Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 6942865, 8 pages
https://doi.org/10.1155/2017/6942865

Research Article

On the Degrees of Freedom of Mixed Matrix Regression

Pan Shang and Lingchen Kong

Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Lingchen Kong; konglchen@126.com

Received 3 April 2017; Revised 13 July 2017; Accepted 7 August 2017; Published 18 September 2017

Academic Editor: Weihai Zhang

With the increasing prominence of big data in modern science, data of interest are more complex and stochastic. To deal with the complex matrix and vector data, this paper focuses on the mixed matrix regression model. We mainly establish the degrees of freedom of the underlying stochastic model, which is one of the important topics to construct adaptive selection criteria for efficiently selecting the optimal model fit. Under some mild conditions, we prove that the degrees of freedom of mixed matrix regression model are the sum of the degrees of freedom of Lasso and regularized matrix regression. Moreover, we establish the degrees of freedom of nuclear-norm regularization multivariate regression. Furthermore, we prove that the estimates of the degrees of freedom of the underlying models process the consistent property.

1. Introduction

With the increasing prominence of large-scale data in modern science, data of interest are more complex, which may be in the form of a matrix, not a vector. At the same time, the random noises are not always normal. These complex stochastic data are frequently collected in a large variety of research areas such as information technology, engineering, medical imaging and diagnosis, and finance [1–7]. For instance, a well-known example is the study of an electroencephalography data set of alcoholism. The study consists of 122 subjects with two groups, an alcoholic group and a normal control group, and each subject was exposed to a stimulus. Voltage values were measured from 64 channels of electrodes placed on the subject's scalp for 256 time points, so each sampling unit is a 256×64 matrix. To address scientific questions arising from those data, sparsity or other forms of regularization are crucial owing to the ultrahigh dimensionality and complex structure of the matrix data. Often, a variety of models in statistics lead to the estimation of matrices with rank constraints. The true signal often has low rank, which can be well approximated by a low rank matrix. Recently, Zhou and Li [5] proposed the so-called regularized matrix regression model to deal with these matrix form data, which is based on spectral regularization. This model includes the well-known Lasso as a special case; see [8] for more details. Moreover, one of the main results in [5] claimed the degrees of freedom of the proposed model under orthonormal assumption.

Degrees of freedom of the underlying stochastic model are one of the important topics. As we know, if we want to evaluate the performance of a model when we use it to analyze data, we need to choose the optimal tuning parameter in the same model. Many methods have been proposed to solve this problem. The popular methods include C_p, AIC, and BIC [9–11]. There is also a computational cost method named cross-validation. Efron [11] showed that C_p is an unbiased estimate of prediction error, and in most cases C_p provides an accurate parameter over cross-validation. Thus, C_p and AIC outperform the cross-validation. The fundamental idea of C_p, AIC, and BIC is connected with the concept of degrees of freedom.

Degrees of freedom can be easily understood in linear model. In linear case, the degrees of freedom are the number of prediction variables. However, if there exist constraints on the prediction variables, the degrees of freedom do not exactly correspond to the number of variables; see, for example, [5, 12–18]. After Stein [12] got Stein's unbiased estimation, analytical forms of the degrees of freedom of different models have been studied for vector case. For instance, Hastie and Tibshirani [13] showed that the degrees of freedom of a linear smoother equal the trace of the prediction matrix. In general, it is difficult to get the degrees of freedom of

many models. In 1998 Ye [15] and in 2002 Shen and Ye [16] used the computational method to predict the degrees of freedom. However, there is a deficient thing that the more data, the more cost of computation. For high-dimension vector case, Zou et al. [14] gave the degrees of freedom of Lasso. Furthermore, Tibshirani and Taylor [17, 18] gave the degrees of freedom of generalized Lasso.

However, for matrix case, there are a few results about the degrees of freedom of matrix regression. One can see that getting the analytical form of the degrees of freedom of our model is very essential both in theory and in practice. Thus, it is important to study the degrees of freedom in matrix case in the big data era. Notice that, besides Zhou and Li's work [5] about the degrees of freedom of regularized matrix regression, Yuan [19] got the degrees of freedom in low rank matrix estimation, which includes the cases of the rank constraints and nuclear-norm regularization. Note that Yuan [19] just considered the rank constraints of multivariate regression, and Zhou and Li [5] did not consider the mixed case, which is combined with matrix and vector. If we use the nuclear norm as the penalty, what are the degrees of freedom of that model? If the variables are mixed, what are the degrees of freedom of that model?

We will answer the above questions affirmatively in our paper. Firstly, we prove that the degrees of freedom of mixed matrix regression model are the sum of the degrees of freedom of Lasso and regularized matrix regression; this result can be useful to construct adaptive selection criteria for efficiently selecting the optimal model fit. Then, following the same idea we establish the degrees of freedom of nuclear-norm regularization multivariate regression. It is worth noticing that Zou et al. [14] not only gave the unbiased estimate of the degrees of freedom of Lasso model, but also proved the following consistency of the estimate. This is an interesting and important work on the estimates of the degrees of freedom of Lasso. Based on their work, we finally prove that the estimates of the degrees of freedom given in this paper are consistent.

Our paper is organized as follows. In Section 2, we introduce the primary model, basic concepts, and notations used in our paper. In Section 3, we show the process of computing the degrees of freedom of model (3). In Section 4, we give the degrees of freedom of multivariate regression with nuclear-norm regularization. In Section 5, we verify the consistent property of the estimates. We conclude the paper with a discussion of potential future research in Section 6.

2. Preliminaries

In this section, we mainly introduce our model and basic concepts. First we present mixed matrix regression model. Then for convenient discussion and understanding of our work, we give some basic knowledge and notations.

