Journal of Statistics: Advances in Theory and Applications
Volume 26, Number 1, 2021, Pages 1-11
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100122218

MAXIMUM LIKELIHOOD ESTIMATION FOR MULTIVARIATE NORMAL POPULATION WITH MISSING DATA USING AUXILIARY INFORMATION

JIANQI YU, SHAOLING DING
and XIANG WANG

Guilin University of Technology
Guilin, Guangxi 540041
P. R. China
e-mail: jxy2416@yahoo.com

Abstract

Closed forms are obtained for maximum likelihood estimates (MLE) of multivariate normal with missing data using auxiliary information. The likelihood function is obtained as product of several independent normal and conditional normal likelihood functions. The parameters are transformed into a new set of parameters of which the MLEs are easy to derive. Since the MLEs are invariant, the MLEs of the original parameters are derived using the inverse transformation.

2020 Mathematics Subject Classification: 62A01, 62H10.
Keywords and phrases: auxiliary variables, incomplete data, maximum likelihood estimator.
Received July 29, 2021

1. Introduction

In statistical practice, missing data arises in many situations, especially in public survey. For example, during data gathering and recording, when the experiment is involved a group of individuals over a period of time like in clinical trials or in a planned experiment where the variables that are expensive to measure are collected only from a subset of a sample. The causes for missing data are not our concern, but to ignore the process that causes missing data it is assumed that the data are missing at random (MAR). Lu and Copas [13] pointed out that inference from the likelihood method ignoring the missing data mechanism is valid if and only if the missing data mechanism is MAR. For formal definition and exposition of MAR or missing completely at random, we refer to Little and Rubin [11] or Little [12].

Anderson [1], one of the earliest papers in this area, gives a simple approach to derive the MLEs and present them for a special case of monotone pattern. Krishnamoorthy and Pannala [6, 8] provided an accurate, simple approach to construct a confidence region for a normal mean vector. Hao and Krishnamoorthy [3] developed an inferential procedure on a normal covariance matrix. Yu et al. [23] considered the problem of testing equality of two normal mean vectors with the assumption that the two covariance matrices are equal. Based on the work of Krishnamoorthy and Yu [9] on the multivariate Behrens Fisher problem with complete data, Krishnamoorthy and Yu [10] provided an approximate solution to the multivariate Behrens-Fisher problem with missing data. Krishnamoorthy [7] considered the inference on simple correlation coefficients with monotone missing data, and pointed out that the inference based on incomplete samples and those based on samples after listwise or pairwise deletion are similar, and the loss of efficiency by ignoring additional data is not appreciable. Yagi et al. [20] obtained an asymptotic solution of ANOVA with monotone missing data using Taylor expansion. Yu et al. [24] considered the problem of testing equality of two normal covariance matrices with monotone missing data. Yamada et al.

[21] used the Mardia statistics to test whether the monotone missing data is multivariate normal, while Raykov et al. [15] solved the same problem through combining several simple tests.

On the other hand, auxiliary variables are also very common in practice. They are usually highly related to the research variables. For example, if the research variable is sleeping time, auxiliary variables can be age, blood pressure, gender, etc. Making full use of auxiliary information can effectively improve the accuracy of inference. For instance, we usually use sample mean to estimate the population mean. However, if there are auxiliary variables, other estimates using the auxiliary information are much better. Cochran [2] proposed the ratio estimation of the population mean in simple random sample survey, and pointed out that the ratio estimation reached the best when the research variables and auxiliary variables were highly positively correlated and the regression line passed through the origin. The product estimation was first proposed by Robson [16] and rediscovered by Murthy [14], which is suitable for the situation where the research variables and auxiliary variables are highly negatively correlated. The regression estimation proposed by Watson [19] is suitable for the case that the regression line of the research variable and auxiliary variable does not pass through the origin. In later years, many scholars proposed various methods to improve the estimation of population mean in Simple Random Sampling. For details, see Singh and Tailor [17], Singh et al. [18], Yan and Tian [22], Khan et al. [5] Kadilar [4], etc.

In this paper, we consider the MLEs for multivariate normal with incomplete data in presence of auxiliary information. Suppose that the p dimensional variable $\boldsymbol{y}$ with expectation $\boldsymbol{\mu}$ and covariance matrix $\sum$ is the research variable that we are interested in, but its observations are not complete. In addition, suppose that there is q dimensional auxiliary variables $\boldsymbol{x}$ with a known expectation c. For simplicity, we assume that the samples of $\boldsymbol{y}$ are of two-step monotone missing pattern. It is easy to generalize the ideas and results to higher steps, but the notation will become very complicated.

