Annals of the Institute of Statistical Mathematics (2025) 77:491–516
https://doi.org/10.1007/s10463-025-00926-z

Tuning parameter selection for the adaptive nuclear norm regularized trace regression

Yiting Ma[1] · **Pan Shang**[1,2] · **Lingchen Kong**[1]

Received: 25 June 2024 / Revised: 16 December 2024 / Accepted: 6 February 2025 /
Published online: 27 March 2025

Abstract

Regularized models have been applied in lots of areas in recent years, with high dimensional data sets being popular. Because that tuning parameter decides the theoretical performance and computational efficiency of the regularized models, tuning parameter selection is a basic and important issue. We consider the tuning parameter selection for adaptive nuclear norm regularized trace regression, which achieves by the Bayesian information criterion (BIC). The proposed BIC is established with the help of an unbiased estimator of degrees of freedom. Under some regularized conditions, this BIC is proved to achieve the rank consistency of the tuning parameter selection. That is the model solution under selected tuning parameter converges to the true solution and has the same rank with that of the true solution in probability. Some numerical results are presented to evaluate the performance of the proposed BIC on tuning parameter selection.

Keywords Tuning parameter selection · Adaptive nuclear norm regularized trace regression · Bayesian information criterion · Degrees of freedom

This work was supported by the National Natural Science Foundation of China (12401430,12371322), and the Postdoctoral Fellowship Program of CPSF under Grant GZB20240801.

✉ Pan Shang
pshang@amss.ac.cn

Yiting Ma
22121645@bjtu.edu.cn

Lingchen Kong
konglchen@126.com

[1] School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

1 Introduction

To deal with high-dimensional data sets with special structures, there are plenties of works that focus on regularized models. These models can be unified as a minimization problem, whose objective function is composed with a loss function and some regularizer. So, existing works consider different loss functions and regularizer under different data settings. For instance, the famous Lasso in Tibshirani (1996) and lots of similar models in Tibshirani et al. (2005); Zou and Hastie (2005); Yuan and Lin (2006); Zou (2006); Zou and Zhang (2009); Fan and Li (2001); Zhang (2010) inducing sparsity, regularized matrix regression in Zhou and Li (2014) and some other models inducing low-rank solution in Yuan et al. (2007); Rothman et al. (2010); Lu et al. (2012); Liu et al. (2014); Yuan (2016); Zhu (2020); Wei and Lee (2022); Zou et al. (2022); Mazumder et al. (2010); Zhou and Li (2014); Zhao et al. (2017); Elsener and van de Geer (2018); Fan et al. (2019); Hamidi and Bayati (2022) and so on. Undoubtedly, these regularized models have been applied to lots of areas, see e.g., bioinformatics in Zhao et al. (2010), engineering in Noorossana et al. (2010), signal process in Davenport and Romberg (2016), system identification in Liu and Vandenberghe (2010) and so on. Among these models, there is an interesting one called the adaptive nuclear norm regularized trace regression, which is proposed by Bach (2008) and achieves the rank consistency. Therefore, we focus on this model in the paper.

For regularized models, their loss function and regularizer are summed together by the so-called tuning parameter. It seems that tuning parameter is insignificant, while the selection of this parameter decides the theory performance and computational effect of these models. Therefore, there are some tools can be used to select this parameter in Wu and Wang (2020), such as Akaike information criterion (AIC in Akaike (1970)), Bayesian information criterion (BIC in Schwarz (1978); Wang and Leng (2007); Wang et al. (2009, 2007); Wang and Zhu (2011); Kim et al. (2012); Fan and Tang (2013); Hirose et al. (2013); Sun et al. (2013); Yaguang et al. (2019); Abbruzzo et al. (2019)), cross-validation in Homrighausen and McDonald (2013, 2017); Lei (2020); Datta and Zou (2020); Chetverikov et al. (2021), and so on. Here, we review some related works about existing tuning parameter selection methods for regularized models, which include regularized liner models with the unknown variable being in vector form and regularized matrix models with matrix as variable, respectively. The following results mainly focus on tuning parameter selection for regularized linear models. Wang et al. (2007) established a BIC for SCAD under the case that the sample size n and feature size p satisfying $n > p$ and p fixed. Wang et al. (2009) considered the BIC for adaptive Lasso under $n > p$ and p diverged. Wang and Zhu (2011) builded up a high dimensional BIC (HBIC) for adaptive elastic net under $n < p$ and p diverged. Kim et al. (2012) builded a generalized information criterion for SCAD and MCP under different data settings, including $n < p$ with p fixed and diverged. Hirose et al. (2013) considered the extended regularized information criterion for adaptive Lasso. Yaguang et al. (2019) established the regularized information criterion for adaptive group regularized generalized

linear models. These proposed information criteria all were proved to have the consistency property, that is the selected tuning parameter will lead to a solution with the same active index as the true one in probability, when the sample size increases. So, these regularized linear models all have an important property, which is the selection consistency. In addition, there are a few tuning parameter selection results for regularized matrix regression. Zhou and Li (2014) proposed the AIC and BIC for regularized matrix regression under the orthogonal assumption of prediction matrix. Yuan (2016) calculated the degree of freedom of regularized matrix regression, which is the basic for some tuning parameter selection methods. However, these results of regularized matrix regression have no theoretical guarantees of consistency.

In this paper, we tend to select the tuning parameter for adaptive nuclear norm regularized trace regression proposed by Bach (2008). The core reason we consider this model is its rank consistency, i.e., the model solution achieved the true rank in probability with the sample size increasing. In addition, this model also includes the regularized matrix regression models in Zhou and Li (2014) and Yuan and Lin (2006) as special cases. We first establish some basic results, including the subdifferential and conjugate function of the adaptive nuclear norm. These results help to analyze the model and provide tools to establish its degrees of freedom, which is an essential element to set up the tuning parameter selection criterion. Based on the definition in Efron (2004), we calculate the unbiased estimation of degrees of freedom, according to the model optimality and basic results of the model. Although the degrees of freedom seems like a complex form, we prove that this result equals to the existing result in Zhou and Li (2014) under a special case. With the help of the estimator of the degrees of freedom, we propose a BIC to select the tuning parameter for the adaptive nuclear norm regularized trace regression. The proposed BIC is proved to achieve the rank selection consistency in probability, under some regularized conditions. That is, the tuning parameter selected by BIC achieves selection consistency and rank selection consistency simultaneously. Actually, there is also a similar AIC can be defined and proved having the same performance as BIC. On numerical experiments, we compare the proposed BIC and AIC with AICc and cross validation for tuning parameter selection on some simulation data and real data sets, which illustrates the efficiency of our criteria.

This paper is organized as follows. We review the adaptive nuclear norm regularized trace regression and present some basic results in Sect. 2. In Sect. 3, we establish the degrees of freedom and BIC with theoretical guarantees. Some numerical experiments and conclusion are showed in Sect. 4 and Sect. 5, respectively.