Suppose $y \in \mathfrak{R}$ is the response variable, $\gamma \in \mathfrak{R}^{p_0}$ is the prediction vector, and $X \in \mathfrak{R}^{p_1 \times p_2}$ is the prediction matrix. They are known. Let $\mathbf{z}$ and B be unknown prediction vector and matrix. The statistical model of matrix regression is given as

$$y = \langle B, X \rangle + \gamma^T \mathbf{z} + \epsilon, \tag{1}$$

where $\langle B, X \rangle$ is the sum of multiply of corresponding element of B and X; ϵ is the prediction error of the model. Suppose we take n samples

$$y_i = \langle B, X_i \rangle + \gamma_i^T \mathbf{z} + \epsilon_i \quad i = 1, \ldots, n. \tag{2}$$

Note that, in the real data case, there are always some special structures of B and $\mathbf{z}$ such that B has low rank and $\mathbf{z}$ is usually sparse. In this case, we define mixed matrix regression model as

$$\min_{(B,\mathbf{z})} \quad \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \gamma_i^T \mathbf{z} - \langle B, X_i \rangle \right)^2 + \lambda_1 \|B\|_* + \lambda_2 \|\mathbf{z}\|_1, \tag{3}$$

where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ are the regularized parameters and $\|B\|_*$ is the nuclear norm of B which is the sum of singular values of B. That is, if B has singular decomposition, $U \operatorname{diag}(\mathbf{b}) V^T$, where $U \in \mathfrak{R}^{p_1 \times p_1}$ and $V \in \mathfrak{R}^{p_2 \times p_2}$, are normal orthogonal matrix, and $\operatorname{diag}(\mathbf{b})$ is a matrix whose elements of main diagonal are singular values of B with $\mathbf{b} = (b_1, b_2, \ldots, b_p, 0, \ldots, 0)$, then $\|B\|_* = |b_1| + |b_2| + \cdots + |b_p|$. $\|\mathbf{z}\|_1$ is defined as the sum of absolute values of every component of $\mathbf{z}$. That is, if $\mathbf{z} = (z_1, z_2, \ldots, z_{p_0})$, $\|\mathbf{z}\|_1 = |z_1| + |z_2| + \cdots + |z_{p_0}|$. Clearly, if $B = 0$ in model (3), we will get Lasso model. For Lasso model, the research is very mature including algorithm and the degrees of freedom. In statistical parlance, Lasso uses an ℓ_1 penalty which has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. We say that Lasso yields sparse models that just involve a subset of the variables, performing variable selection. Lasso has been widely used in statistical and machine learning. In model (3), if $\mathbf{z} = \mathbf{0}$, we will get regularized matrix regression model mainly studied in [5].

Now we review some basic results on the degrees of freedom. Based on Stein's unbiased estimation, Efron et al. [20] showed that the effective degrees of freedom of any fitting procedure δ has a rigorous definition under the differentiability condition on the estimate $\hat{\mathbf{y}}$ of $\mathbf{y}$ based on δ, where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ denotes the response vector. That is, given a method δ, let $\hat{\mathbf{y}} = \delta(\mathbf{y})$ denote its fit. Then under the differentiability of $\hat{\mathbf{y}}$, the degrees of freedom of δ are given by

$$\operatorname{df}(\hat{\mathbf{y}}) = \operatorname{tr}\left\{ D\hat{\mathbf{y}}(\mathbf{y}) \right\}. \tag{4}$$

This means that the degrees of the freedom of δ are the trace of the Jacobian matrix which is a special case of Definition 1. Once we get the degrees of freedom, we can establish three well-known information criteria C_p, AIC, and BIC under the normal noise case. That is,

$$\begin{aligned} C_p &= \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{n\sigma^2} + \frac{2\operatorname{df}}{n}, \\ \mathrm{AIC} &= \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{\sigma^2} + 2\operatorname{df}, \\ \mathrm{BIC} &= \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{\sigma^2} + \ln(n)\operatorname{df}. \end{aligned} \tag{5}$$

In order to deal with the degrees of freedom of mixed matrix regression model, we define an operator to simplify the expression of optimal question and the Jacobian matrix of matrix function.

Definition 1. Suppose there is a matrix function f:

$$f : \mathfrak{R}^{p\times q} \longmapsto \mathfrak{R}^{s\times t}, \quad N \longrightarrow M = f(N). \tag{6}$$

Then one defines the Jacobian matrix as $DM(N) = \partial \operatorname{vec}(M)/\partial \operatorname{vec}^T(N)$.

Suppose $M = \begin{pmatrix} m_{11},\ldots,m_{1t} \\ \vdots \\ m_{s1},\ldots,m_{st} \end{pmatrix}$, $N = \begin{pmatrix} n_{11},\ldots,n_{1q} \\ \vdots \\ n_{p1},\ldots,n_{pq} \end{pmatrix}$. We vectorize the matrix into a vector by column. For example, $\operatorname{vec}(M) = (m_{11},\ldots,m_{s1},\ldots,m_{1t},\ldots,m_{st})^T$. Then the Jacobian matrix of f can be written as

$$DM(N) = \begin{pmatrix} \frac{\partial m_{11}}{\partial n_{11}}, \frac{\partial m_{11}}{\partial n_{21}}, \ldots, \frac{\partial m_{11}}{\partial n_{p1}}, \ldots, \frac{\partial m_{11}}{\partial n_{1q}}, \frac{\partial m_{11}}{\partial n_{2q}}, \ldots, \frac{\partial m_{11}}{\partial n_{pq}} \\ \frac{\partial m_{21}}{\partial n_{11}}, \frac{\partial m_{21}}{\partial n_{21}}, \ldots, \frac{\partial m_{21}}{\partial n_{p1}}, \ldots, \frac{\partial m_{21}}{\partial n_{1q}}, \frac{\partial m_{21}}{\partial n_{2q}}, \ldots, \frac{\partial m_{21}}{\partial n_{pq}} \\ \vdots \\ \frac{\partial m_{st}}{\partial n_{11}}, \frac{\partial m_{st}}{\partial n_{21}}, \ldots, \frac{\partial m_{st}}{\partial n_{p1}}, \ldots, \frac{\partial m_{st}}{\partial n_{1q}}, \frac{\partial m_{st}}{\partial n_{2q}}, \ldots, \frac{\partial m_{st}}{\partial n_{pq}} \end{pmatrix}. \tag{7}$$

Definition 2. Let the operator $\star$ be defined from $\mathfrak{R}^{m\times mn} \times \mathfrak{R}^{m\times n}$ to $\mathfrak{R}^m$ by

$$X \star Y = X\operatorname{vec}(Y). \tag{8}$$

It is easy to verify that the operator $\star$ is linear and $A(X \star Y) = (AX) \star Y$.

Let $\chi = \begin{pmatrix} \operatorname{vec}^T(X_1) \\ \vdots \\ \operatorname{vec}^T(X_n) \end{pmatrix}$, $\gamma = \begin{pmatrix} \gamma_1^T \\ \vdots \\ \gamma_n^T \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$. Then, we can rewrite mixed matrix regression model (3) as

$$\min_{(B,\mathbf{z})} \quad \frac{1}{2}\left\|\mathbf{y} - \chi \star B - \gamma \star \mathbf{z}\right\|_2^2 + \lambda_1 \|B\|_* + \lambda_2 \|\mathbf{z}\|_1. \tag{9}$$

Let $\mathscr{B} = (B, \mathbf{z})$ denote the unknown coefficients, and let $\mathscr{A} = (\chi, \gamma)$ denote the prediction matrix. Our paper is based on the assumptions that $\mathscr{A}^T\mathscr{A}$ is full rank and the matrix data and vector data are independent, that is, $\chi^T\gamma = 0$.

