STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **6**, June 2018, pp 286–291.
Published online in International Academic Press (www.IAPress.org)

The parameter estimation of the multivariate matrix regression models

Zerong Lin, Lingling He, Tian Wu, Changqing Xu*

School of Mathematics and Physics, Suzhou University of Science and Technology, China

Abstract In this paper, we consider the parameter matrix estimation problem of the multivariate matrix regression models. We approximate the parameter matrix B and the covariance matrix by using the method of the maximum likelihood estimation, together with the Kronecker product of matrices, vectorization of matrices and matrix derivatives.

Keywords Multivariate matrix model, High order derivative, Likelihood estimation, Vectorization, Parameter matrix.

AMS 2010 subject classifications 53A45, 62N02

DOI: 10.19139/soic.v6i2.361

1. Introduction

The Linear model (LM), or called linear regression model, is a basic tool for statistical analysis, among which multivariate linear models (MLM) are widely used in many fields such as agriculture, engineering, pharmaceutical chemical engineering, aerospace, theoretical research and data analysis (see e.g. [1, 4, 5, 7, 8, 12, 16, 20]). The MLM is the case where the number of response factors is greater than 1. Similar to the general LM, in the MLM it is always assumed that the response variable is a linear function of some explanatory variables (vectors or matrices). In a LM, the covariance matrix of the response variables and the parameter matrix B (generally consists of the linear regression coefficients) are unknown and to be estimated by some methods such as maximum likelihood (ML) and ordinary least square (OLS) method in terms of the given data of the response variables and the interpretable variables, and the predict is thus followed after the parameter estimation. The application of the linear model mainly involves the following two aspects.

- Prediction and minimization of the errors. The regression function between the response and the explaining factors can be obtained by observations, and the regression coefficients are obtained by some methods.
- The correlation analysis. This may cause the partitioning or clustering of the observed data, leading to a hierarchical dataset, or to a different model such as the Envelope model[10].

The common approach to estimate the parameters in a linear model is the ordinary least square (OLS)[4, 5], the maximum likelihood (ML) estimation[6, 7, 11], the minimization of error norm (such as minimizing absolute deviation regression analysis[4], and the cost function least squares penalty minimization method[16, 20] (l^2-norm penalty) and Lasso (l^1-norm penalty[7,8]) etc.. Note that the OLS can also be used to estimate parameters of the nonlinear regression model[9]. The general MLM can be indicated by

$$Y = XB + E \tag{1.1}$$

where it satisfies the following assumptions

*Correspondence to: Changqing Xu (Email: cqxurichard@usts.edu.cn). School of Mathematics and Physics, Suzhou University of Science and Technology, Suzhou, Jiangsu Province, China (215009).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1. The rows of the random matrix $Y \in R^{n\times d}$, denoted $Y_{i.}$, are mutually independent.
2. The design matrix $X \in R^{n\times p}$ is fixed and known.
3. The parameter matrix $B \in R^{p\times d}$ is unknown and is to be estimated.
4. The response matrix $Y \in R^{p\times d}$ has a covariance matrix, which is fixed unknown.
5. The mean of the random error is 0.
6. The covariance matrix of the random error ϵ is $\Sigma \otimes I_n$.

The multivariate linear model under the above assumptions is called the Gauss-Markov model. Note that

Definition 1.1
Let $A \in R^{m_1\times n_1}$ and $B \in R^{m_2\times n_2}$. Then matrix $C := A \otimes B \in R^{m_1m_2\times n_1n_2}$, called a direct product(Kronecker product) of the matrices A and B is defined as the blocking matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & a_{13B} & \dots & a_{1n-1}B & a_{1n}B \\ a_{21}B & a_{22}B & a_{23B} & \dots & a_{2n-1}B & a_{2n}B \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ a_{m-11}B & a_{m-1,2}B & a_{m-1,3}B & \dots & a_{m-1,n-1}B & a_{m-1,n}B \\ a_{m1}B & a_{m,2}B & a_{m,3}B & \dots & a_{m,n-1}B & a_{m,n}B \end{bmatrix}$$

Now we present here some basic propositions related to the Kronecker product. The reader is referred to the first chapter of [13] for more detail on Kronecker product.