2. Maximum Likelihood Estimation

Let the research variable $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}$, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are p_1 and p_2 dimensional respectively and $p_1 + p_2 = p$. Let the auxiliary variable $\boldsymbol{x}$ be q-dimensional. Assume that

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \sim N_{q+p_1+p_2}\left(\begin{pmatrix} \boldsymbol{c} \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sum_{11} & \sum_{12} & \sum_{13} \\ \sum_{21} & \sum_{22} & \sum_{23} \\ \sum_{31} & \sum_{32} & \sum_{33} \end{pmatrix} \right),$$

where $\boldsymbol{c}$ is a known constant and

$$\sum = \begin{pmatrix} \sum_{22} & \sum_{23} \\ \sum_{32} & \sum_{33} \end{pmatrix}.$$

Suppose there are n_2 independent samples on $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$, but there are $n_2 - n_3$ samples missing on $\boldsymbol{y}_2$, the second component of the research variable. Moreover, there are additional $m = n_1 - n_2$ observations solely on $\boldsymbol{x}$. In other words, we have a random sample like this:

$$\begin{aligned} &\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n_3} \quad \boldsymbol{x}_{n_3+1}, \dots, \boldsymbol{x}_{n_2} \quad \boldsymbol{x}_{n_2+1}, \dots, \boldsymbol{x}_{n_1}, \\ &\boldsymbol{y}_{11}, \dots, \boldsymbol{y}_{1,n_3} \quad \boldsymbol{y}_{1,n_3+1}, \dots, \boldsymbol{y}_{1,n_2}, \\ &\boldsymbol{y}_{21}, \dots, \boldsymbol{y}_{2,n_3}, \end{aligned} \tag{1}$$

where $n_1 = n_2 + m$ and the sample of the research variable is incomplete.

Partition the data in (1) as follows:

$$\mathbf{E}_1 = \left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_3}, \boldsymbol{x}_{n_3+1}, \ldots, \boldsymbol{x}_{n_2}, \boldsymbol{x}_{n_2+1}, \ldots, \boldsymbol{x}_{n_1}\right),$$

$$\mathbf{E}_2 = \begin{pmatrix} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_3} & \boldsymbol{x}_{n_3+1}, \ldots, \boldsymbol{x}_{n_2} \\ \boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{1,n_3} & \boldsymbol{y}_{1,n_3+1}, \ldots, \boldsymbol{y}_{1,n_2} \end{pmatrix},$$

$$\mathbf{E}_3 = \begin{pmatrix} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_3} \\ \boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{1,n_3} \\ \boldsymbol{y}_{21}, \ldots, \boldsymbol{y}_{2,n_3} \end{pmatrix}. \tag{3}$$

Denote the sample mean vector and the sum of squares and sum of products matrix based on $\mathbf{E}_i$ by $\left(\bar{\mathbf{E}}_i, V_i\right)$, $i = 1, 2, 3$, and partition these means and matrices accordingly as follows:

$$\bar{\mathbf{E}}_1 = \bar{\mathbf{E}}_1^{(1)}, \ \bar{\mathbf{E}}_2 = \begin{pmatrix} \bar{\mathbf{E}}_2^{(1)} : q \times 1 \\ \bar{\mathbf{E}}_2^{(2)} : p_1 \times 1 \end{pmatrix}, \ \bar{\mathbf{E}}_3 = \begin{pmatrix} \bar{\mathbf{E}}_3^{(1)} : q \times 1 \\ \bar{\mathbf{E}}_3^{(2)} : p_1 \times 1 \\ \bar{\mathbf{E}}_3^{(3)} : p_2 \times 1 \end{pmatrix},$$

$$\mathbf{V}_1 = \mathbf{V}_1^{(1,1)}, \quad \mathbf{V}_2 = \begin{pmatrix} \mathbf{V}_2^{(1,1)} : q \times q & \mathbf{V}_2^{(1,2)} : q \times p_1 \\ \mathbf{V}_2^{(2,1)} : p_1 \times q & \mathbf{V}_2^{(2,2)} : p_1 \times p_1 \end{pmatrix},$$

and

$$\mathbf{V}_3 = \begin{pmatrix} \mathbf{V}_3^{(1,1)} : q \times q & \mathbf{V}_3^{(1,2)} : q \times p_1 & \mathbf{V}_3^{(1,3)} : q \times p_2 \\ \mathbf{V}_3^{(2,1)} : p_1 \times q & \mathbf{V}_3^{(2,2)} : p_1 \times p_1 & \mathbf{V}_3^{(2,3)} : p_1 \times p_2 \\ \mathbf{V}_3^{(3,1)} : p_2 \times q & \mathbf{V}_3^{(3,2)} : p_2 \times p_1 & \mathbf{V}_3^{(3,3)} : p_2 \times p_2 \end{pmatrix}.$$