3. The Unbiased Estimate of the Degrees of Freedom

We begin with the least squares estimate of our mixed matrix regression, which is the optimal solution of the problem

$$\min \quad \frac{1}{2}\left\|\mathbf{y} - \chi \star B - \gamma \star \mathbf{z}\right\|_2^2. \tag{10}$$

By taking the partial deviation of the minimal question, we can know that

$$\begin{aligned} -\chi^T\left(\mathbf{y} - \chi \star \widehat{B} - \gamma \star \widehat{\mathbf{z}}\right) = 0, \\ -\gamma^T\left(\mathbf{y} - \chi \star \widehat{B} - \gamma \star \widehat{\mathbf{z}}\right) = 0. \end{aligned} \tag{11}$$

From our definitions in Section 2, we can easily verify that $D\chi \star B(B) = \chi$, $D\gamma \star \mathbf{z}(\mathbf{z}) = \gamma$. From the relationship $\widehat{\mathbf{y}} = \chi \star \widehat{B} + \gamma \star \widehat{\mathbf{z}}$, we obtain that

$$D\widehat{\mathbf{y}}\left(\widehat{\mathscr{B}}\right) = \left(D\widehat{\mathbf{y}}\left(\widehat{B}\right), D\widehat{\mathbf{y}}\left(\widehat{\mathbf{z}}\right)\right) = (\chi, \gamma). \tag{12}$$

By taking the partial deviation of the implicit functions on $\mathbf{y}$ above, we get

$$\begin{aligned} -\chi^T + \chi^T\chi D\widehat{B}_{LS}(\mathbf{y}) + \chi^T\gamma D\widehat{\mathbf{z}}_{LS}(\mathbf{y}) = 0, \\ -\gamma^T + \gamma^T\chi D\widehat{B}_{LS}(\mathbf{y}) + \gamma^T\gamma D\widehat{\mathbf{z}}_{LS}(\mathbf{y}) = 0. \end{aligned} \tag{13}$$

Thus, we derive $\mathscr{A}^T\mathscr{A}D\widehat{\mathscr{B}}_{LS}(\mathbf{y}) = \mathscr{A}^T$. If $\mathscr{A}^T\mathscr{A}$ is a full rank matrix, we can get $D\widehat{\mathscr{B}}_{LS}(\mathbf{y}) = (\mathscr{A}^T\mathscr{A})^{-1}\mathscr{A}^T$.

Based on the definition of the degrees of freedom, we know that if the estimation $\widehat{\mathbf{y}}$ is differentiable on $\mathbf{y}$, $\widehat{\mathrm{df}} = \operatorname{tr}\{D\widehat{\mathbf{y}}(\mathbf{y})\}$ is an unbiased estimate of the degrees of freedom. Combining with the chain rule and the Jacobian matrix of fitted value with respect to responses, we can get

$$\begin{aligned} \widehat{\mathrm{df}} &= \operatorname{tr}\left\{D\widehat{\mathbf{y}}(\mathbf{y})\right\} \\ &= \operatorname{tr}\left\{D\widehat{\mathbf{y}}\left(\widehat{\mathscr{B}}\right)D\widehat{\mathscr{B}}\left(\widehat{\mathscr{B}}_{LS}\right)D\widehat{\mathscr{B}}_{LS}(\mathbf{y})\right\}. \end{aligned} \tag{14}$$

This together with the above arguments, we can get

$$\begin{aligned} \widehat{\mathrm{df}} &= \operatorname{tr}\left\{\mathscr{A}D\widehat{\mathscr{B}}\left(\widehat{\mathscr{B}}_{LS}\right)\left(\mathscr{A}^T\mathscr{A}\right)^{-1}\mathscr{A}^T\right\} \\ &= \operatorname{tr}\left\{D\widehat{\mathscr{B}}\left(\widehat{\mathscr{B}}_{LS}\right)\left(\mathscr{A}^T\mathscr{A}\right)^{-1}\mathscr{A}^T\mathscr{A}\right\} \\ &= \operatorname{tr}\left\{D\widehat{\mathscr{B}}\left(\widehat{\mathscr{B}}_{LS}\right)\right\}. \end{aligned} \tag{15}$$

Because $D\widehat{\mathscr{B}}(\widehat{\mathscr{B}}_{LS}) = \begin{pmatrix} D\widehat{B}(\widehat{B}_{LS}), D\widehat{B}(\widehat{\mathbf{z}}_{LS}) \\ D\widehat{\mathbf{z}}(\widehat{B}_{LS}), D\widehat{\mathbf{z}}(\widehat{\mathbf{z}}_{LS}) \end{pmatrix}$, it is easy to yield

$$\begin{aligned} \widehat{\mathrm{df}} &= \operatorname{tr}\left\{D\widehat{\mathscr{B}}\left(\widehat{\mathscr{B}}_{LS}\right)\right\} \\ &= \operatorname{tr}\left\{D\widehat{B}\left(\widehat{B}_{LS}\right)\right\} + \operatorname{tr}\left\{D\widehat{\mathbf{z}}\left(\widehat{\mathbf{z}}_{LS}\right)\right\}. \end{aligned} \tag{16}$$

We are ready to present our main result in this section.

Theorem 3. *Let $\widehat{B}_{LS}$ be the usual least squares estimate of B and assume that it has distinct positive singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_p > 0$; then the unbiased estimate of the degrees of freedom of model (9) is*

$$\widehat{df} = \|\widehat{\mathbf{z}}\|_0 + \sum_{i=1}^{p}\mathbf{1}_{\{\sigma_i>\lambda_1\}}\left\{1 + \sum_{j=1,j\neq i}^{p_1}\frac{\sigma_i(\sigma_i-\lambda_1)}{\sigma_i^2-\sigma_j^2} + \sum_{j=1,j\neq i}^{p_2}\frac{\sigma_i(\sigma_i-\lambda_1)}{\sigma_i^2-\sigma_j^2}\right\}, \tag{17}$$

where $\widehat{\mathbf{z}}$ is the estimate of $\mathbf{z}$ and $\|\widehat{\mathbf{z}}\|_0$ is the number of nonzero elements in $\widehat{\mathbf{z}}$. Clearly, $df = E(\widehat{df})$ is the degrees of freedom of mixed matrix regression.

Theorem 3 is an immediate result of the following two propositions whose proofs are relegated to the Appendix for the sake of presentation.

