, because missing values are encountered in many practical situations (see [4]). Two commonly used approaches to analyze incomplete data are the likelihood based approach and the multiple imputation. The imputation method is to impute the missing data to form complete data and then use the standard methods available for the complete data analysis. For a good exposition of the imputation procedures and validity of imputation inferences in practice, we refer to [5].

In this paper we consider samples with monotone missing data pattern. Let $(X_1,\ldots,X_n)^t$ be a random vector with multivariate normal distribution $N_n(\mu,\Sigma)$, where the mean vector μ and the covariance matrix Σ are unknown. Suppose that we have $k_1+\cdots+k_n$ independent observations, k_1 of which are on $(X_1,\ldots,X_n)^t$, k_2 - on $(X_2,\ldots,X_n)^t$ and so on, k_n on the random variable X_n. Assume that $k_j \geq 0$ and $m_j = k_1+\cdots+k_j > j$, $j=1,\ldots,n$. The data can be written in the following pattern, known as a monotone pattern

$$\begin{matrix} x_{1,1} & \cdots & x_{1,m_1} & & & & & & \\ \vdots & & \vdots & & & & & & \\ x_{n-1,1} & \cdots & x_{n-1,m_1} & x_{n-1,m_1+1} & \cdots & x_{n-1,m_{n-1}} & & & \\ x_{n,1} & \cdots & x_{n,m_1} & x_{n,m_1+1} & \cdots & x_{n,m_{n-1}} & x_{n,m_{n-1}+1} & \cdots & x_{n,m_n} \end{matrix} \quad . \tag{1}$$

In the literature on inference for μ and Σ, it is noticeable that the exact distributions of $\hat{\mu}$ and $\hat{\Sigma}$, the maximum likelihood estimators of μ and Σ, have remained unknown. This problem is basic to inference with incomplete data when large samples are infeasible or impractical (see [1]). In [1] the authors initiate a program of research on inference for μ and Σ with the goal of deriving explicit results analogous to those existing in the classical complete case.

In this paper we consider hypotheses $H_0 : \Sigma = \sigma^2\Sigma_0$ against $H_a : \Sigma \neq \sigma^2\Sigma_0$, where Σ_0 is known positive definite matrix and σ^2 is unknown positive constant, propose an imputation procedure for the missing data in (1) and give the exact distribution of the corresponding likelihood ratio test statistic from the classical complete case.

PRELIMINARY NOTES

Let us denote by $\mathbf{z}_i$ the vector of available observations on X_i in (1), i.e. $\mathbf{z}_i = (x_{i,1},\ldots,x_{i,m_i})^t$, $i=1,\ldots,n$. Let $\bar{\mathbf{z}}_i$ be the mean of the elements of the vector $\mathbf{z}_i$, $i=1,\ldots,n$. Consider the data matrix

$$\mathbf{X}=\begin{pmatrix} x_{1,1} & \cdots & x_{1,m_1} & \bar{\mathbf{z}}_1 & \cdots & \bar{\mathbf{z}}_1 & \bar{\mathbf{z}}_1 & \cdots & \bar{\mathbf{z}}_1 \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{n-1,1} & \cdots & x_{n-1,m_1} & x_{n-1,m_1+1} & \cdots & x_{n-1,m_{n-1}} & \bar{\mathbf{z}}_{n-1} & \cdots & \bar{\mathbf{z}}_{n-1} \\ x_{n,1} & \cdots & x_{n,m_1} & x_{n,m_1+1} & \cdots & x_{n,m_{n-1}} & x_{n,m_{n-1}+1} & \cdots & x_{n,m_n} \end{pmatrix}, \qquad (2)$$

in which we substitute $\bar{\mathbf{z}}_1$, …, $\bar{\mathbf{z}}_{n-1}$ for the missing values in 1, …, $n-1$'th row respectively of the data in (1). The next Proposition is proved in [7].