Notations: Let $M \in \mathbb{R}^{p_1 \times p_2}$ be a matrix. vec(M) denotes the vector in $\mathbb{R}^{p_1 p_2}$ obtained by stacking its columns into a single vector. For any $j \in \{1, 2, \cdots, p_2\}$ and $k \in \{1, 2, \cdots, p_1\}$, $M_{:,j}$ means the j_{th} column and $M_{k,:}$ means the k_{th} row of M. Suppose M has a singular value decomposition with nonincreasing singular values $\sigma_1(M) \geq \cdots \geq \sigma_r(M) > 0$ and $r \leq \min\left\{p_1, p_2\right\}$ is the rank of M. There are some related norms with singular values of M. The Frobenius norm $\|\cdot\|_F$ is defined as $\|M\|_F = \sqrt{\sum_{k=1}^{p_1} \sum_{j=1}^{p_2} M_{k,j}^2} = \sqrt{\sigma_1^2(M) + \cdots + \sigma_r^2(M)}$. The nuclear norm $\|\cdot\|_*$ is the

sum of non-zero singular values, i.e., $\|M\|_* = \sum_{i=1}^{r} \sigma_i(M)$. The spectral norm $\|\cdot\|_2$ is the largest singular value, i.e., $\|M\|_2 = \sigma_1(M)$. A symmetric matrix $M \in \mathbb{R}^{p\times p}$ is called positive semidefinite (positive definite), denoted as $M \succeq 0$ ($M \succ 0$), if $\boldsymbol{x}^\top M\boldsymbol{x} \geq 0\,(\boldsymbol{x}^\top M\boldsymbol{x} > 0)$ holds for any $0 \neq \boldsymbol{x} \in \mathbb{R}^p$. Any $M \in \mathbb{O}^p$ means that $M \in \mathbb{R}^{p\times p}$ and $M^\top M = I_p$. For any $M, N \in \mathbb{R}^{p_1\times p_2}$, their inner product $\langle M, N\rangle$ is defined as $\langle M, N\rangle = \sum_{k=1}^{p_1}\sum_{j=1}^{p_2} M_{kj}N_{kj}$. For any vector $\boldsymbol{x} \in \mathbb{R}^p$, the norm $\|\cdot\|$ is defined as $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{p} x_i^2}$. The notation $\boldsymbol{x} \in \mathbb{R}^p_{++}$ means all elements of $\boldsymbol{x}$ are positive. The indictor function of a set $\mathscr{A}$ is denoted as $\delta_{\mathscr{A}}(\cdot)$, which means the value of this function is zero when variable in the set $\mathscr{A}$ and is ∞ otherwise. For any index set I, its cardinal number is denoted as $|I|$ that counts the number in the index set I, and its complementary set is denoted as I^c. In this paper, the notation 0 may represent scalar, vector and matrix, which can be inferred from the context.

2 Preliminaries

This section presents the adaptive nuclear norm regularized trace regression with its analysis and show some optimal results of this model, which are foundations of the tuning parameter selection.

2.1 Model analysis

The statistical model of the trace regression is

$$y = \langle X, B^*\rangle + \epsilon,$$

where $X \in \mathbb{R}^{p_1\times p_2}$ is the predictor, $y \in \mathbb{R}$ is the responsor, $\epsilon \in \mathbb{R}$ is a random error and $B^* \in \mathbb{R}^{p_1\times p_2}$ is the true coefficient. By sampling n times, we get

$$y_i = \langle X_i, B^*\rangle + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

Let $\mathscr{X} = \left(\mathrm{vec}\left(X_1\right), \cdots, \mathrm{vec}\left(X_n\right)\right)^\top \in \mathbb{R}^{n\times p_1p_2}$, $\mathbf{y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ and the random error vector $\varepsilon = (\epsilon_1, \cdots, \epsilon_n)^\top \in \mathbb{R}^n$. These sample models can be written as

$$\mathbf{y} = \mathscr{X}\mathrm{vec}\left(B^*\right) + \varepsilon.$$

To estimate the unknown matrix B^*, there are some literatures focus on the nuclear norm regularized trace regression, such as Vladimir et al. (2011); Negahban and Wainwright (2011); Zhou and Li (2014) and so on, which is

$$\min_{B\in\mathbb{R}^{p_1\times p_2}} \left\{ \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle X_i, B\rangle\right)^2 + \lambda\|B\|_* \right\}.$$

Here, $\lambda > 0$ is the tuning parameter. However, this model can hardly achieve the rank consistent solution in high-dimensional case as in Bach (2008). So, an adaptive version was proposed as

$$\min_{B\in\mathbb{R}^{p_1\times p_2}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\langle X_i,B\rangle\right)^2+\lambda\|W_1BW_2\|_*\right\},\tag{1}$$

where W_1 and W_2 are given weight matrixes based on the solution of least square trace regression. The detailed results of these matrixes are showed as follows. Let

$$\hat{B}_{LS}=\arg\min_{B\in\mathbb{R}^{p_1\times p_2}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\langle X_i,B\rangle\right)^2\right\}.$$

Suppose $\hat{B}_{LS}$ has full rank and its singular value decomposition is

$$\hat{B}_{LS}=U_{LS}Diag\left(s^{LS}\right)V_{LS}^{\top},$$

where $U_{LS}\in\mathbb{O}^{p_1}$, $V_{LS}\in\mathbb{O}^{p_2}$, $s^{LS}\in\mathbb{R}^{\min\{p_1,p_2\}}$ and $\mathrm{Diag}\left(s^{LS}\right)$ is a matrix whose diagonal vector is the singular vector s^{LS}. s^{LS} can be completed by $n^{-1/2}$ to reach dimensions p_1 or p_2. Then,

$$W_1=U_{LS}\mathrm{Diag}\left(s^{LS}\right)^{-\gamma}U_{LS}^{\top},W_2=V_{LS}\mathrm{Diag}\left(s^{LS}\right)^{-\gamma}V_{LS}^{\top},\gamma\in(0,1].$$

Based on expressions of W_1 and W_2, they are obvious symmetric matrixes, i.e., $W_1^{\top}=W_1$ and $W_2^{\top}=W_2$. In addition, it is clear that $W_1^{-1}=U_{LS}\mathrm{Diag}\left(s^{LS}\right)^{\gamma}U_{LS}^{\top}$ and $W_2^{-1}=V_{LS}\mathrm{Diag}\left(s^{LS}\right)^{\gamma}V_{LS}^{\top}$.

To illustrate that the solution of (1) depends on the tuning parameter λ, we denote the solution as $\hat{B}(\lambda)$ in the following sections.

2.2 Basic results

In this section, some related results are proposed and reviewed, which provide theoretical guarantees for the rest of this paper. The following definitions are from Rockafellar (2015).

Definition 1 Let $f:\mathbb{R}^{p_1\times p_2}\to(-\infty,+\infty]$ be a proper closed convex function and let $M\in\mathbb{R}^{p_1\times p_2}$. The subdifferential of f at M is denoted by $\partial f(M)$ and

$$\partial f(M)=\left\{G\in\mathbb{R}^{p_1\times p_2}:f(N)\geq f(M)+\langle G,N-M\rangle,\forall N\in\mathbb{R}^{p_1\times p_2}\right\}.$$

According to this definition, we compute the subdifferential of $f(M)=\|W_1MW_2\|_*$ as follows.