Proposition 4. *For any $\lambda_1 \geq 0$, the unbiased estimate of the degrees of freedom of regularized matrix regression model equals* $\mathrm{tr}\{D\widehat{B}(\widehat{B}_{LS})\}$ *given by*

$$\mathrm{tr}\left\{D\widehat{B}\left(\widehat{B}_{LS}\right)\right\} = \sum_{i=1}^{p}\mathbf{1}_{\{\sigma_i>\lambda_1\}}\left\{1+\sum_{j=1,j\neq i}^{p_1}\frac{\sigma_i\left(\sigma_i-\lambda_1\right)}{\sigma_i^2-\sigma_j^2} + \sum_{j=1,j\neq i}^{p_2}\frac{\sigma_i\left(\sigma_i-\lambda_1\right)}{\sigma_i^2-\sigma_j^2}\right\}, \tag{18}$$

where $\widehat{B}_{LS}$ is the usual least squares estimate of B and assume that it has distinct positive singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_p > 0$.

Proposition 5. *$\forall\lambda_2 \geq 0$, the unbiased estimate of the degrees of freedom of Lasso equals* $\mathrm{tr}\{D\widehat{\mathbf{z}}(\widehat{\mathbf{z}}_{LS})\}$ *given by*

$$\mathrm{tr}\left\{D\widehat{\mathbf{z}}\left(\widehat{\mathbf{z}}_{LS}\right)\right\} = \|\mathbf{z}\|_0. \tag{19}$$

4. Multivariate Regression with Nuclear-Norm Regularization

This section considers the multivariate regression, which has the following statistical model

$$Y = XB + E, \tag{20}$$

where $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^T$ is an $n \times q$ response matrix, $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ is an $n \times p$ prediction matrix, B is a $p \times q$ unknown coefficient matrix, and the regression random noise $E \sim N(0, \tau^2 I_n \otimes I_q)$.

Very recently, Yuan [19] studied the degrees of freedom of multivariate regression with low rank constraint via the following optimal model:

$$\min_{B,\mathrm{rank}(B)\leq k} \|Y - XB\|_F^2. \tag{21}$$

Since the above optimal model with the low rank constraint is difficult to compute, it is NP-hard problem. In this case, we usually relax the rank constraint to nuclear-norm regularization. Then we get the nuclear-norm regularization multivariate regression model

$$\min_{B} \|Y - XB\|_F^2 + \lambda\|B\|_*. \tag{22}$$

Following the same technique as in the proof of Theorem 3, we can easily obtain the degrees of freedom of the nuclear-norm regularization multivariate regression. We omit its proof for brevity.

Theorem 6. *Assume that* $\mathrm{rank}(X^TX) = p$ *in (22). Let $\widehat{B}_{LS}$ be the usual least squares estimate and assume that it has distinct positive singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_p > 0$, where $p = \min\{p_1, p_2\}$. With the convention $\sigma_i = 0$ for $i > p$, the following expression is an unbiased estimate of the degrees of freedom of the regularized fit (22):*

$$\widehat{df}(\lambda) = \sum_{i=1}^{p}\mathbf{1}_{\{\sigma_i>\lambda\}}\left\{1+\sum_{j=1,j\neq i}^{p_1}\frac{\sigma_i\left(\sigma_i-\lambda\right)}{\sigma_i^2-\sigma_j^2} + \sum_{j=1,j\neq i}^{p_2}\frac{\sigma_i\left(\sigma_i-\lambda\right)}{\sigma_i^2-\sigma_j^2}\right\}. \tag{23}$$

Thus $df = E(\widehat{df}(\lambda))$ is the degrees of freedom of the nuclear-norm regularization multivariate regression.

5. Consistency of the Unbiased Estimate

The consistency of an estimate is important because it implies that the estimate is convergent to true value in probability. Suppose the estimated random variable is $T(X)$; we use statistical methods to get an estimate $\widehat{T}_n(X)$, which is a function of the size of sample. If $\widehat{T}_n(X)$ is a consistent estimate of $T(X)$, it means that, with the sample size increasing, $\widehat{T}_n(X)$ equals $T(X)$ almost everywhere. That is, for any $\epsilon > 0$, we can get

$$\lim_{n\to\infty} P\left\{\left|\widehat{T}_n(X) - T(X)\right| < \epsilon\right\} = 1. \tag{24}$$

In this section, we prove the consistent property of the estimates of the degrees of freedom given in the former sections. We will first prove the consistency of the unbiased estimate $\widehat{\mathrm{df}}$ of regularized matrix regression. To do so, we need the following proposition on the continuous property of $\widehat{\mathrm{df}}$.

Proposition 7. *An unbiased estimate of the degrees of freedom of regularized matrix regression model is*

$$\widehat{df} = \sum_{i=1}^{p}\mathbf{1}_{\{\sigma_i>\lambda\}}\left\{1+\sum_{j=1,j\neq i}^{p_1}\frac{\sigma_i\left(\sigma_i-\lambda\right)}{\sigma_i^2-\sigma_j^2} + \sum_{j=1,j\neq i}^{p_2}\frac{\sigma_i\left(\sigma_i-\lambda\right)}{\sigma_i^2-\sigma_j^2}\right\}, \tag{25}$$

where σ_i is the singular value of the least square estimate. In this case, the degrees of freedom $\widehat{df}$ are only continuous in $\{\lambda \mid \lambda \neq \sigma_i, i = 1, 2, \ldots, p\}$.

Proof. For any $\lambda \in (\sigma_m, \sigma_{m-1})$, we know that $\lambda < \sigma_{m-1} < \sigma_{m-2} < \cdots < \sigma_1$. So the degrees of freedom of regularized matrix regression model are written as

$$\widehat{\mathrm{df}} = \sum_{i=1}^{m-1}\left\{1+\sum_{j=1,j\neq i}^{p_1}\frac{\sigma_i\left(\sigma_i-\lambda\right)}{\sigma_i^2-\sigma_j^2} + \sum_{j=1,j\neq i}^{p_2}\frac{\sigma_i\left(\sigma_i-\lambda\right)}{\sigma_i^2-\sigma_j^2}\right\}. \tag{26}$$

It is obvious that $\widehat{\mathrm{df}}$ is a linear function on $\lambda \in (\sigma_m, \sigma_{m-1})$. Thus, $\widehat{\mathrm{df}}$ is continuous in $\{\lambda \mid \lambda \neq \sigma_i, i = 1, 2, \ldots, p\}$.