Proposition 1.2
Let $A_i \in R^{m_i\times n_i}, B_i \in R^{n_i\times p_i}$ for $i = 1, 2$. Then we have

$$(A_1 \otimes A_2)(B_1 \otimes B_2) = (A_1B_1) \otimes (A_2B_2) \tag{1.2}$$

and

$$(A_1 \otimes A_2)' = (A_1)' \otimes (A_2)' \tag{1.3}$$

Proposition 1.3
Let $A \in R^{m\times m}, B \in R^{n\times n}$ be both invertible. Then $A \otimes B$ is invertible and

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \tag{1.4}$$

The vectorization is an operation such that any matrix $A \in R^{m\times n}$ can be made into a column vector $\mathtt{vec}(A) \in R^{mn}$ by vertically stacking in order all the columns of A. Thus the $\mathtt{vec}$ can be regarded as a 1-1 correspondence from the matrix space $R^{m\times n}$ to R^{mn}. From the vectorization, we have

Proposition 1.4
Let $A \in R^{m\times n}, B \in R^{n\times p}, C \in R^{p\times q}$. Then we have

$$\mathtt{vec}(ABC) = (C' \otimes B)\mathtt{vec}(A) \tag{1.5}$$

and if $p = m$, then

$$\mathtt{Tr}(AB) = \mathtt{vec}(B)'\mathtt{vec}(A) \tag{1.6}$$

Matrix vectorization plays an important role in solving the regression model of multivariate linear matrix. We will use this method in the following to figure out the parameter estimation in the model. Firstly, we introduce the derivative of the matrix function. The matrix derivative is one of the key notions in matrix theory for multivariate analysis such as in some extreme problems, maximum likelihood estimation, parameter asymptotic expression of multivariate limit distributions etc.. The real starting point for matrix derivative is by Dwyer MacPhail[9], then further developed by Bargmann[2] and MacRae[14]. A useful tool employed in matrix derivatives is the vectorization of matrices, see Neudecker[17] and Tracy Dwyer[18], McDonald Swaminathan [15] and Bentler Lee

[3]. The notion of a matrix derivative is a realization of the Fréchet derivative known from functional analysis.

We first recall the definition of the derivative of a vector $\mathbf{y} = (y_1, \ldots, y_n)'$ w.r.t. another vector $\mathbf{x} = (x_{1,\ldots,x_m})$. Then $\frac{d\mathbf{y}}{d\mathbf{x}} = (J_{ij}) \in R^{m\times n}$ where $J_{ij} = \frac{dy_j}{dx_i}$ for $i = 1, \ldots, m; j = 1, \ldots, n$. Now consider matrix $Y = (y_{ij}) \in R^{m\times n}$ each of whose entries y_{ij} is a differentiable function of $X = (x_{st}) \in R^{p\times q}$, i.e., y_{ij} can be regarded as a function with pq arguments x_{st}. Then we define

$$\frac{dY}{dX} = \frac{d(\mathtt{vec}(Y)')}{d\mathtt{vec}(X)} \in R^{pq\times mn}$$

The second order derivative, $\frac{d^2Y}{dX^2}$, is defined by

$$\frac{d^2Y}{dX^2} := \frac{d}{dX}(\frac{dY}{dX}) = \frac{d}{d\mathtt{vec}(X)}\mathtt{vec}(\frac{dY}{dX})' \tag{1.7}$$

Thus we have $\frac{d^2Y}{dX^2} \in R^{pq\times mnpq}$. We can also define any order derivative by using the induction on the order. Actually we already defined the 1st and the 2nd order derivative of Y w.r.t. X. Now suppose we have defined the $(k-1)$th order derivative, i.e.,

$$\frac{d^{k-1}Y}{dX^{k-1}} = A^{(k-1)} \in R^{pq\times mn(pq)^{k-2}}$$

Then the kth order derivative of Y w.r.t. X is defined by

$$\frac{d^kY}{dX^k} = \frac{d}{d\mathtt{vec}(X)}(\frac{d^{k-1}Y}{dX^{k-1}}) \in R^{pq\times mn(pq)^{k-1}}$$

We have

Proposition 1.5
Let $X = (x_{ij}) \in R^{m\times n}$ and the elements of X are all independent variables, and A be a matrix of proper size with constant elements, and c is a constant. Then we have

(1) $\frac{dX}{dX} = I_{mn}$, where I_k represents the $k \times k$ identity matrix.

(2) $\frac{d(cX)}{dX} = cI_{mn}$ and $\frac{d(AX)}{dX} = I_n \otimes A'$.