Proposition 1. *Let the matrix* $\mathbf{X}$*, defined by (2) presents the observations (1) on a random vector* $(X_1,\ldots,X_n)^t$ *with multivariate normal distribution* $N_n(\mu,\mathrm{I}_n)$*, where* I_n *is the identity matrix of size* n*. Let* $\mathbf{x}_i$*,* $i=1,\ldots,m_n$ *be the column vectors of the matrix* $\mathbf{X}$*,* $\bar{\mathbf{x}}$ *be the vector* $\bar{\mathbf{x}}=(\bar{\mathbf{z}}_1,\ldots,\bar{\mathbf{z}}_n)^t$ *and* $\mathbf{S}$ *be the matrix*

$$\mathbf{S}=\sum_{i=1}^{m_n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^t. \qquad (3)$$

Then $\bar{\mathbf{x}}$ *and* $\mathbf{S}$ *are independent,* $\bar{\mathbf{x}}\sim N_n(\mu,diag(m_1^{-1},\ldots,m_n^{-1}))$ *and* $\mathbf{S}$ *has Bellman gamma distribution* $BG_n^l\left(\frac{m_1-1}{2},\ldots,\frac{m_n-1}{2};\frac{1}{2}\mathrm{I}_n\right)$.

The Bellman gamma distribution is a matrix variate distribution, which is a generalization of the Wishart and the matrix gamma distributions. The next definition is given in [3]. By $\Gamma_n^*(a_1,\ldots,a_n)$ is denoted the generalized multivariate gamma function,

$$\Gamma_n^*(a_1,\ldots,a_n)=\pi^{n(n-1)/4}\prod_{j=1}^{n}\Gamma\left(a_j-\frac{1}{2}(j-1)\right),\ \text{for}\ a_j>\frac{1}{2}(j-1),\ j=1,\ldots,n.$$

If A is a square matrix of order n, by $\det \mathrm{A}[\{i_1,\ldots,i_k\}]$ is denoted the principal minor of A, composed of the rows and columns of A with numbers $i_1,\ldots,i_k$, $1\le i_1<\cdots<i_k\le n$.

Definition 1. *A random positive definite matrix* $\mathbf{U}$ *(*$n\times n$*) is said to follow Bellman gamma type I distribution, denoted by* $\mathbf{U}\sim BG_n^l(a_1,\ldots,a_n;\mathrm{C})$*, if its probability density function is given by*

$$\frac{\prod_{i=1}^{n}(\det \mathrm{C}[\{i,\ldots,n\}])^{a_i-a_{i-1}}}{\Gamma_n^*(a_1,\ldots,a_n)}\,\frac{(\det \mathrm{U})^{a_n-(n+1)/2}}{\prod_{i=2}^{n}(\det \mathrm{U}[\{1,\ldots,i-1\}])^{a_i-a_{i-1}}}\,etr(-\mathrm{CU}),$$

where C *(*$n\times n$*) is a positive definite constant matrix,* $a_0=0$ *and* $a_j>\frac{1}{2}(j-1)$*,* $j=1,\ldots,n$.

The next two Propositions can be found in [8].

Proposition 2. *Let* $\mathbf{U}\sim BG_n^l(a_1,\ldots,a_n;\mathrm{C})$ *and* L *be an arbitrary lower triangular constant matrix of size* n*. Then the matrix* $\mathbf{W}=\mathrm{L}\mathbf{U}\mathrm{L}^t$ *has distribution* $BG_n^l(a_1,\ldots,a_n;(\mathrm{L}^t)^{-1}\mathrm{C}\mathrm{L}^{-1})$.