Proposition 1 *For any $M\in\mathbb{R}^{p_1\times p_2}$, let r be the rank of W_1MW_2. The subdifferential of $\|W_1MW_2\|_*$ is*

$$\partial\|W_1MW_2\|_*=\left\{W_1\left(U_rV_r^{\top}+N\right)W_2:W_1MW_2=U_r\mathrm{Diag}\left(\sigma_r\right)V_r^{\top},\|N\|_2\leq1,U_r^{\top}N=0,NV_r=0\right\},$$

where $U_r\in\mathbb{R}^{p_1\times r}$, $V_r\in\mathbb{R}^{p_2\times r}$, $\sigma_r\in\mathbb{R}_{++}^{r}$ and $N\in\mathbb{R}^{p_1\times p_2}$.

Proof It is clear that $f(M) = h(g(M))$ with $g(M) = W_1 M W_2$ and $h(M) = \lambda \|M\|_*$. Because g is a linear transformation, according to (Beck 2017, Theorem 3.43),

$$\partial f(M) = g^\top \left(\partial_{g(M)} h(g(M))\right)$$

with $g^\top$ being the adjoint operator of g.

According to the definition of the adjoint operator, $\langle g(M), N\rangle = \left\langle M, g^\top(N)\right\rangle$, which leads to

$$g^\top(N) = W_1^\top N W_2^\top = W_1 N W_2.$$

Based on the singular value decomposition of $W_1 M W_2$ and the subdifferential of the nuclear norm in Watson (1992), it holds that

$$\partial_{g(M)} h(g(M)) = \left\{U_r V_r^\top + N \,:\, \|N\|_2 \le 1, U_r^\top N = 0, N V_r = 0\right\}.$$

Combing these results, the desired result can be obtained. □

There are some basic calculation rule in Horn and Johnson (2012); Magnus and Neudecker (2019) that are foundations of our result, which we review them in the next proposition.

Proposition 2 *Let A, B, C, D be any matrix in compatible dimensions, the following claims hold.*

i) $vec(ABC) = \left(C^\top \bigotimes A\right) vec(B)$, *where* $\bigotimes$ *means the Kronecker product.*
ii) $vec\left(A^\top\right) = K vec(A)$, *where K is the permutation matrix.*
iii) $(A \bigotimes B)(C \bigotimes D) = AC \bigotimes BD$.
iv) $\frac{\partial vec(ABC)}{\partial vec^\top(B)} = C^\top \bigotimes A$.

3 Bayesian information criterion

In this section, we tend to build up a Bayesian Information Criterion (BIC) and prove its consistency to select the appropriate tuning parameter for (1).

According to the definition of traditional BIC, a key basis is degrees of freedom of the model. So, we calculate degrees of freedom of (1) in the next theorem.

Theorem 1 *For any $\lambda \ge 0$, let r be the rank of $W_1 \hat{B}(\lambda) W_2$. Suppose the singular value decomposition of $W_1 \hat{B}(\lambda) W_2$ is $W_1 \hat{B}(\lambda) W_2 = U_r \mathrm{Diag}\left(b_r\right) V_r^\top$ with $U_r \in \mathbb{R}^{p_1 \times r}$, $V_r \in \mathbb{R}^{p_2 \times r}$ and $b_r \in \mathbb{R}^r_{++}$. An unbiased estimation of degrees of freedom of* (1) *under λ satisfies that*

$$\hat{\mathrm{df}}_\lambda = \frac{1}{n}\sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top M_r^+ \mathrm{vec}\big(X_k\big).$$

Here, $M_r = \frac{1}{n}\sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)\mathrm{vec}\big(X_k\big)^\top + \lambda M_r^{(1)} + M_r^{(2)}$ *is assumed full rank or symmetric, with* $M_r^{(1)}$ *and* $M_r^{(2)}$ *defined in Lemma 1 and Lemma 2 respectively, and* M_r^+ *denotes the Moore-Penrose generalized inverse matrix of* M_r.

Proof As the result in Stein (1981), an unbiased estimation of degrees of freedom of (1) under λ is

$$\hat{\mathrm{df}}_\lambda = \sum_{k=1}^{n} \frac{\partial \hat{y}_k}{\partial y_k} = \sum_{k=1}^{n} \left\langle X_k, \frac{\partial \hat{B}(\lambda)}{\partial y_k} \right\rangle = \sum_{k=1}^{n} \left\langle \mathrm{vec}\big(X_k\big), \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} \right\rangle.$$

According to the optimal condition of (1) and the subdifferential of $\|W_1 B W_2\|_*$ in Proposition 1, there is a matrix N such that $\|N\|_2 \le 1, U_r^\top N = 0, NV_r = 0$ and

$$-\frac{1}{n}\sum_{i=1}^{n} \big(y_i - \big\langle X_i, \hat{B}(\lambda)\big\rangle\big) X_i + \lambda W_1 G\big(\hat{B}(\lambda)\big) W_2 + \lambda W_1 N W_2 = 0, \tag{2}$$

where $G\big(\hat{B}(\lambda)\big) = U_r V_r^\top$. Vectorizing this equality and using *i)* in Proposition 2, we get that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} y_i \mathrm{vec}\big(X_i\big) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top \mathrm{vec}\big(\hat{B}(\lambda)\big) \\ + \lambda\Big(W_2 \bigotimes W_1\Big)\mathrm{vec}\big(G\big(\hat{B}(\lambda)\big)\big) + \lambda\Big(W_2 \bigotimes W_1\Big)\mathrm{vec}(N).\end{aligned} \tag{3}$$

The result of the partial differential of the equation (3) with respect of y_k is

$$\frac{1}{n}\mathrm{vec}(X_k) = \left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top + \lambda M_r^{(1)} + M_r^{(2)}\right] \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k}, \tag{4}$$

where

$$M_r^{(1)} = \Big(W_2 \bigotimes W_1\Big)\frac{\partial \mathrm{vec}\big(G\big(\hat{B}(\lambda)\big)\big)}{\partial \mathrm{vec}^\top\big(\hat{B}(\lambda)\big)}, M_r^{(2)} = \lambda\Big(W_2 \bigotimes W_1\Big)\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^\top\big(\hat{B}(\lambda)\big)}.$$

Denote $M_r = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top + \lambda M_r^{(1)} + M_r^{(2)}$. Then, the equation (4) is transformed as

$$\frac{1}{n}\mathrm{vec}\big(X_k\big) = M_r \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k}. \tag{5}$$

To get the result of $\hat{\mathrm{df}}_\lambda$, we assume that M_r has full rank or it is a symmetric matrix. The results are illustrated as follows.

