The Sphericity Test  
For Samples with Monotone Missing Data  

Evelina Veleva  

Abstract: We consider samples with monotone missing data, drawn from a normal population 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 \(\mu\) and the covariance matrix \(\Sigma\) are unknown. Suppose that we have \(k_1 + \cdots + k_n\) independent observations, \(k_i\) 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  
\[
\begin{array}{cccc}
  x_{1,1} & \cdots & x_{1,m_1} \\
  \vdots & & \vdots \\
  x_{n-1,1} & \cdots & x_{n-1,m_1} & \cdots & x_{n-1,m_{n-1}} \\
  x_{n,1} & \cdots & x_{n,m_1} & \cdots & x_{n,m_{n-1}} & \cdots & x_{n,m_n} \\
\end{array}
\]  
(1)  

In the literature on inference for \(\mu\) and \(\Sigma\), it is noticeable that the exact distributions of \(\hat{\mu}\) and \(\hat{\Sigma}\), the maximum likelihood estimators of \(\mu\) and \(\Sigma\), 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 \(\mu\) and \(\Sigma\) 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 \(\Sigma_0\) is known positive definite matrix and \(\sigma^2\) 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 \(z_i\) the vector of available observations on \(X_i\) in (1), i.e. \(z_i = (x_{i,1},\ldots,x_{i,m_i})^T\), \(i = 1,\ldots,n\). Let \(z_i\) be the mean of the elements of the vector \(z_i\), \(i = 1,\ldots,n\). Consider the data matrix
in which we substitute \( \bar{z}_1, \ldots, \bar{z}_{n-1} \) for the missing values in \( 1, \ldots, n-1 \)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 \( (X_1, \ldots, X_n) \) with multivariate normal distribution \( N_n(\mathbf{I}, \mathbf{I}) \), where \( \mathbf{I} \) 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 \( \bar{\mathbf{x}}=(\bar{z}_1, \ldots, \bar{z}_n) \) and \( \mathbf{S} \) be the matrix

\[
\mathbf{S} = \sum_{i=1}^{m_n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.
\]

Then \( \bar{\mathbf{x}} \) and \( \mathbf{S} \) are independent, \( \bar{\mathbf{x}} \sim N_n(\mu, \text{diag}(m_1, \ldots, m_n)) \) and \( \mathbf{S} \) has Bellman gamma distribution \( \text{BG}_{m_n}^n\left(\frac{m_1-1}{2}, \ldots, \frac{m_n-1}{2}, \frac{1}{2} \mathbf{I}_n\right) \).

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 \( \Gamma_n^\ast(a_1, \ldots, a_n) \) is denoted the generalized multivariate gamma function,

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

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

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

\[
\frac{\prod_{i=1}^{n} \left( \text{det} \mathbf{C}[i, \ldots, n] \right)^{a_i-a_{i-1}}}{\Gamma_n^\ast(a_1, \ldots, a_n)} \frac{\left( \text{det} \mathbf{U} \right)^{n-(n+1)/2}}{\prod_{j=2}^{n} \left( \text{det} \mathbf{U}[\{1, \ldots, j-1]\] \right)^{a_j-a_{j-1}}} \text{etr}(-\mathbf{C} \mathbf{U}),
\]

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

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

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

Let \( \mathcal{P}(n, \mathbb{R}) \) be the set of all real, symmetric, positive definite matrices of order \( n \). Let us denote by \( \mathcal{D}(n, \mathbb{R}) \) 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 \((\text{one-to-one correspondence})\) \( h : \mathcal{D}(n, \mathbb{R}) \rightarrow \mathcal{P}(n, \mathbb{R}) \) (see [6]). image of an arbitrary matrix \( \mathbf{X}=(x_{i,j}) \) from \( \mathcal{D}(n, \mathbb{R}) \) by the bijection \( h \), is a matrix \( \mathbf{Y}=(y_{i,j}) \) from \( \mathcal{P}(n, \mathbb{R}) \), such that

\[
y_{j,i} = x_{j,i}, \quad j = 1, \ldots, n,
\]
The next relation between the elements of the matrices $X$ and $Y$ is proved in [6]:

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

(5)

**Proposition 3.** Let $a_1, \ldots, a_n$ be real numbers, such that $a_j > \frac{1}{2}(j-1), \ j = 1, \ldots, n$. Let $\xi = (\xi_{i,j})$ be a symmetric $n \times n$ random matrix. Suppose that $\xi_{i,j} \sim \text{Beta}(a_j - i/2, a_j - i/2 - 1), 1 \leq i < j \leq n$ and $\xi_{i,i} \sim \text{G}(a_1, 1), j = 1, \ldots, n$. Let $U$ be the matrix $U = h(\xi)$. Then the matrix $U$ has Bellman gamma type I distribution $BG_i(\sigma_{a}, \ldots, \sigma_{a}; I_{n})$.

**THE SPHERICITY TEST**

For the data in (1), let us consider the hypotheses $H_0 : \Sigma = \sigma^2 I_n$ against $H_a : \Sigma \neq \sigma^2 I_n$, where $\Sigma_0$ is a known positive definite matrix of size $n$ and $\sigma^2$ is an 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 $\Sigma_0 = I_n$.

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

Let us consider the modified likelihood ratio test statistic $\lambda^*$:

$$\lambda^* = n^{(m_n - 1)n/2} \frac{(\det S)^{(m_n - 1)/2}}{(tr S)^{(m_n - 3)n/2}},$$

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

$$tr S = tr (2\sigma^2 h(\xi)) = 2\sigma^2 tr (h(\xi)) = 2\sigma^2 (\xi_{11} + \cdots + \xi_{n,n}),$$

$$\det S = 2^n \sigma^{2n} \det h(\xi) = 2^n \sigma^{2n} \xi_{11} \cdots \xi_{n,n} \left( \prod_{1 \leq i < j \leq n} (1 - \xi_{i,j}^2) \right).$$

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

- 14 -
Consequently, under $H_0: \Sigma = \sigma^2 I_n$, the variable $\tau = n^{-n} \lambda^2 \frac{m_n}{m_n - 1} \frac{\det S}{(tr S)^n}$ is distributed as the product

$$\tau = \frac{\det S}{(tr S)^n} \sim \xi_1 \cdots \xi_{n-1} \prod_{j=1}^{n} \eta_j \prod_{j=1}^{n} (\eta_1 + \cdots + \eta_n)^n,$$

where $\xi_1, \ldots, \xi_{n-1}$, $\eta_1, \ldots, \eta_n$ are mutually independent, $\xi_j \sim Beta((m_{j,1} - j - 1)/2, j/2,0,1)$, $j = 1,\ldots,n-1$ and $\eta_j \sim G((m_j - 1)/2,1)$, $j = 1,\ldots,n$.

CONCLUSIONS

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

REFERENCES


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.