We next prove that $\widehat{\mathrm{df}}$ is not continuous in $\{\sigma_i, i = 1, 2, \ldots, p\}$. If $\lambda \in [\sigma_m, \sigma_{m-1})$ and $\lambda \to \sigma_m^+$, $\lambda < \sigma_{m-1} < \sigma_{m-2} < \cdots < \sigma_1$, we obtain

$$\lim_{\lambda \to \sigma_m^+} \widehat{\mathrm{df}} = \sum_{i=1}^{m-1} \left\{ 1 + \sum_{j=1, j\neq i}^{p_1} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} + \sum_{j=1, j\neq i}^{p_2} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} \right\}. \tag{27}$$

If $\lambda \in (\sigma_{m+1}, \sigma_m)$ and $\lambda \to \sigma_m^-$, $\lambda < \sigma_m < \sigma_{m-1} < \cdots < \sigma_1$, we have

$$\begin{aligned} \lim_{\lambda \to \sigma_m^-} \widehat{\mathrm{df}} &= \sum_{i=1}^{m} \left\{ 1 + \sum_{j=1, j\neq i}^{p_1} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} + \sum_{j=1, j\neq i}^{p_2} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} \right\} = \sum_{i=1}^{m-1} \left\{ 1 \right. \\ &\quad \left. + \sum_{j=1, j\neq i}^{p_1} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} + \sum_{j=1, j\neq i}^{p_2} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} \right\} + \left(1 \right. \\ &\quad \left. + \sum_{j=1, j\neq m}^{p_1} \frac{\sigma_m (\sigma_m - \sigma_m)}{\sigma_m^2 - \sigma_j^2} + \sum_{j=1, j\neq m}^{p_2} \frac{\sigma_m (\sigma_m - \sigma_m)}{\sigma_m^2 - \sigma_j^2} \right) \\ &= \sum_{i=1}^{m-1} \left\{ 1 + \sum_{j=1, j\neq i}^{p_1} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} \right. \\ &\quad \left. + \sum_{j=1, j\neq i}^{p_2} \frac{\sigma_i (\sigma_i - \sigma_m)}{\sigma_i^2 - \sigma_j^2} \right\} + 1. \end{aligned} \tag{28}$$

Therefore, we get $\lim_{\lambda \to \sigma_m^-} \widehat{\mathrm{df}} = \lim_{\lambda \to \sigma_m^+} \widehat{\mathrm{df}} + 1$. Clearly, $\widehat{\mathrm{df}}$ is not continuous in $\{\sigma_i, i = 1, 2, \ldots, p\}$. □

Now, we show the unbiased estimate $\widehat{\mathrm{df}}$ is consistent to the true degrees of freedom.

Theorem 8. *Suppose σ_i is the singular value of the least square estimate of the regularized matrix regression model, and $\lambda_n^* \to \lambda^* > 0$, where λ^* is not equal to the singular values, that means $\{\lambda^* \neq \sigma_i, i = 1, 2, \ldots, p\}$. Then, $\widehat{df}(\lambda_n^*) - df(\lambda_n^*) \to 0$ in probability.*

Proof. By assumption and Proposition 7, it holds that $\widehat{\mathrm{df}}$ is a continuous function in $\{\lambda \mid \lambda \neq \sigma_i, i = 1, 2, \ldots, p\}$. If we have a sequence λ_n^* satisfying $\lambda_n^* \to \lambda^* \neq \sigma_i, i = 1, 2, \ldots, p$, the continuous mapping theorem implies that $\lim_{n\to\infty} \widehat{\mathrm{df}}(\lambda_n^*) = \widehat{\mathrm{df}}(\lambda^*)$. Immediately, we see $\widehat{\mathrm{df}}(\lambda_n^*) \to_p \widehat{\mathrm{df}}(\lambda^*)$. By using the dominated convergence theorem, we can get

$$\mathrm{df}(\lambda_n^*) = E\left[\widehat{\mathrm{df}}(\lambda_n^*)\right] \longrightarrow \widehat{\mathrm{df}}(\lambda^*). \tag{29}$$

Hence, $\widehat{\mathrm{df}}(\lambda_n^*) - \mathrm{df}(\lambda_n^*) \to_p 0$. □

Notice that, for the vector case, Zou et al. [14] not only gave the unbiased estimate of the degrees of freedom of the Lasso model, but also proved the following consistency of the estimate.

Proposition 9. *For the Lasso model, if $\lambda_n^*/n \to \lambda^* > 0$ with λ^* being a nontransition point, $\widehat{df}(\lambda_n^*) - df(\lambda_n^*) \to 0$ in probability.*

Based on Theorems 3, 8 and Proposition 9, we can easily show the following theorem.

Theorem 10. *If $(\lambda_1^n, \lambda_2^n/n) \to (\lambda_1^*, \lambda_2^*) > 0$, where λ_1^* and λ_2^* satisfy the assumptions in Theorem 8 and Proposition 9, then, $\widehat{df}(\lambda_1^n, \lambda_2^n/n) - df(\lambda_1^n, \lambda_2^n/n) \to 0$ in probability.*

6. Conclusions

In this paper, we mainly obtain the degrees of freedom of mixed matrix regression model. Moreover, we prove that the obtained estimates of degrees of freedom are consistent. Note that our results of the degrees of freedom are given under the assumption that the prediction matrix and vector are independent. However, if they are not independent but in linear relationship or another nonlinear relationship, or the number of samples is less than the number of variables, what is the analytical form of degrees of freedom? We will leave this as a future research topic.

Appendix

In this part, we give the proofs of Propositions 4 and 5. We first give the proof of Proposition 4. To do so, we need the following results. See [5] for more details.

Proposition A.1. *For a given matrix A with singular value decomposition $A = U\,\mathrm{diag}(a)V^T$, $f \circ \sigma(B)$ denotes any function of singular vectors of B. The optimal solution to*

$$\min_B \quad \frac{1}{2}\|B - A\|_F^2 + f \circ \sigma(B) \tag{A.1}$$

shares the same singular vectors as A and its ordered singular values are the solution to

$$\min_{\mathbf{b}} \quad \frac{1}{2}\|\mathbf{b} - \mathbf{a}\|_2^2 + f(\mathbf{b}). \tag{A.2}$$

An immediate consequence of the above proposition is the well-known singular value thresholding formula for nuclear-norm regularization.

Corollary A.2. *For a given matrix A with singular value decomposition $A = U\,\mathrm{diag}(a)V^T$. The optimal solution to*

$$\min_B \quad \frac{1}{2}\|B - A\|_F^2 + \lambda\|B\|_* \tag{A.3}$$

shares the same singular vectors as A and its singular values are $b_i = (a_i - \lambda)_+$.