Define

$$\mathbf{B}_{2.1} = \sum_{21}\sum_{11}^{-1},\ \mu_{2.1} = \mu - \mathbf{B}_{2.1}c,\ \sum_{2.1} = \sum_{22} - \mathbf{B}_{2.1}\sum_{12}$$

$$\mathbf{B}_{3.21} = \left(\sum_{31}, \sum_{32}\right)\begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}^{-1},\ \mu_{3.21}$$

$$= \mu - \mathbf{B}_{3.21}\begin{pmatrix} \boldsymbol{c} \\ \mu_1 \end{pmatrix},\ \sum_{3.21} = \sum_{33} - \mathbf{B}_{3.21}\begin{pmatrix} \sum_{13} \\ \sum_{23} \end{pmatrix}. \tag{4}$$

Consider the density function of data in (1). We note that the density of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as the marginal density of $\boldsymbol{x}$ times the conditional density of $\boldsymbol{y}_1$ given $\boldsymbol{x}$ times the conditional density of $\boldsymbol{y}_2$ given $\boldsymbol{x}$, $\boldsymbol{y}_1$ (we denote the density of normal distribution by $n(\cdot)$ here). After some calculation, the likelihood function can be written as

$$L\left(\mu, \sum\right) = \prod_{i=1}^{n_1} n\left(x_i \mid c, \sum_{11}\right) \prod_{i=1}^{n_2} n\left(y_{1i} \mid \mu_{2.1} + \mathbf{B}_{2.1}x_i, \sum_{2.1}\right)$$

$$\times \prod_{i=1}^{n_3} n\left(\boldsymbol{y}_{2i} \mid \mu_{3.21} + \mathbf{B}_{3.21}\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_{1i} \end{pmatrix}, \sum_{3.21}\right). \tag{5}$$

The maximum likelihood estimates of $\sum_{11}$, $\mu_{2.1}$, $\mathbf{B}_{2.1}$, $\sum_{2.1}$, $\mu_{3.21}$, $\mathbf{B}_{3.21}$, $\sum_{3.21}$ are those values that maximize (5). To maximize (5) with respect to $\sum_{11}$, we maximize $\prod_{i=1}^{n+m} n\left(x_i \mid c, \sum_{11}\right)$. This gives us the usual maximum likelihood estimates of the parameters of a normal distribution based on $n + m$ observations, namely,

$$\widehat{\sum}_{11} = \frac{1}{n+m}\sum_{i=1}^{n+m} (\boldsymbol{x}_i - \boldsymbol{c})' (\boldsymbol{x}_i - \boldsymbol{c}). \tag{6}$$

To maximize (5) with respect to $\boldsymbol{\mu}_{2.1}$, $\mathbf{B}_{2.1}$ and $\sum_{2.1}$, we maximize the second term of the right hand side of (5). This gives the usual estimates of regression parameters, namely,

$$\hat{\mathbf{B}}_{2.1} = \mathbf{S}_2^{(2,1)}\left(\mathbf{S}_2^{(1,1)}\right)^{(-1)}$$

$$\hat{\boldsymbol{\mu}}_{2.1} = \overline{\mathbf{E}}_2^{(2)} - \hat{\mathbf{B}}_{2.1}\overline{\mathbf{E}}_2^{(1)}$$

$$\hat{\sum}_{2.1} = \left(\mathbf{S}_2^{(2,2)} - \hat{\mathbf{B}}_{2.1}\mathbf{S}_2^{(1,2)}\right)/\, n_2. \tag{7}$$

To maximize (5) with respect to $\boldsymbol{\mu}_{3.21}$, $\mathbf{B}_{3.21}$ and $\sum_{3.21}$, we maximize the third term of the right hand side of (5). This gives the usual estimates of regression parameters, namely,

$$\hat{\mathbf{B}}_{3.21} = \left(\mathbf{S}_3^{(3,1)},\ \mathbf{S}_3^{(3,2)}\right)\begin{pmatrix} \mathbf{S}_3^{(1,1)} & \mathbf{S}_3^{(1,2)} \\ \mathbf{S}_3^{(2,1)} & \mathbf{S}_3^{(2,2)} \end{pmatrix}^{-1}$$

$$\hat{\boldsymbol{\mu}}_{3.21} = \overline{\mathbf{E}}_3^{(3)} - \hat{\mathbf{B}}_{3.21}\begin{pmatrix} \overline{\mathbf{E}}_3^{(1)} \\ \overline{\mathbf{E}}_3^{(2)} \end{pmatrix}$$

$$\hat{\sum}_{3.21} = \left(\mathbf{S}_3^{(3,3)} - \hat{\mathbf{B}}_{3.21}\begin{pmatrix} \mathbf{S}_3^{(1,3)} \\ \mathbf{S}_3^{(2,1)} \end{pmatrix}\right)/\, n_3. \tag{8}$$