(3) $\frac{d(AXB)}{dX} = B \otimes A'$.

The following results concerns the derivatives of the inverse, determinant and the trace of a random matrix.

Proposition 1.6
Let $X = (x_{ij}) \in R^{n\times n}$ be invertible and A be a matrix of proper size. Then

(1) $\frac{dX^{-1}}{dX} = -X^{-1} \otimes (X')^{-1}$.

(2) $\frac{d(\det(X)}{dX} = \det(X)\mathtt{vec}((X^{-1})')$.

(3) $\frac{d(\mathtt{Tr}(Y)}{dX} = \frac{dY}{dX}\mathtt{vec}(I)$.

(4) $\frac{d(\mathtt{Tr}(AXBX')}{dX} = \mathtt{vec}(A'XB') + \mathtt{vec}(AXB)$.

2. Maximum Likelihood Estimation of B in (1.1)

In this section, we use the maximum likelihood (ML) function and combine the results we obtained in the last section to estimate B in (1.1). Let the random matrix Y satisfying model (1.1) obeys the normal distribution with parameter matrix B and the covariance matrix Σ. Then the corresponding distribution density function is

$$L(B,\Sigma,Y) := (2\pi)^{-nd/2}(\det(\Sigma))^{-n/2}exp\left\{-\frac{1}{2}\mathtt{Tr}((Y-XB)\Sigma^{-1}(Y-XB)')\right\} \tag{2.8}$$

Theorem 2.1
Suppose $r := \mathtt{rank}(X) = p \le n$ in (1.1). Then the maximum likelihood estimation of B is

$$\hat{B} = (X'X)^{-1}X'Y \tag{2.9}$$

Proof
We regard $L(B,\Sigma,Y)$ as a function of B. To get the maximum likelihood of B, we compute the derivative on the logarithm of L w.r.t. B , since

$$\log(L) = -\frac{1}{2}nd\log(2\pi) - \frac{1}{2}n\log(\det\Sigma) - \frac{1}{2}S$$

We have

$$\frac{\partial\log(L)}{\partial B} = \frac{1}{2}\frac{\partial S}{\partial B}$$

Note that

$$\begin{aligned}\frac{\partial S}{\partial B} &= \frac{\partial(Y-XB)}{\partial B}\frac{\partial\mathtt{Tr}((Y-XB)\Sigma^{-1}(Y-XB)')}{\partial(Y-XB)}\\ &= -\frac{\partial(XB)}{\partial B}\mathtt{vec}[(Y-XB)\Sigma^{-1}]\\ &= -2(I\otimes X')\mathtt{vec}[(Y-XB)\Sigma^{-1}] = -2[X'(Y-XB)\Sigma^{-1}]\\ &= -2(X'Y-X'XB)\Sigma^{-1}\end{aligned}$$

Thus $\frac{\partial\log(L)}{\partial B} = 0$ is equivalent to $(X'Y - X'XB)\Sigma^{-1} = 0$. It follows that $(X'X)B = X'Y$. Consequently we get (2.9) under the hypothesis of $r := \mathtt{rank}(X) = p \le n$. □

In order to estimate the parameter matrix B in (1.1) for the case when X is not full rank, i.e., $\mathtt{rank}(X) < p$, we need to introduce a class of generalized inverse–the group inverse or g-inverse, which is also called a $\{1\}$ −inverse. There are a lot of literatures on generalized inverses of matrices. We refer the reader to the first chapter in [13] for reference.

Given a matrix $A \in \mathcal{C}^{m\times n}$, the g-inverse of A, denoted A^-, is an $n \times m$ matrix satisfying condition

$$AA^-A = A \tag{2.10}$$

An equivalent definition for the g-inverse is:

Proposition 2.2
Let $A \in \mathcal{C}^{m\times n}, B \in \mathcal{C}^{n\times m}$. Then B is an g-inverse of A if and only if for any vector $b \in \mathtt{Col}(A)$, $x = Bb$ is a solution to the linear system $Ax = b$, where $\mathtt{Col}(A) := \{y = Ax : x \in \mathcal{C}^n\}$ denotes the range space of A.

Note that the g-inverse of a matrix is usually non-unique. Another useful fact we will utilize in the proof of the next result is that when matrix A is a square nonsingular matrix, the g-inverse is exactly the inverse matrix (and therefore it is unique). The following proposition presents a general form of g-inverses of a given matrix A after given a specific g-inverse.