Let $P(n,\Re)$ be the set of all real, symmetric, positive definite matrices of order n. Let us denote by $D(n,\Re)$ the set of all real, symmetric matrices of order n, with positive diagonal elements, which off-diagonal elements are in the interval (-1,1). There exist a bijection (one-to-one correspondence) $h: D(n,\Re)\to P(n,\Re)$ (see [6]). image of an arbitrary matrix $\mathrm{X}=(x_{i,j})$ from $D(n,\Re)$ by the bijection h, is a matrix $\mathrm{Y}=(y_{i,j})$ from $P(n,\Re)$, such that

$$y_{j,j}=x_{j,j},\ j=1,\ldots,n, \qquad (4)$$

$$y_{1,j} = x_{1,j}\sqrt{x_{1,1}x_{j,j}},\ \ j = 2,\ldots,n,$$

$$y_{i,j} = \sqrt{x_{ij}x_{j,j}}\left[\sum_{r=1}^{i-1}\left(x_{r,i}x_{r,j}\prod_{q=1}^{r-1}\sqrt{(1-x_{q,i}^2)(1-x_{q,j}^2)}\right) + x_{i,j}\prod_{q=1}^{i-1}\sqrt{(1-x_{q,i}^2)(1-x_{q,j}^2)}\right],\ 2 \le i < j \le n.$$

The next relation between the elements of the matrices X and Y is proved in [6]:

$$\det \mathrm{Y} = x_{1,1}\ldots x_{n,n}\left(\prod_{1\le i<j\le n}(1-x_{i,j}^2)\right). \tag{5}$$

Proposition 3. *Let a_1, …, a_n be real numbers, such that $a_j > \frac{1}{2}(j-1)$, $j = 1,\ldots,n$. Let* $\boldsymbol{\xi}=(\xi_{i,j})$ *be a symmetric $n\times n$ random matrix. Suppose that $\xi_{i,j}$, $1\le i \le j \le n$ are mutually independent, $\xi_{i,j} \sim Beta\left(a_j - i/2, a_j - i/2, -1, 1\right)$, $1\le i < j \le n$ and $\xi_{i,i} \sim G(a_i, 1)$, $i = 1,\ldots,n$. Let* $\mathbf{U}$ *be the matrix* $\mathbf{U} = h(\boldsymbol{\xi})$. *Then the matrix* $\mathbf{U}$ *has Bellman gamma type I distribution* $BG_n^I\left(a_1,\ldots,a_n;\mathrm{I}_n\right)$.

THE SPHERICITY TEST

For the data in (1), let us consider the hypotheses $H_0 : \Sigma = \sigma^2\Sigma_0$ against $H_a : \Sigma \ne \sigma^2\Sigma_0$, where Σ_0 is a known positive definite matrix of size n and σ^2 is a unknown constant. It is shown in [4], that the testing problem is invariant under a suitable transformation of the data in (1) and without loss of generality we can assume that $\Sigma_0 = \mathrm{I}_n$.

It is easy to see that under $H_0 : \Sigma = \sigma^2\mathrm{I}_n$, the maximum likelihood estimations for the missing values in the j'th row in (1) are equal to $\bar{\mathbf{z}}_j$, $j = 1,\ldots,n$. The matrix $\frac{1}{(m_n-1)}\mathbf{S}$ is actually the empirical covariance matrix, obtained from the data matrix (2).

Let us consider the modified likelihood ratio test statistic λ^*

$$\lambda^* = n^{(m_n-1)n/2}\frac{(\det \mathbf{S})^{(m_n-1)/2}}{(tr\mathbf{S})^{(m_n-1)n/2}},$$

which is unbiased in the classical case of fully observed data matrix (see [2]). From Propositions 1 and 2 it follows that under $H_0 : \Sigma = \sigma^2\mathrm{I}_n$, the distribution of the matrix $\mathbf{S}$ is $BG_n^I\left(\frac{m_1-1}{2},\ldots,\frac{m_n-1}{2};\frac{1}{2\sigma^2}\mathrm{I}_n\right)$. Consequently, applying again Proposition 2, the distribution of $\frac{1}{2\sigma^2}\mathbf{S}$ is $BG_n^I\left(\frac{m_1-1}{2},\ldots,\frac{m_n-1}{2};\mathrm{I}_n\right)$. Therefore $\mathbf{S} = 2\sigma^2 h(\boldsymbol{\xi})$, where the matrix $\boldsymbol{\xi}$ is defined by Proposition 3, with $a_j = (m_j-1)/2$, $j = 1,\ldots,n$. Applying the relations (4) and (5) we get