Before proving Proposition 4, we also need some lemmas.

Lemma A.3. *Suppose that $\widehat{B}_{LS}$ has singular decomposition $\widehat{B}_{LS} = U\Sigma V^T = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, $p = \min\{p_1, p_2\}$. The estimate $\widehat{B} = U\Sigma_{\lambda_1} V^T$, where Σ_{λ_1} has diagonal entries $(\sigma_i - \lambda_1)_+$.*

Proof. According to the result of Corollary A.2, we just need to show that

$$\left\|\mathbf{y} - \chi \star B\right\|_2^2 = \left\|B - B_{LS}\right\|_F^2. \tag{A.4}$$

Note that, for any matrix A, $\|A\|_F^2 = \langle \mathrm{vec}(A), \mathrm{vec}(A)\rangle$. Thus, we can get

$$\begin{aligned}\left\|B - B_{LS}\right\|_F^2 &= \left\langle \mathrm{vec}\left(B - B_{LS}\right), \mathrm{vec}\left(B - B_{LS}\right)\right\rangle \\ &= \left\langle \mathrm{vec}\left(B\right) - \chi^T\mathbf{y}, \mathrm{vec}\left(B\right) - \chi^T\mathbf{y}\right\rangle \\ &= \|B\|_F^2 - 2\left\langle \mathrm{vec}\left(B\right), \chi^T\mathbf{y}\right\rangle + \|\mathbf{y}\|_2^2.\end{aligned} \tag{A.5}$$

Direct calculation yields that $\langle \mathrm{vec}(B), \chi^T\mathbf{y}\rangle = \mathrm{vec}(B)^T\chi^T\mathbf{y} = (\chi\,\mathrm{vec}(B))^T\mathbf{y} = (\chi \star B)^T\mathbf{y} = \langle \chi \star B, \mathbf{y}\rangle = \langle \mathbf{y}, \chi \star B\rangle$. We then derive

$$\begin{aligned}\left\|\mathbf{y} - \chi \star B\right\|_2^2 &= \left\langle \mathbf{y} - \chi \star B, \mathbf{y} - \chi \star B\right\rangle \\ &= \|\mathbf{y}\|_2^2 - 2\left\langle \mathbf{y}, \chi \star B\right\rangle + \|B\|_F^2.\end{aligned} \tag{A.6}$$

□

Lemma A.4. *One has*

$$\begin{aligned}\mathrm{tr}\left\{D\widehat{B}\left(\mathbf{v}_i\right) D\mathbf{v}_i\left(\widehat{B}_{LS}\right)\right\} &= \mathbf{1}_{\{\sigma_i > \lambda_1\}} \sum_{j=1, j\neq i}^{j=p_2} \frac{\sigma_i\left(\sigma_i - \lambda_1\right)}{\sigma_i^2 - \sigma_j^2}, \\ \mathrm{tr}\left\{D\widehat{B}\left(\mathbf{u}_i\right) D\mathbf{u}_i\left(\widehat{B}_{LS}\right)\right\} &= \mathbf{1}_{\{\sigma_i > \lambda_1\}} \sum_{j=1, j\neq i}^{j=p_1} \frac{\sigma_i\left(\sigma_i - \lambda_1\right)}{\sigma_i^2 - \sigma_j^2}.\end{aligned} \tag{A.7}$$

Proof. Since $\widehat{B}_{LS} = U\Sigma V^T$, the eigenvectors of the symmetric matrix $\widehat{B}_{LS}^T\widehat{B}_{LS} = V\Sigma^2 V^T$ coincide with the right singular vectors of $\widehat{B}_{LS}$. Then, by the chain rule,

$$\begin{aligned}&D\widehat{B}\left(\mathbf{v}_i\right) D\mathbf{v}_i\left(\widehat{B}_{LS}\right) \\ &\quad = D\widehat{B}\left(\mathbf{v}_i\right) D\mathbf{v}_i\left(\widehat{B}_{LS}^T\widehat{B}_{LS}\right) D\left(\widehat{B}_{LS}^T\widehat{B}_{LS}\right)\left(\widehat{B}_{LS}\right).\end{aligned} \tag{A.8}$$

Now $D\widehat{B}(\mathbf{v}_i) = (\sigma_i - \lambda_1)\mathbf{1}_{\{\sigma_i > \lambda_1\}} I_{p_2} \otimes \mathbf{u}_i$.

By the well-known formula for the differential of eigenvector, $D\mathbf{v}_i(\widehat{B}_{LS}^T\widehat{B}_{LS}) = \mathbf{v}_i^T \otimes (\sigma_i I_{p_2} - \widehat{B}_{LS}^T\widehat{B}_{LS})^+$, where C^+ is the Moore-Penrose generalized inverse of a matrix C.

The Jacobian matrix of the symmetric product is $D(\widehat{B}_{LS}^T\widehat{B}_{LS})(\widehat{B}_{LS}) = (I_{p_2^2} + K_{p_2p_2})(I_{p_2} \otimes \widehat{B}_{LS}^T)$, where $K_{p_2p_2}$ is the commutation matrix.

Now, by cycle permutation invariance of the trace function, we have

$$\begin{aligned}&\mathrm{tr}\left\{\mathbf{1}_{\{\sigma_i > \lambda_1\}}\left(\sigma_i - \lambda\right)\left(I_{p_2} \otimes \mathbf{u}_i\right)\mathbf{v}_i^T\right. \\ &\quad \left.\otimes\left(\sigma_i I_{p_2} - \widehat{B}_{LS}^T\widehat{B}_{LS}\right)^+ I_{p_2^2} \otimes \widehat{B}_{LS}\right\} = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\left(\sigma_i\right. \\ &\quad \left.- \lambda_1\right)\mathrm{tr}\left\{\mathbf{v}_i \otimes \left(\sigma_i I_{p_2} - \widehat{B}_{LS}^T\widehat{B}_{LS}\right)^+ \widehat{B}_{LS}^T\mathbf{u}_i\right\} \\ &\quad = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\sigma_i\left(\sigma_i - \lambda_1\right)\mathrm{tr}\left(\mathbf{v}_i^T \otimes 0_{p_2}\right) = 0.\end{aligned} \tag{A.9}$$