If M_r has full rank, it is obvious that $\frac{\partial \mathrm{vec}\left(\hat{B}(\lambda)\right)}{\partial y_k} = \frac{1}{n} M_r^{-1} \mathrm{vec}\left(X_k\right)$ and

$$\hat{\mathrm{df}}_\lambda = \frac{1}{n} \sum_{k=1}^{n} \mathrm{vec}\left(X_k\right)^\top M_r^{-1} \mathrm{vec}\left(X_k\right).$$

If M_r is symmetric, suppose the eigenvalue decomposition of M_r is

$$M_r = V \begin{pmatrix} \mathrm{Diag}\left(\sigma\left(M_r\right)\right), & 0 \\ 0, & 0 \end{pmatrix} V^\top,$$

where $V = \left(\mathbf{v}_1, \cdots, \mathbf{v}_{r_M}, \mathbf{v}_{r_M+1}, \cdots, \mathbf{v}_{p_1 p_2}\right) = \left(V_{r_M}, V_{r_M^\perp}\right) \in \mathbb{O}^{p_1 p_2}$ with $V_{r_M^\perp}^\top V_{r_M} = 0$ and $\sigma\left(M_r\right) \in \mathbb{R}^{r_M}$ with r_M being the rank of M_r. Then, the Moore-Penrose generalized inverse matrix of M_r is

$$M_r^+ = V \begin{pmatrix} \mathrm{Diag}\left(\sigma\left(M_r\right)\right)^{-1}, & 0 \\ 0, & 0 \end{pmatrix} V^\top. \tag{6}$$

Based on the equation (5), we have

$$\begin{aligned} \frac{1}{n} V_{r_M^\perp}^\top \mathrm{vec}\left(X_k\right) &= V_{r_M^\perp}^\top M_r \frac{\partial \mathrm{vec}\left(\hat{B}(\lambda)\right)}{\partial y_k} = V_{r_M^\perp}^\top V \begin{pmatrix} \mathrm{Diag}\left(\sigma\left(M_r\right)\right), & 0 \\ 0, & 0 \end{pmatrix} V^\top \frac{\partial \mathrm{vec}\left(\hat{B}(\lambda)\right)}{\partial y_k} \\ &= V_{r_M^\perp}^\top V \begin{pmatrix} \mathrm{Diag}\left(\sigma\left(M_r\right)\right) V_{r_M}^\top \frac{\partial \mathrm{vec}\left(\hat{B}(\lambda)\right)}{\partial y_k} \\ 0 \end{pmatrix} \\ &= \left(V_{r_M^\perp}^\top V_{r_M}, V_{r_M^\perp}^\top V_{r_M^\perp}\right) \begin{pmatrix} \mathrm{Diag}\left(\sigma\left(M_r\right)\right) V_{r_M}^\top \frac{\partial \mathrm{vec}\left(\hat{B}(\lambda)\right)}{\partial y_k} \\ 0 \end{pmatrix} \\ &= \left(0, I_{p_1 p_2 - r_M}\right) \begin{pmatrix} \mathrm{Diag}\left(\sigma\left(M_r\right)\right) V_{r_M}^\top \frac{\partial \mathrm{vec}\left(\hat{B}(\lambda)\right)}{\partial y_k} \\ 0 \end{pmatrix} = 0. \end{aligned}$$

Combining the fact that $M_r^+ M_r = V_{r_M} V_{r_M}^\top$ and (6), we have

$$
\begin{aligned}
\hat{\mathrm{df}}_\lambda &= \sum_{k=1}^{n} \left\langle \mathrm{vec}\big(X_k\big), \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} \right\rangle \\
&= \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top V V^\top \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} \\
&= \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top M_r^+ M_r \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} + \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top V_{r_M^\perp} V_{r_M^\perp}^\top \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} \\
&\underset{(5)}{=} \frac{1}{n} \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top M_r^+ \mathrm{vec}\big(X_k\big) + \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top V_{r_M^\perp} V_{r_M^\perp}^\top \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} \\
&= \frac{1}{n} \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top M_r^+ \mathrm{vec}\big(X_k\big) + \sum_{k=1}^{n} \left\langle V_{r_M^\perp}^\top \mathrm{vec}\big(X_k\big), V_{r_M^\perp}^\top \frac{\partial \mathrm{vec}\big(\hat{B}(\lambda)\big)}{\partial y_k} \right\rangle \\
&\underset{(6)}{=} \frac{1}{n} \sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top M_r^+ \mathrm{vec}\big(X_k\big).
\end{aligned}
$$

These two results leads to Theorem 1. □

Next, we give detailed results of $M_r^{(1)}$ and $M_r^{(2)}$ in following lemmas.

Lemma 1 *For any $\lambda \geq 0$, let r be the rank of $W_1\hat{B}(\lambda)W_2$. Suppose the singular value decomposition of $W_1\hat{B}(\lambda)W_2$ is $W_1\hat{B}(\lambda)W_2 = U_r \mathrm{Diag}\big(b_r\big)V_r^\top$ with $U_r \in \mathbb{R}^{p_1 \times r}$, $V_r \in \mathbb{R}^{p_2 \times r}$ and $b_r \in \mathbb{R}_{++}^r$. Then,*

$$
\begin{aligned}
M_r^{(1)} &= \Big(W_2 \bigotimes W_1\Big) \frac{\partial \mathrm{vec}\big(G(\hat{B}(\lambda))\big)}{\partial \mathrm{vec}^\top\big(\hat{B}(\lambda)\big)} \\
&= \Big[W_2 V_r \mathrm{Diag}\big(b_r\big)^{-1} V_r^\top W_2\Big] \bigotimes W_1^2 + W_2^2 \bigotimes \Big[W_1 U_r \mathrm{Diag}\big(b_r\big)^{-1} U_r^\top W_1\Big].
\end{aligned} \tag{7}
$$

Proof Note that $M_r^{(1)} = \big(W_2 \bigotimes W_1\big) \frac{\partial \mathrm{vec}\left(G\left(\hat{B}(\lambda)\right)\right)}{\partial \mathrm{vec}^\top\left(\hat{B}(\lambda)\right)}$. It is obvious that

$$
\frac{\partial \mathrm{vec}\left(G\left(\hat{B}(\lambda)\right)\right)}{\partial \mathrm{vec}^\top\left(\hat{B}(\lambda)\right)} = \frac{\partial \mathrm{vec}\left(G\left(\hat{B}(\lambda)\right)\right)}{\partial \mathrm{vec}^\top\left(U_r\right)} \frac{\partial \mathrm{vec}\left(U_r\right)}{\partial \mathrm{vec}^\top\left(\hat{B}(\lambda)\right)} + \frac{\partial \mathrm{vec}\left(G\left(\hat{B}(\lambda)\right)\right)}{\partial \mathrm{vec}^\top\left(V_r^\top\right)} \frac{\partial \mathrm{vec}\left(V_r^\top\right)}{\partial \mathrm{vec}^\top\left(\hat{B}(\lambda)\right)}. \tag{8}
$$

(i) According to the fact that *i*) in Proposition 2, the following results hold.