Proposition 2.3
Let $A \in \mathcal{C}^{m\times n}$ and A_0^- be a given g-inverse of A. Then any g-inverse of A is in form

$$A^- = A_0^- + Z - A_0^- AZAA_0^- \tag{2.11}$$

where $Z \in \mathcal{C}^{n\times m}$ is arbitrary.

We now generalize the result in (2.1) :

Theorem 2.4
Let $Y \in R^{n\times d}, X \in R^{n\times p}, B \in R^{p\times d}$ in (1.1). Then the maximum likelihood estimation of B is

$$\hat{B} = (X'X)^- X'Y \tag{2.12}$$

Proof
For $r = \mathtt{rank}(X) = p$, the matrix $X'X$ is invertible. In this case, we have $(X'X)^- = (X'X)^{-1}$. By Theorem (2.1), we get the result. Now we consider the case $r = \mathtt{rank}(X) < p$. By the similar argument as in the proof of Theorem (2.1), we obtain $(X'X)B = X'Y$. This is equivalent to $(I_p \otimes X'X)\mathtt{vec}(B) = \mathtt{vec}(X'Y)$ by Proposition 1.2, where I_p is the $p \times p$ identity matrix. Thus we have by Proposition 1.3, 1.4 and the Proposition 2.2 that

$$\begin{aligned} \mathtt{vec}(B) &= (I \otimes X'X)^- \mathtt{vec}(X'Y) \\ &= [I_p \otimes (X'X)^-]\mathtt{vec}(X'Y) \end{aligned}$$

The result (2.12) follows by using again Proposition 1.3. □

We now investigate the covariance matrix of the random matrix B in model (1.1). The covariance matrix of a random vector $\mathbf{x} \in R^n$ is a symmetric positive (semi-)definite matrix $D[\mathbf{x}] = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'] \in R^{n\times n}$. For a random matrix $X = (X_{ij}) \in \mathcal{C}^{m\times n}$ (i.e., each element of X is a random variable), we define its covariance matrix $D[X]$ as the covariance matrix of $\mathtt{vec}(X)$, that is, $D[X] = D[\mathtt{vec}(X)] \in \mathcal{C}^{mn\times mn}$. It follows that

Lemma 2.5
Let $X = (X_{ij}) \in R^{m\times n}$ be a random matrix. If all rows $X^{(i)}$'s are uniformly distributed with $\mathtt{Cov}(X^{(i)}) = \Sigma$ for all $i \in [m]$, and all columns X_j's are uniformly distributed with $\mathtt{Cov}(X_j) = \Phi$ for all $j \in [n]$. Then $D[X] = \Sigma \otimes \Phi$.

Given $\mu \in R^{m\times n}, \Sigma \in R^{n\times n}, \Phi \in R^{m\times m}$ where Σ, Φ are both positive definite matrices. We say X obeys a matrix normal distribution with parameter matrices μ, Σ, Φ , or denoted $X \sim Normal_{m,n}\ (\mu, \Sigma, \Phi)$. Note that $X \sim Normal_{m,n}\ (\mu, \Sigma, \Phi)$is equivalent to $\mathtt{vec}(X) \sim Normal_{mn}\ (\mathtt{vec}(\mu), \Sigma \otimes \Phi)$.

Theorem 2.6
Let $Y \in R^{n\times d}, X \in R^{n\times p}, B \in R^{p\times d}$ in (1.1) satisfying condition (1-6), and let $r = \mathtt{rank}(X) = p$. Then the covariance matrix Ω is

$$\Omega = \Sigma \otimes (X'X)^{-1} \tag{2.13}$$

Proof
For $r = \mathtt{rank}(X) = p$, the matrix $X'X$ is invertible. In this case, we have

$$\begin{aligned} \hat{(B)} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(XB + E) \\ &= B + ME \end{aligned}$$

where $M = (X'X)^{-1}X'$. Therefore we have

$$\begin{aligned} \Omega &= \mathtt{Cov}(\mathtt{vec}(\hat{(B)})) = \mathtt{Cov}(\mathtt{vec}(B) + \mathtt{vec}(ME)) \\ &= \mathtt{Cov}(\mathtt{vec}(ME)) = \mathtt{Cov}[(I_d \otimes M)\mathtt{vec}(E)] \\ &= (I_d \otimes M)\mathtt{vec}(E)(I_d \otimes M)' \\ &= (I_d \otimes M)(\Sigma \otimes I_n)(I_d \otimes M)' \\ &= \Sigma \otimes MM' = \Sigma \otimes (X'X)^{-1} \end{aligned}$$

Thus we get (2.13). The proof is completed. □

We end the paper by presenting without proof a result on the estimation of the covariance matrix Ω based upon Theorem 2.13.