Then,

$$\begin{aligned}&\mathrm{tr}\left\{\left(\sigma_i - \lambda\right)\mathbf{1}_{\{\sigma_i > \lambda_1\}}\left(I_{p_2} \otimes \mathbf{u}_i\right)\mathbf{v}_i^T\right. \\ &\quad \left.\otimes\left(\sigma_i I_{p_2} - \widehat{B}_{LS}^T\widehat{B}_{LS}\right)^+ K_{p_2p_2}\left(I_{p_2^2} \otimes \widehat{B}_{LS}\right)\right\} = \left(\sigma_i\right. \\ &\quad \left.- \lambda_1\right)\mathbf{1}_{\{\sigma_i > \lambda_1\}}\mathrm{tr}\left\{\left(\sigma_i I_{p_2} - \widehat{B}_{LS}^T\widehat{B}_{LS}\right)^+ \otimes \mathbf{u}_i\mathbf{v}_i^T\widehat{B}_{LS}^T\right\} \\ &\quad = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\sigma_i\left(\sigma_i - \lambda_1\right)\mathrm{tr}\left\{\left(\sigma_i I_{p_2} - \widehat{B}_{LS}^T\widehat{B}_{LS}\right)^+\right. \\ &\quad \left.\otimes \mathbf{u}_i\mathbf{u}_i^T\right\} = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\sum_{j=1, j\neq i}^{j=p_2}\frac{\sigma_i\left(\sigma_i - \lambda_1\right)}{\sigma_i^2 - \sigma_j^2}\mathrm{tr}\left\{\mathbf{u}_i\mathbf{u}_i^T\right\} \\ &\quad = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\sum_{j=1, j\neq i}^{j=p_2}\frac{\sigma_i\left(\sigma_i - \lambda_1\right)}{\sigma_i^2 - \sigma_j^2}.\end{aligned} \tag{A.10}$$

By symmetry, we also have

$$\mathrm{tr}\left\{D\widehat{B}\left(\mathbf{u}_i\right) D\mathbf{u}_i\left(\widehat{B}_{LS}\right)\right\} = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\sum_{j=1, j\neq i}^{j=p_1}\frac{\sigma_i\left(\sigma_i - \lambda_1\right)}{\sigma_i^2 - \sigma_j^2}. \tag{A.11}$$

□

Lemma A.5. *One has*

$$\mathrm{tr}\left\{D\widehat{B}\left(\sigma_i\right) D\sigma_i\left(\widehat{B}_{LS}\right)\right\} = \mathbf{1}_{\{\sigma_i > \lambda\}}. \tag{A.12}$$

Proof. As in the proof of Lemma A.4, we utilize the fact that σ_i is the positive square root of the eigenvalues η_i of the symmetric matrix $\widehat{B}_{LS}^T\widehat{B}_{LS}$. Then, by the chain rule and the Jacobian matrix of fitted value with respect to responses,

$$\begin{aligned}&D\widehat{B}\left(\sigma_i\right) D\sigma_i\left(\widehat{B}_{LS}\right) \\ &\quad = D\widehat{B}\left(\sigma_i\right) D\sigma_i\left(\eta_i\right) D\eta_i\left(\widehat{B}_{LS}^T\widehat{B}_{LS}\right) D\widehat{B}_{LS}^T\widehat{B}_{LS}\left(\widehat{B}_{LS}\right).\end{aligned} \tag{A.13}$$

By combining $D\widehat{B}(\sigma_i) = \mathbf{1}_{\{\sigma_i > \lambda_1\}}\mathbf{v}_i \otimes \mathbf{u}_i$, $D\sigma_i(\eta_i) = 1/2\sqrt{\eta_i} = 1/2\sigma_i$, $D\eta_i(\widehat{B}_{LS}^T\widehat{B}_{LS}) = \mathbf{v}_i^T \otimes \mathbf{v}_i^T$, and

$$D\left(\widehat{B}_{LS}^T\widehat{B}_{LS}\right)\left(\widehat{B}_{LS}\right) = \left(I_{p_2^2} + K_{p_2p_2}\right)\left(I_{p_2} \otimes \widehat{B}_{LS}^T\right), \tag{A.14}$$

we obtain that

$$\begin{aligned} D\widehat{B}(\sigma_i)\, D\sigma_i\left(\widehat{B}_{LS}\right) &= \mathbf{1}_{\{\sigma_i>\lambda_1\}}\frac{1}{2\sigma_i} \\ &\cdot \operatorname{tr}\left\{\mathbf{v}_i\otimes\mathbf{u}_i\cdot\mathbf{v}_i^T\otimes\mathbf{v}_i^T\mathbf{1}_{p_2^2}\left(I_{p_2}\otimes\widehat{B}_{LS}^T\right)\right\}+\mathbf{1}_{\{\sigma_i>\lambda_1\}} \\ &\cdot\frac{1}{2\sigma_i}\operatorname{tr}\left\{\mathbf{v}_i\otimes\mathbf{u}_i\cdot\mathbf{v}_i^T\otimes\mathbf{v}_i^T K_{p_2p_2}\left(I_{p_2}\otimes\widehat{B}_{LS}^T\right)\right\} \\ &= \mathbf{1}_{\{\sigma_i>\lambda_1\}}\frac{1}{2\sigma_i}\operatorname{tr}\left\{\mathbf{v}_i\mathbf{v}_i^T\otimes\mathbf{u}_i\mathbf{v}_i^T\widehat{B}_{LS}^T\right\}+\mathbf{1}_{\{\sigma_i>\lambda_1\}}\frac{1}{2\sigma_i} \\ &\cdot\operatorname{tr}\left\{\mathbf{v}_i\mathbf{v}_i^T\otimes\mathbf{u}_i\mathbf{v}_i^T\widehat{B}_{LS}^T\right\}=\mathbf{1}_{\{\sigma_i>\lambda_1\}}\frac{1}{\sigma_i} \\ &\cdot\operatorname{tr}\left\{\sigma_i\mathbf{v}_i\mathbf{v}_i^T\otimes\mathbf{u}_i\mathbf{u}_i^T\right\}=\mathbf{1}_{\{\sigma_i>\lambda_1\}}. \end{aligned} \tag{A.15}$$

□

Proof of Proposition 4. We only need to show that the optimal $\widehat{B}$ of our model is the solution to the following problem:

$$\min_B \quad \frac{1}{2}\left\|\mathbf{y}-\chi\star B\right\|_2^2+\lambda_1\|B\|_*. \tag{A.16}$$

The least square estimate of B in model (9) is the solution of the following:

$$\min_B \quad \frac{1}{2}\left\|(\mathbf{y}-\gamma\star\mathbf{z})-\chi\star B\right\|_2^2. \tag{A.17}$$