$$
\begin{aligned}
\frac{\partial \mathrm{vec}\left(G\left(\hat{B}(\lambda)\right)\right)}{\partial \mathrm{vec}^\top\left(U_r\right)} &= \frac{\partial \mathrm{vec}\left(I_{p_1} U_r V_r^\top\right)}{\partial \mathrm{vec}^\top\left(U_r\right)} = \frac{\partial\left[\left(V_r \bigotimes I_{p_1}\right)\mathrm{vec}\left(U_r\right)\right]}{\partial \mathrm{vec}^\top\left(U_r\right)} = V_r \bigotimes I_{p_1}. \\
\frac{\partial \mathrm{vec}\left(G\left(\hat{B}(\lambda)\right)\right)}{\partial \mathrm{vec}^\top\left(V_r^\top\right)} &= \frac{\partial \mathrm{vec}\left(U_r V_r^\top I_{p_2}\right)}{\partial \mathrm{vec}^\top\left(V_r^\top\right)} = \frac{\partial\left[\left(I_{p_2} \bigotimes U_r\right)\mathrm{vec}\left(V_r^\top\right)\right]}{\partial \mathrm{vec}^\top\left(V_r^\top\right)} = I_{p_2} \bigotimes U_r.
\end{aligned}
$$

(ii) According to the fact that U_r and V_r have full column rank, there are $U_r^\top U_r = I_r$ and $V_r^\top V_r = I_r$. Combining these facts with $W_1\hat{B}(\lambda)W_2 = U_r\mathrm{Diag}(b_r)V_r^\top$, we have

$$U_r = W_1\hat{B}(\lambda)W_2V_r\mathrm{Diag}(b_r)^{-1},$$
$$V_r^\top = \mathrm{Diag}(b_r)^{-1}U_r^\top W_1\hat{B}(\lambda)W_2,$$

which leads to

$$\mathrm{vec}(U_r) = \left[\left(\mathrm{Diag}(b_r)^{-1}V_r^\top W_2\right)\bigotimes W_1\right]\mathrm{vec}(\hat{B}(\lambda)),$$
$$\mathrm{vec}(V_r^\top) = \left[W_2\bigotimes\left(\mathrm{Diag}(b_r)^{-1}U_r^\top W_1\right)\right]\mathrm{vec}(\hat{B}(\lambda)).$$

Thus,

$$\frac{\partial\mathrm{vec}(U_r)}{\partial\mathrm{vec}^\top(\hat{B}(\lambda))} = \left[\mathrm{Diag}(b_r)^{-1}V_r^\top W_2\right]\bigotimes W_1, \frac{\partial\mathrm{vec}(V_r^\top)}{\partial\mathrm{vec}^\top(\hat{B}(\lambda))} = W_2\bigotimes\left[\mathrm{Diag}(b_r)^{-1}U_r^\top W_1\right].$$

Combining results in (i) and (ii), we know that

$$\begin{aligned}M_r^{(1)} &= \left(W_2\bigotimes W_1\right)\frac{\partial\mathrm{vec}(G(\hat{B}(\lambda)))}{\partial\mathrm{vec}^\top(\hat{B}(\lambda))}\\ &= \left[W_2V_r\mathrm{Diag}(b_r)^{-1}V_r^\top W_2\right]\bigotimes W_1^2 + W_2^2\bigotimes\left[W_1U_r\mathrm{Diag}(b_r)^{-1}U_r^\top W_1\right].\end{aligned}$$

□

There are some semi-positive definite matrices, such as

$$W_2V_rDiag(b_r)^{-1}V_r^\top W_2 \succeq 0, W_1^2 \succ 0, W_2^2 \succ 0 and W_1U_rDiag(b_r)^{-1}U_r^\top W_1 \succeq 0.$$

Due to the result in (Magnus and Neudecker 2019, Corollary 2.1), we have

$$(W_2V_rDiag(b_r)^{-1}V_r^\top W_2)\bigotimes W_1^2 \succeq 0 and W_2^2\bigotimes(W_1U_rDiag(b_r)^{-1}U_r^\top W_1) \succeq 0,$$

which leads to the matrix $M_r^{(1)}$ is positive semidefinite.

Now, we calculate the result of $M_r^{(2)}$, which is decided by $\frac{\partial\mathrm{vec}(N)}{\partial\mathrm{vec}^\top(\hat{B}(\lambda))}$.

Lemma 2 *For any $\lambda \geq 0$, let r be the rank of $W_1\hat{B}(\lambda)W_2$. Suppose the singular value decomposition of $W_1\hat{B}(\lambda)W_2$ is $W_1\hat{B}(\lambda)W_2 = U_r\mathrm{Diag}(b_r)V_r^\top$ with $U_r \in \mathbb{R}^{p_1\times r}$, $V_r \in \mathbb{R}^{p_2\times r}$ and $b_r \in \mathbb{R}^r_{++}$. Then,*

$$
\begin{aligned}
M_r^{(2)} = &-\left[I_{p_2} \bigotimes \left(W_1\left(I_{p_1} - U_r U_r^{\top}\right)W_1^{-1}\right)\right] \cdot \frac{\mathscr{X}^{\top}\mathscr{X}}{n} \cdot \left[\left(W_2^{-1}V_r V_r^{\top}W_2\right) \bigotimes I_{p_1}\right] \\
&-\left[\left(W_2\left(I_{p_2} - V_r V_r^{\top}\right)W_2^{-1}\right) \bigotimes I_{p_1}\right] \cdot \frac{\mathscr{X}^{\top}\mathscr{X}}{n} \cdot \left[I_{p_2} \bigotimes \left(W_1^{-1}U_r U_r^{\top}W_1\right)\right] \\
&-\left[\left(W_2\left(I_{p_2} - V_r V_r^{\top}\right)W_2^{-1}\right) \bigotimes I_{p_1}\right] \cdot \frac{\mathscr{X}^{\top}\mathscr{X}}{n} \cdot \left[\left(W_2^{-1}V_r V_r^{\top}W_2\right) \bigotimes \left(W_1^{-1}U_r U_r^{\top}W_1\right)\right] \\
&-\lambda\left[\left(E_{\lambda}^{\top}W_1^{-1}\right) \bigotimes W_1\right]\left(I_{p_1^2} + K_{p_1^2}\right)\left[\left(\left((W_1\hat{B}(\lambda)W_2)^{+}\right)^{\top}W_2\right) \bigotimes W_1\right] \\
&-\lambda\left[W_2 \bigotimes \left(E_{\lambda}W_2^{-1}\right)\right]\left(I_{p_2^2} + K_{p_2^2}\right)\left[W_2 \bigotimes \left(\left(W_1\hat{B}(\lambda)W_2\right)^{+}W_1\right)\right],
\end{aligned}
\tag{9}
$$

where $E_{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i$ *and* $\left(W_1\hat{B}(\lambda)W_2\right)^{+} = V_r \mathrm{Diag}\left(b_r\right)^{-1}U_r^{\top}$.