It is easy to see that the maximum likelihood estimates of the original parameters $\boldsymbol{\mu}_1$, $\sum_{12}$, $\sum_{22}$, $\boldsymbol{\mu}_2$, $\sum_{13}$, $\sum_{23}$, $\sum_{33}$ are obtained by solving (4), where $\boldsymbol{\mu}_{2.1} = \hat{\boldsymbol{\mu}}_{2.1}$, $\mathbf{B}_{2.1} = \hat{\mathbf{B}}_{2.1}$, $\sum_{2.1} = \hat{\sum}_{2.1}$, $\boldsymbol{\mu}_{3.21} = \hat{\boldsymbol{\mu}}_{3.21}$, $\mathbf{B}_{3.21} = \hat{\mathbf{B}}_{3.21}$, and $\sum_{3.21} = \hat{\sum}_{3.21}$. Hence, the MLE of $\boldsymbol{\mu}$ and $\sum$ is

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{E}}_2^{(2)} - \hat{\mathbf{B}}_{2.1}\left(\bar{\mathbf{E}}_2^{(1)} - c\right) \\ \bar{\mathbf{E}}_3^{(3)} - \hat{\mathbf{B}}_{3.21}\begin{pmatrix} \bar{\mathbf{E}}_3^{(1)} - c \\ \bar{\mathbf{E}}_3^{(2)} - \hat{\mu}_1 \end{pmatrix} \end{pmatrix}$$

$$\hat{\sum} = \begin{pmatrix} \hat{\sum}_{22} & \hat{\sum}_{23} \\ \hat{\sum}_{32} & \hat{\sum}_{33} \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\sum}_{2.1} + \hat{\mathbf{B}}_{2.1}\hat{\sum}_{12} & \left(\hat{\sum}_{21}, \hat{\sum}_{22}\right)\hat{\mathbf{B}}'_{3.21} \\ \hat{\mathbf{B}}_{3.21}\begin{pmatrix} \hat{\sum}_{12} \\ \hat{\sum}_{22} \end{pmatrix} & \hat{\sum}_{3.21} + \hat{\mathbf{B}}_{2.1}\begin{pmatrix} \hat{\sum}_{31} \\ \hat{\sum}_{32} \end{pmatrix} \end{pmatrix} \tag{9}$$

with

$$\hat{\sum}_{12} = \hat{\mathbf{B}}_{2.1}\hat{\sum}_{11}, \hat{\sum}_{22} = \hat{\sum}_{2.1} + \hat{\mathbf{B}}_{2.1}\hat{\sum}_{12}, \begin{pmatrix} \hat{\sum}_{31} \\ \hat{\sum}_{32} \end{pmatrix}$$

$$= \hat{\sum}_{3.21}\begin{pmatrix} \hat{\sum}_{11} & \hat{\sum}_{12} \\ \hat{\sum}_{21} & \hat{\sum}_{22} \end{pmatrix}.$$

It is obvious that $\hat{\mu}$ is determined completely by $\mathbf{E}_2$ and $\mathbf{E}_3$, which are not related to the last $n_1 - n_2$ observations of the auxiliary variables $\boldsymbol{x}$. However, $\hat{\sum}$ is different. It is not only related to $\mathbf{E}_2$ and $\mathbf{E}_3$, but also related to the extra observations on x in $\mathbf{E}_1$.

References

[1] T. W. Anderson, Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, Journal of the American Statistical Association 52(278) (1957), 200-203.

DOI: https://doi.org/10.2307/2280845

[2] W. G. Cochran, The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce, Journal of Agriculture Science 30(2) (1940), 262-275.

DOI: https://doi.org/10.1017/S0021859600048012

[3] J. Hao and K. Krishnamoorthy, Inferences on a normal covariance matrix and generalized variance with monotone missing data, Journal of Multivariate Analysis 78(1) (2001), 62-82.

DOI: https://doi.org/10.1006/jmva.2000.1939

[4] G. O. Kadilar, A new exponential type estimator for the population mean in simple random sampling, Journal of Modern Applied Statistical Methods 15(2) (2016), 207-214.