Corollary 2.7
Let $Y \in R^{n\times d}, X \in R^{n\times p}, B \in R^{p\times d}$ in (1.1) satisfying condition (1-6), and let $r = \mathtt{rank}(X) = p$. Then the covariance matrix is $\Omega = \Sigma \otimes (X'X)^{-1}$ where Σ can be estimated by

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})'(Y - X\hat{B})$$

3. Conclusion

Based on the generalised inverse and the properties of the matrix inverse, we get the covariance matrix for the error distribution of the matrix regression model (1.1). We also present the covariance matrix for the model (1.1) when all the rows (columns) of the design matrix X are uniformly distributed.

Acknowledgement

This work was supported by the Graduate student Research Foundation of USTS and partially supported by Hong Kong Research fund(No. PolyU 502111, 501212).

REFERENCES

1. S.G. Baker, *Regression analysis of grouped survival data with incomplete covariates: nonignorable missing-data and censoring mechanisms*, Biometrics, vol. 50, pp. 821–826, 1994.
2. R. E. Bargmann, *Matrices and determinants. In: CRC Handbook of Tables for Mathematics*, Ed. S.M. Selby. Chemical Rubber Co, Cleveland, pp. 146–148, 1964.
3. P. M. Bentler, and S. Y. Lee, *Some extensions of matrix calculus*. General Systems vol. 20, pp. 145–150,1975.
4. M. Bilodeau, and D. Brenner, *Theory of Multivariate Statistics*, Springer., New York, 1961.
5. D.B. Cox, *Regression models and life-tables*, J. Roy. Statist. Soc. B, vol. 34, pp. 187–220,1972.
6. A.P. Dempster, N.M. Laird, D.B. Rubin *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. B, vol. 39, pp. 1–38,1977.
7. K. A. Bollen, and P. J. Curran, *Latent curve models: A structure equation perspective*, Wiley, NJ, 2006.
8. A. S. Bryk, and S. W. Raudenbush, *Application of hierarchical linear models to assessing change*, Psychological Bulletin., vol. 101, pp. 147–158,1987.
9. P. S. Dwyer, and M. S. Macphail, *Symbolic Matrix Derivatives*, Ann. Math. Statist. vol.19, pp. 517–534,1948.
10. R. D. Cook, and X. Zhang, *Simultaneous envelopes for multivariate linear regression*, Technometrics, vol. 57, pp. 11–25, 2015.
11. T. Hastie, R. Tibshirani, and J. Friedman, *The elements of Statistical Learning*, 2nd ed., Springe, New York, 2009.
12. J. D. Kalbfleisch and R. L. Prentice, The Statistical Analysis of Failure Time Data. Wiley, New York,1980.
13. T. Kollo, and D. Rosen, *Advanced Multivariate Statistics with Matrices*, Springer., New York, 2005.
14. E. C. MacRae, *Matrix derivatives with an application to an adaptive linear decision problem*, Ann. Statist. vol. 2, pp. 337–346,1974.
15. R. P. McDonald, and H. Swaminathan, *A simple matrix calculus with applications to multivariate analysis*, General Systems, vol.18, pp. 37–54,1973.
16. K. E. Muller, and P. W. Stewart, *Linear Model Theory: Univariate, Multivariate, and Mixed Models*, Wiley Blackwell, 2012.
17. H. Neudecker, *Some theorems on matrix differentiations with special reference to Kronecker matrix products*, J. Amer. Statist. Assoc, vol.64, pp. 953–963,1969.
18. D. S. Tracy, and P. S. Dwyer, *Multivariate maxima and minima with matrix derivates*, J. Amer. Statist. Assoc., vol. 64, pp. 1576–1594, 1969.
19. S. W. Raudenbush, and A. S. Bryk, *Hierarchical linear models: Applications and data analysis methods 2*, Sage Publications, Thousand Oaks, CA, 2002.
20. J. D. Singer, *Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth models*, Journal of Educational and Behavioral Statistics, vol. 23, pp. 323–355,1989.