Proof Note that $M_r^{(2)} = \lambda\left(W_2 \bigotimes W_1\right)\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}$. Based on the optimality condition (2), the matrix N can be expressed as

$$
N = \frac{1}{n\lambda}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1} - U_r V_r^{\top}.
$$

According to $U_r^{\top}N = 0$ and $NV_r = 0$, the following results hold.

$$
\begin{aligned}
&\frac{1}{n\lambda}U_r^{\top}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1} \\
&= V_r^{\top}, \frac{1}{n\lambda}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1}V_r = U_r.
\end{aligned}
$$

Replacing these results into N, then N can be expressed as

$$
\begin{aligned}
N = &\frac{1}{n\lambda}W_1^{-1}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i W_2^{-1} \\
&-\frac{1}{n^2\lambda^2}W_1^{-1}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i W_2^{-1}V_r U_r^{\top}W_1^{-1}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i W_2^{-1} \\
= &\frac{1}{n\lambda}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1}\left[I_{p_2} - \frac{1}{n\lambda}V_r U_r^{\top}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1}\right] \\
= &\left[I_{p_1} - \frac{1}{n\lambda}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1}V_r U_r^{\top}\right]\frac{1}{n\lambda}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1}.
\end{aligned}
$$

Let $\tilde{N} = \frac{1}{n\lambda}W_1^{-1}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right)W_2^{-1}$. Then,

$$N = \frac{1}{n\lambda} W_1^{-1} \left(\sum_{i=1}^{n} \left(y_i - \left\langle X_i, \hat{B}(\lambda) \right\rangle \right) X_i \right) W_2^{-1} \left[I_{p_2} - V_r V_r^{\top} \right] = \tilde{N} \left[I_{p_2} - V_r V_r^{\top} \right]$$
$$= \frac{1}{n\lambda} \left[I_{p_1} - U_r U_r^{\top} \right] W_1^{-1} \left(\sum_{i=1}^{n} \left(y_i - \left\langle X_i, \hat{B}(\lambda) \right\rangle \right) X_i \right) W_2^{-1} = \left[I_{p_1} - U_r U_r^{\top} \right] \tilde{N}. \tag{10}$$

To verify this N satisfies the required conditions, there is only one left part needed to be proved, i.e, $\|N\|_2 \leq 1$. According to the equation (2),

$$\tilde{N} = \frac{1}{n\lambda} W_1^{-1} \left(\sum_{i=1}^{n} \left(y_i - \left\langle X_i, \hat{B}(\lambda) \right\rangle \right) X_i \right) W_2^{-1} = \partial \left\| \hat{B}(\lambda) \right\|_*,$$

which leads to $\left\| \tilde{N} \right\|_2 \leq 1$. Based on the singular inequalities in Horn and Johnson (2012),

$$\|N\|_2 = \left\| \tilde{N} \left[I_{p_2} - V_r V_r^{\top} \right] \right\|_2 \leq \|\tilde{N}\|_2 \cdot \left\| I_{p_2} - V_r V_r^{\top} \right\|_2 \leq 1.$$

So, the matrix N in (2) can be expressed as in (10).

Now, we calculate the result of $\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}$. It is sure that

$$\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)} = \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(U_r\right)} \cdot \frac{\partial \mathrm{vec}\left(U_r\right)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)} + \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)} \cdot \frac{\partial \mathrm{vec}\left(V_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}$$
$$+ \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\mathrm{Diag}\left(b_r\right)\right)} \cdot \frac{\partial \mathrm{vec}\left(\mathrm{Diag}\left(b_r\right)\right)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}.$$

(a) Based on the expression of N,

$$\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(U_r\right)} = \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\tilde{N}\right)} \cdot \frac{\partial \mathrm{vec}\left(\tilde{N}\right)}{\partial \mathrm{vec}^{\top}\left(U_r\right)} + \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(I_{p_1} - U_r U_r^{\top}\right)} \cdot \frac{\partial \mathrm{vec}\left(I_{p_1} - U_r U_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(U_r\right)}.$$

$$\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\tilde{N}\right)} = \frac{\partial \mathrm{vec}\left(\left(I_{p_2} \bigotimes \left(I_{p_1} - U_r U_r^{\top}\right)\right) \mathrm{vec}\left(\tilde{N}\right)\right)}{\partial \mathrm{vec}^{\top}\left(\tilde{N}\right)} = I_{p_2} \bigotimes \left(I_{p_1} - U_r U_r^{\top}\right).$$

Due to the fact that

$$\mathrm{vec}(\tilde{N}) = \frac{1}{n\lambda} \left(W_2^{-1} \bigotimes W_1^{-1} \right) \mathrm{vec} \left(\sum_{i=1}^{n} \left(y_i - \left\langle X_i, \hat{B}(\lambda) \right\rangle \right) X_i \right)$$
$$= \frac{1}{n\lambda} \left(W_2^{-1} \bigotimes W_1^{-1} \right) \left[\mathcal{X}^{\top} \mathbf{y} - \mathcal{X}^{\top} \mathcal{X} \mathrm{vec}\left(\hat{B}(\lambda)\right) \right]$$
$$= \frac{1}{n\lambda} \left(W_2^{-1} \bigotimes W_1^{-1} \right) \left[\mathcal{X}^{\top} \mathbf{y} - \mathcal{X}^{\top} \mathcal{X} \left(\left(W_2^{-1} V_r \mathrm{Diag}\left(b_r\right) \right) \bigotimes W_1^{-1} \right) \mathrm{vec}\left(U_r\right) \right],$$

the result of $\frac{\partial \mathrm{vec}(\tilde{N})}{\partial \mathrm{vec}^{\top}\left(U_r\right)}$ is

$$\frac{\partial \mathrm{vec}(\tilde{N})}{\partial \mathrm{vec}^{\top}(U_r)} = -\frac{1}{n\lambda}\left(W_2^{-1} \bigotimes W_1^{-1}\right)\mathscr{X}^{\top}\mathscr{X}\left[\left(W_2^{-1}V_r\mathrm{Diag}(b_r)\right) \bigotimes W_1^{-1}\right].$$

It is easy to get that

$$\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(I_{p_1} - U_r U_r^{\top}\right)} = \frac{\partial\left(\left(\tilde{N}^{\top} \bigotimes I_{p_1}\right)\mathrm{vec}\left(I_{p_1} - U_r U_r^{\top}\right)\right)}{\partial \mathrm{vec}^{\top}\left(I_{p_1} - U_r U_r^{\top}\right)} = \tilde{N}^{\top} \bigotimes I_{p_1}.$$

The claim *ii*) in Proposition 2 leads to $\mathrm{vec}\left(U_r^{\top}\right) = K_{rp_1}\mathrm{vec}\left(U_r\right)$ and

$$\begin{aligned}\frac{\partial \mathrm{vec}\left(I_{p_1} - U_r U_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(U_r\right)} &= -\frac{\partial \mathrm{vec}\left(U_r U_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(U_r\right)} - \frac{\partial \mathrm{vec}\left(U_r U_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(U_r^{\top}\right)} \cdot \frac{\partial \mathrm{vec}(U_r^{\top})}{\partial \mathrm{vec}^{\top}\left(U_r\right)} \\ &= -U_r \bigotimes I_{p_1} - \left(I_{p_1} \bigotimes U_r\right)K_{rp_1} \\ &= -U_r \bigotimes I_{p_1} - K_{p_1^2}\left(U_r \bigotimes I_{p_1}\right) = -\left(I_{p_1^2} + K_{p_1^2}\right)\left(U_r \bigotimes I_{p_1}\right).\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}(U_r)} = &-\frac{1}{n\lambda}\left[W_2^{-1} \bigotimes (I_{p_1} - U_r U_r^{\top})W_1^{-1}\right]\mathscr{X}^{\top}\mathscr{X}\left[\left(W_2^{-1}V_r\mathrm{Diag}(b_r)\right) \bigotimes W_1^{-1}\right] \\ &- \left(\tilde{N}^{\top} \bigotimes I_{p_1}\right)\left(I_{p_1^2} + K_{p_1^2}\right)\left(U_r \bigotimes I_{p_1}\right).\end{aligned}$$