DOI: https://doi.org/10.22237/jmasm/1478002380

[5] S. Khan, H. Ali and S. Manzoor Alamgir, A class of transformed efficient ratio estimators of finite population mean, Pakistan Journal of Statistics 31(4) (2015), 353-362.

[6] K. Krishnamoorthy and M. K. Pannala, Some simple test procedures for normal mean vector with incomplete data, Annals of the Institute of Statistical Mathematics 50(3) (1998), 531-542.

DOI: https://doi.org/10.1023/A:1003581513299

[7] K. Krishnamoorthy, Comparison of confidence intervals for correlation coefficients based on incomplete monotone samples and those based on listwise deletion, Journal of Multivariate Analysis 114 (2013), 378-388.

DOI: https://doi.org/10.1016/j.jmva.2012.08.003

[8] K. Krishnamoorthy and M. K. Pannala, Confidence estimation of a normal mean vector with incomplete data, Canadian Journal of Statistics 27(2) (1999), 395-407.

DOI: https://doi.org/10.2307/3315648

[9] K. Krishnamoorthy and J. Yu, Modified Nel and Van der Merwe test for the multivariate Behrens-Fisher problem, Statistics & Probability Letters 66(2) (2004), 161-169.

DOI: https://doi.org/10.1016/j.spl.2003.10.012

[10] K. Krishnamoorthy and J. Yu, Multivariate Behrens-Fisher problem with missing data, Journal of Multivariate Analysis 105(1) (2012), 141-150.

DOI: https://doi.org/10.1016/j.jmva.2011.08.019

[11] R. J. A. Little and D. Rubin, Statistical Analysis with Missing Data, New York: Wiley, 1987.

DOI: https://doi.org/10.1002/9781119013563

[12] R. J. A. Little, A test of missing completely at random for multivariate data with missing values, Journal of American Statistical Association 83(404) (1988), 1198-1202.

DOI: https://doi.org/10.2307/2290157

[13] G. B. Lu and J. B. Copas, Missing at random, likelihood ignorability and model completeness, Annals of Statistics 32(2) (2004), 754-765.

DOI: https://doi.org/10.1214/009053604000000166

[14] M. N. Murthy, Product method of estimation, Sankhyā: The Indian Journal of Statistics, Series A 26(1) (1964), 294-307.

[15] T. Raykov, C. Lee and G. A. Marcoulides, Examining normality in incomplete data sets: A note on a multiple testing approach, Structural Equation Modeling: A Multidisciplinary Journal 23(4) (2016), 595-600.

DOI: https://doi.org/10.1080/10705511.2014.938218

[16] D. S. Robson, Applications of multivariate Polykays to the theory of unbiased ratio-type estimation, Journal of American Statistical Association 52(280) (1957), 511-522.

DOI: https://doi.org/10.2307/2281700

[17] H. P. Singh and R. Tailor, Use of known correlation coefficient in estimating the finite population means, Statistics in Transition 6(4) (2003), 555-560.

[18] R. Singh, P. Chauhan, N. Sawan, and F. Smarandache, Improvement in estimating the population mean using exponential estimator in simple random sampling, Bulletin of Statistics and Economy 3(A09) (2009), 13-18.

[19] D. J. Watson, The estimation of leaf area in field crops, Journal of Agriculture Science 27(3) (1937), 474-483.

DOI: https://doi.org/10.1017/S002185960005173X

[20] A. Yagi, T. Seo and M. S. Srivastava, Testing equality of mean vectors in a one-way MANOVA with monotone missing data, Communications in Statistics-Theory and Methods 47(22) (2018), 5534-5546.

DOI: https://doi.org/10.1080/03610926.2017.1397168

[21] T. Yamada, M. M. Romer and D. St. P. Richards, Kurtosis tests for multivariate normality with monotone, Test 24(3) (2015), 532-557.

DOI: https://doi.org/10.1007/s11749-014-0423-1

[22] Z. Yan and B. Tian, Ratio method to the mean estimation using coefficient of skewness of auxiliary variable, International Conference on Information Computing and Applications: Part II 106 (2010), 103-110.

DOI: https://doi.org/10.1007/978-3-642-16339-5_14

[23] J. Yu, K. Krishnamoorthy and M. K. Pannala, Two-sample inference for normal mean vectors based on monotone missing data, Journal of Multivariate Analysis 97(10) (2006), 2162-2176.

DOI: https://doi.org/10.1016/j.jmva.2006.07.002

[24] J. Yu, K. Krishnamoorthy and Y. He, Testing equality of two normal covariance matrices with monotone missing data, Communications in Statistics-Theory and Methods 49(16) (2019), 3911-3918.

DOI: https://doi.org/10.1080/03610926.2019.1591453

■