So $\operatorname{vec}(\widehat{B}_{LS})=(\chi^T\chi)^{-1}[\chi^T(\mathbf{y}-\gamma\star\mathbf{z})]=(\chi^T\chi)^{-1}\chi^T\mathbf{y}$. Under the assumption, it is interesting to find that it has no relationship with γ and can be get from the following model:

$$\min_B \quad \left\|\mathbf{y}-\chi\star B\right\|_2^2. \tag{A.18}$$

Thus, by Lemma A.3, we have

$$\begin{aligned} \operatorname{tr}\left\{D\widehat{B}\left(\widehat{B}_{LS}\right)\right\} &= \operatorname{tr}\left\{\sum_{i=1}^{p}\left[D\widehat{B}(\mathbf{v}_i)\, D\mathbf{v}_i\left(\widehat{B}_{LS}\right)\right.\right. \\ &\left.\left.+D\widehat{B}(\mathbf{u}_i)\, D\mathbf{u}_i\left(\widehat{B}_{LS}\right)+D\widehat{B}(\sigma_i)\, D\sigma_i\left(\widehat{B}_{LS}\right)\right]\right\}. \end{aligned} \tag{A.19}$$

By Lemmas A.4 and A.5, we easily yield the desired conclusion.

It is worth noting that Zou et al. [14] showed that the degrees of freedom of Lasso fit are that $\operatorname{df}(\lambda)=E|\mathscr{B}_\lambda|$, where $\mathscr{B}_\lambda$ is the effective set of the Lasso coefficient estimates $\widehat{\beta}$. Thus, we know that $\widehat{\operatorname{df}}=\|\widehat{\beta}\|_0$ is an unbiased estimation of the degrees of freedom. □

Proof of Proposition 5. As we mentioned in Section 3, under a differentiability condition on $\widehat{\mathbf{y}}(\lambda)$, $\widehat{\operatorname{df}}=\operatorname{tr}\{D\widehat{\mathbf{y}}(\mathbf{y})\}$ is an unbiased estimation of the degrees of freedom. By the chain rule,

$$\begin{aligned} \widehat{\operatorname{df}} &= \operatorname{tr}\left\{D\widehat{\mathbf{y}}(\mathbf{y})\right\} \\ &= \operatorname{tr}\left\{D\widehat{\mathbf{y}}\left(\widehat{\beta}\right)D\widehat{\beta}\left(\widehat{\beta}_{LS}\right)D\widehat{\beta}_{LS}(\mathbf{y})\right\}. \end{aligned} \tag{A.20}$$

Because $\widehat{\mathbf{y}}=X\widehat{\beta}$, we can get $D\widehat{\mathbf{y}}(\widehat{\beta})=X$. The usual least square estimate for Lasso model is defined by

$$\min_\beta \quad \|y-X\beta\|_2^2. \tag{A.21}$$

So $\widehat{\beta}_{LS}=(X^TX)^{-1}X^T\mathbf{y}$. If $X^TX=I$, $\widehat{\beta}_{LS}(y)=X^T$, then we have

$$\begin{aligned} \widehat{\operatorname{df}} &= \operatorname{tr}\left\{XD\widehat{\beta}\left(\widehat{\beta}_{LS}\right)X^T\right\}=\operatorname{tr}\left\{X^TXD\widehat{\beta}\left(\widehat{\beta}_{LS}\right)\right\} \\ &= \operatorname{tr}\left\{D\widehat{\beta}\left(\widehat{\beta}_{LS}\right)\right\}. \end{aligned} \tag{A.22}$$

So we can get

$$\widehat{\operatorname{df}}=\operatorname{tr}\left\{D\widehat{\beta}\left(\widehat{\beta}_{LS}\right)\right\}=\left\|\widehat{\beta}\right\|_0. \tag{A.23}$$

In the mixed case, under the assumptions, we obtain that the optimal $\widehat{\mathbf{z}}$ is the solution of the following:

$$\min_{\mathbf{z}} \quad \left\|y-\gamma\star\mathbf{z}\right\|_2^2+\lambda_2\|\mathbf{z}\|_1. \tag{A.24}$$

It has no relationship with χ. Thus, in a similar way, we easily obtain

$$\operatorname{tr}\left\{D\widehat{\mathbf{z}}\left(\widehat{\mathbf{z}}_{LS}\right)\right\}=\|\widehat{\mathbf{z}}\|_0. \tag{A.25}$$

The proof is completed. □

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities (2017JBM323) and the National Natural Science Foundation of China (11671029).

References

[1] B. Pete and V. Sara, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[2] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[3] Y. Li, W. Zhang, and X. Liu, "Stability of nonlinear stochastic discrete-time systems," *Journal of Applied Mathematics*, vol. 2013, Article ID 356746, 2013.

[4] X. Liu, Y. Li, and W. Zhang, "Stochastic linear quadratic optimal control with constraint for discrete-time systems," *Applied Mathematics and Computation*, vol. 228, pp. 264–270, 2014.

[5] H. Zhou and L. Li, "Regularized matrix regression," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 76, no. 2, pp. 463–483, 2014.

[6] Y. Zhao and W. Zhang, "Observer-based controller design for singular stochastic markov jump systems with state dependent noise," *Journal of Systems Science and Complexity*, vol. 29, pp. 946–958, 2016.

[7] H. Ma and Y. Jia, "Stability analysis for stochastic differential equations with infinite Markovian switchings," *Journal of Mathematical Analysis and Applications*, vol. 435, no. 1, pp. 593–605, 2016.

[8] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[9] C. L. Mallows, "Some comments on Cp," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[10] H. Akaike, "Information theory and an extension of tha maximum likelihood principle," in *Part of the Series Springer Series in Statistics, In second international symposium on information theory*, pp. 267–281, Springer, New York, NY, USA, 1973.

[11] B. Efron, "The estimation of prediction error: covariance penalties and cross-validation," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 619–642, 2004.

[12] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[13] T. Hastie and R. Tibshirani, *Generalized Additive Models*, Chapman & Hall, New York, NY, USA, 1990.

[14] H. Zou, T. Hastie, and R. Tibshirani, "On the "degrees of freedom" of the lasso," *The Annals of Statistics*, vol. 35, no. 5, pp. 2173–2192, 2007.

[15] J. Ye, "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 120–131, 1998.

[16] X. Shen and J. Ye, "Adaptive model selection," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 210–221, 2002.

[17] R. J. Tibshirani and J. Taylor, "Degrees of freedom in lasso problems," *The Annals of Statistics*, vol. 40, no. 2, pp. 1198–1232, 2012.

[18] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *The Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[19] M. Yuan, "Degrees of freedom in low rank matrix estimation," *Science China. Mathematics*, vol. 59, no. 12, pp. 2485–2502, 2016.

[20] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.