(b) Based on the expression of N, $\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)} = \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}(\tilde{N})} \cdot \frac{\partial \mathrm{vec}(\tilde{N})}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)} + + \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(I_{p_2} - V_r V_r^{\top}\right)} \cdot \frac{\partial \mathrm{vec}\left(I_{p_2} - V_r V_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)}$.

$$\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\tilde{N}\right)} = \frac{\partial \mathrm{vec}\left(\left(\left(I_{p_2} - V_r V_r^{\top}\right) \otimes I_{p_1}\right)\mathrm{vec}(\tilde{N})\right)}{\partial \mathrm{vec}^{\top}\left(\tilde{N}\right)} = \left(I_{p_2} - V_r V_r^{\top}\right) \bigotimes I_{p_1}.$$

Due to the fact that

$$\begin{aligned}\mathrm{vec}\left(\tilde{N}\right) &= \frac{1}{n\lambda}\left(W_2^{-1} \bigotimes W_1^{-1}\right)\mathrm{vec}\left(\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)X_i\right) \\ &= \frac{1}{n\lambda}\left(W_2^{-1} \bigotimes W_1^{-1}\right)\left[\mathscr{X}^{\top}\mathbf{y} - \mathscr{X}^{\top}\mathscr{X}\mathrm{vec}(\hat{B}(\lambda))\right] \\ &= \frac{1}{n\lambda}\left(W_2^{-1} \bigotimes W_1^{-1}\right)\left[\mathscr{X}^{\top}\mathbf{y} - \mathscr{X}^{\top}\mathscr{X}\left(W_2^{-1} \bigotimes \left(W_1^{-1}U_r\mathrm{Diag}(b_r)\right)\right)\mathrm{vec}\left(V_r^{\top}\right)\right],\end{aligned}$$

the result of $\frac{\partial \mathrm{vec}(\tilde{N})}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)}$ is

$$\frac{\partial \mathrm{vec}\left(\tilde{N}\right)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)} = -\frac{1}{n\lambda}\left(W_2^{-1} \bigotimes W_1^{-1}\right)\mathscr{X}^{\top}\mathscr{X}\left[W_2^{-1} \bigotimes \left(W_1^{-1}U_r\mathrm{Diag}(b_r)\right)\right].$$

It is easy to get that

$$\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(I_{p_2}-V_r V_r^{\top}\right)}=\frac{\partial \mathrm{vec}\left(\left(I_{p_2}\bigotimes \tilde{N}\right)\mathrm{vec}\left(I_{p_2}-V_r V_r^{\top}\right)\right)}{\partial \mathrm{vec}^{\top}\left(I_{p_2}-V_r V_r^{\top}\right)}=I_{p_2}\bigotimes \tilde{N}.$$

From the result that $\mathrm{vec}\left(V^{\top}\right)=K_{rp_2}\mathrm{vec}\left(V_r\right)$, we have

$$\begin{aligned}\frac{\partial \mathrm{vec}\left(I_{p_2}-V_r V_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)}&=-\frac{\partial \mathrm{vec}\left(V_r V_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)}-\frac{\partial \mathrm{vec}\left(V_r V_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(V_r\right)}\cdot\frac{\partial \mathrm{vec}\left(V_r\right)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)}\\&=-I_{p_2}\bigotimes V_r-\left(V_r\bigotimes I_{p_1}\right)K_{rp_2}\\&=-I_{p_2}\bigotimes V_r-K_{p_2^2}\left(I_{p_2}\bigotimes V_r\right)=-\left(I_{p_2^2}+K_{p_2^2}\right)\left(I_{p_2}\bigotimes V_r\right).\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(V_r^{\top}\right)}=&-\frac{1}{n\lambda}\left[\left(\left(I_{p_2}-V_r V_r^{\top}\right)W_2^{-1}\right)\bigotimes W_1^{-1}\right]\mathscr{Z}^{\mathrm{T}}\mathscr{Z}\left[W_2^{-1}\bigotimes\left(W_1^{-1}U_r\mathrm{Diag}\left(b_r\right)\right)\right]\\&-\left(I_{p_2}\bigotimes \tilde{N}\right)\left(I_{p_2^2}+K_{p_2^2}\right)\left(I_{p_2}\bigotimes V_r\right).\end{aligned}$$

(c) In the similar way, the result of $\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\mathrm{Diag}\left(b_r\right)\right)}$ is showed as follows.

$$\begin{aligned}&\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\mathrm{Diag}\left(b_r\right)\right)}=\frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^{\top}\left(\tilde{N}\right)}\cdot\frac{\partial \mathrm{vec}\left(\tilde{N}\right)}{\partial \mathrm{vec}^{\top}\left(\mathrm{Diag}\left(b_r\right)\right)}\\&=-\frac{1}{n\lambda}\left[\left(I_{p_2}-V_r V_r^{\top}\right)\bigotimes I_{p_1}\right]\cdot\left(W_2^{-1}\bigotimes W_1^{-1}\right)\mathscr{Z}^{\mathrm{T}}\mathscr{Z}\left[\left(W_2^{-1}V_r\right)\bigotimes\left(W_1^{-1}U_r\right)\right]\\&=-\frac{1}{n\lambda}\left[\left(\left(I_{p_2}-V_r V_r^{\top}\right)W_2^{-1}\right)\bigotimes W_1^{-1}\right)\mathscr{Z}^{\mathrm{T}}\mathscr{Z}\left[\left(W_2^{-1}V_r\right)\bigotimes\left(W_1^{-1}U_r\right)\right].\end{aligned}$$

(d) Because $W_1\hat{B}(\lambda)W_2=U_r\mathrm{Diag}(b_r)V_r^{\top}$, then $\mathrm{Diag}(b_r)=U_r^{\top}W_1\hat{B}(\lambda)W_2V_r$ and

$$\frac{\partial \mathrm{vec}\left(Diag\left(b_r\right)\right)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}=\left(V_r^{\top}W_2\right)\bigotimes\left(U_r^{\top}W_1\right).$$

In addition, results of $\frac{\partial \mathrm{vec}\left(U_r\right)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}$ and $\frac{\partial \mathrm{vec}\left(V_r^{\top}\right)}{\partial \mathrm{vec}^{\top}\left(\hat{B}(\lambda)\right)}$ have been computed in Lemma 1.