$$tr\mathbf{S} = tr\left(2\sigma^2 h(\boldsymbol{\xi})\right) = 2\sigma^2 tr\left(h(\boldsymbol{\xi})\right) = 2\sigma^2(\xi_{1,1}+\cdots+\xi_{n,n}),$$

$$\det \mathbf{S} = 2^n\sigma^{2n}\det h(\boldsymbol{\xi}) = 2^n\sigma^{2n}\xi_{1,1}\ldots\xi_{n,n}\left(\prod_{1\le i<j\le n}(1-\xi_{i,j}^2)\right).$$

It is shown in [8] that $\prod_{1\le i<j\le n}(1-\xi_{i,j}^2)$ is distributed as the product $\zeta_1\ldots\zeta_{n-1}$ of $n-1$ mutually independent random variables, $\zeta_j \sim Beta\left((m_{j+1}-j-1)/2, j/2, 0, 1\right)$, $j = 1,\ldots,n-1$..

Consequently, under $H_0 : \Sigma = \sigma^2 \mathrm{I}_n$ the variable $\tau = n^{-n} \lambda^{*\frac{2}{m_n - 1}} = \frac{\det \mathbf{S}}{(tr\mathbf{S})^n}$ is distributed as the product

$$\tau = \frac{\det \mathbf{S}}{(tr\mathbf{S})^n} \sim \zeta_1 \dots \zeta_{n-1} \frac{\eta_1 \dots \eta_n}{(\eta_1 + \dots + \eta_n)^n}, \tag{6}$$

where $\zeta_1, \dots, \zeta_{n-1}, \eta_1, \dots, \eta_n$ are mutually independent, $\zeta_j \sim Beta\left((m_{j+1} - j - 1)/2, j/2, 0, 1\right)$, $j = 1, \dots, n-1$ and $\eta_j \sim G((m_j - 1)/2, 1)$, $j = 1, \dots, n$.

CONCLUSIONS

Since we obtain the exact distribution of the variable $\tau = n^{-n} \lambda^{*\frac{2}{m_n - 1}}$, it can be used for testing the hypothesis of sphericity of the covariance matrix for small samples with monotone missing data.

REFERENCES

[1] Wan – Ying Chang, Donald St.P. Richards, Finite – sample inference with monotone incomplete multivariate normal data I, Journal of Multivariate Analysis, In Press, Corrected Proof, Available online 14 May 2009.

[2] N.C. Giri, Multivariate Statistical Analysis, Marcel Dekker Inc., New York, 2004.

[3] A.K. Gupta and D.K. Nagar, Matrix variate distributions, Chapman & Hall/CRC, 2000.

[4] J. Hao and K. Krishnamoorthy, Inferences on a normal covariance matrix and generalized variance with monotone missing data, J. Multivariate Anal., 78 (2001), 62 – 82.

[5] J.A. Little and D.B. Rubin, Statistical Analysis With Missing Data, 2nd edition, Wiley – Interscience, Hoboken, NJ, 2002.

[6] E. Veleva, A representation of the Wishart distribution by functions of independent random variables, Annuaire de l'Univerite de Sofia "St.Kliment Ohridski", Faculte des Sciences Economiques et de Gestion, Vol. 6, (2007), 59 – 68.

[7] E. Veleva, Testing a normal covariance matrix for small samples with monotone missing data, Applied Mathematical Sciences, Vol. 3, No. 54, (2009), 2671 - 2679.

[8] E. Veleva, Stochastic representations of the Bellman gamma distribution, Int. Conf. on Theory and Applications in Mathematics and Informatics, ICTAMI 2009, Alba Iulia, Romania, 3 – 6 September, 2009.

ABOUT THE AUTHOR

Principal Assistant, Evelina Veleva, Department of Numerical Methods and Statistics, University of Rousse, Phone: +359 82 888 466, E-mail: eveleva@ru.acad.bg.

Докладът е рецензиран.