From (a)-(d), we know that

$$
\begin{aligned}
M_r^{(2)} &= \lambda\Big(W_2 \bigotimes W_1\Big) \cdot \frac{\partial \mathrm{vec}(N)}{\partial \mathrm{vec}^\top\big(\hat{B}(\lambda)\big)} \\
&= -\Big[I_{p_2} \bigotimes \big(W_1(I_{p_1} - U_r U_r^\top)W_1^{-1}\big)\Big] \cdot \frac{\mathscr{X}^\mathsf{T}\mathscr{X}}{n} \cdot \Big[\big(W_2^{-1}V_r V_r^\top W_2\big) \bigotimes I_{p_1}\Big] \\
&\quad - \Big[\big(W_2(I_{p_2} - V_r V_r^\top)W_2^{-1}\big) \bigotimes I_{p_1}\Big] \cdot \frac{\mathscr{X}^\mathsf{T}\mathscr{X}}{n} \cdot \Big[I_{p_2} \bigotimes \big(W_1^{-1}U_r U_r^\top W_1\big)\Big] \\
&\quad - \Big[\big(W_2(I_{p_2} - V_r V_r^\top)W_2^{-1}\big) \bigotimes I_{p_1}\Big] \cdot \frac{\mathscr{X}^\mathsf{T}\mathscr{X}}{n} \cdot \Big[\big(W_2^{-1}V_r V_r^\top W_2\big) \bigotimes \big(W_1^{-1}U_r U_r^\top W_1\big)\Big] \\
&\quad - \lambda\Big[\Big(W_2\tilde{N}^\top\Big) \bigotimes W_1\Big]\Big(I_{p_1^2} + K_{p_1^2}\Big)\Big[\big(U_r \mathrm{Diag}(b_r)^{-1}V_r^\top W_2\big) \bigotimes W_1\Big] \\
&\quad - \lambda\Big[W_2 \bigotimes \big(W_1\tilde{N}\big)\Big]\Big(I_{p_2^2} + K_{p_2^2}\Big)\Big[W_2 \bigotimes \big(V_r \mathrm{Diag}(b_r)^{-1}U_r^\top W_1\big)\Big].
\end{aligned}
$$

By simple computation, we know that $\big(W_1\hat{B}(\lambda)W_2\big)^+ = V_r \mathrm{Diag}\big(b_r\big)^{-1}U_r^\top$. Replace the expression of $\tilde{N}$ and $E_\lambda = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \big\langle X_i, \hat{B}(\lambda)\big\rangle\big)X_i$ into $M_r^{(2)}$. Then, we have

$$
\begin{aligned}
M_r^{(2)} &= -\Big[I_{p_2} \bigotimes \big(W_1(I_{p_1} - U_r U_r^\top)W_1^{-1}\big)\Big] \cdot \frac{\mathscr{X}^\mathsf{T}\mathscr{X}}{n} \cdot \Big[\big(W_2^{-1}V_r V_r^\top W_2\big) \bigotimes I_{p_1}\Big] \\
&\quad - \Big[\big(W_2(I_{p_2} - V_r V_r^\top)W_2^{-1}\big) \bigotimes I_{p_1}\Big] \cdot \frac{\mathscr{X}^\mathsf{T}\mathscr{X}}{n} \cdot \Big[I_{p_2} \bigotimes \big(W_1^{-1}U_r U_r^\top W_1\big)\Big] \\
&\quad - \Big[\big(W_2(I_{p_2} - V_r V_r^\top)W_2^{-1}\big) \bigotimes I_{p_1}\Big] \cdot \frac{\mathscr{X}^\mathsf{T}\mathscr{X}}{n} \cdot \Big[\big(W_2^{-1}V_r V_r^\top W_2\big) \bigotimes \big(W_1^{-1}U_r U_r^\top W_1\big)\Big] \\
&\quad - \lambda\Big[\big(E_\lambda^\top W_1^{-1}\big) \bigotimes W_1\Big]\Big(I_{p_1^2} + K_{p_1^2}\Big)\Big[\Big(\Big(\big(W_1\hat{B}(\lambda)W_2\big)^+\Big)^\top W_2\Big) \bigotimes W_1\Big] \\
&\quad - \lambda\Big[W_2 \bigotimes \big(E_\lambda W_2^{-1}\big)\Big]\Big(I_{p_2^2} + K_{p_2^2}\Big)\Big[W_2 \bigotimes \Big(\big(W_1\hat{B}(\lambda)W_2\big)^+ W_1\Big)\Big].
\end{aligned}
$$

□

Remark 1 There are some facts about the result of this degrees of freedom.

In (1), if $\lambda = 0$ and $\hat{B}_{LS}$ has full rank, we have $M_r^{(2)} = 0$ and $M_r = \frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top$, which leads to

$$
\hat{\mathrm{df}}_\lambda = \frac{1}{n}\sum_{k=1}^{n}\mathrm{vec}(X_k)^\top\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top\right]^+\mathrm{vec}(X_k) = \mathrm{rank}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top\right).
$$

When λ is sufficiently large such that $\hat{B}(\lambda) = 0$, we have $M_r^{(1)} = 0$ and $M_r^{(2)} = 0$, which also leads to

$$
\hat{\mathrm{df}}_\lambda = \mathrm{rank}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\big(X_i\big)\mathrm{vec}\big(X_i\big)^\top\right).
$$

In Zhou and Li (2014), they considered the special case of (1) with the weight matrixes $W_1 = I_{p_1}$ and $W_2 = I_{p_2}$. Mathematically, their model can be regarded as the a special case of (1). In Zhou and Li (2014), they proved the value of degrees of

freedom of their model, under assumptions $\sum_{i=1}^{n} \mathrm{vec}(X_i)\mathrm{vec}(X_i)^\top = I_{p_1p_2}$ and the singular values of $\hat{B}_{LS}$ are not same. Practically, the setting $W_1 = I_{p_1}$ and $W_2 = I_{p_2}$ means that all singular values of $\hat{B}_{LS}$ being 1, which contradicts the assumption in Zhou and Li (2014). Ideally, under the condition $\sum_{i=1}^{n} \mathrm{vec}(X_i)\mathrm{vec}(X_i)^\top = I_{p_1p_2}$ and $\lambda = 0$, our degrees of freedom is $\hat{\mathrm{df}}_0 = p_1p_2$, which equals to the degrees of freedom in their paper.

Remark 2 The proof pattern of Theorem 1 not only fits (1) with W_1 and W_2 generated from the solution of least square trace regression, but also can be easily applied to other choices. There are three cases to be analyzed. Cases I: One can see that the proof pattern utilizes that the model (1) is a convex problem, and matrixes W_1 and W_2 are symmetric and invertible, which implies the result in Theorem 1 still holds when other adaptive choices satisfy these properties. Cases II: Matrixes W_1 and W_2 are not symmetric or invertible, but the resulting model (1) is a convex problem. In this case, the proof pattern can be directly applied, and the result may be slightly different with the inverse matrix being the Moore-Penrose generalized inverse matrix and some delicate changes about the transpose of these matrixes, i.e., for $i = 1, 2$, W_i^{-1} needs to be replaces with W_i^{+}, and some W_i need to be replaced with $W_i^{\top}$ according to the context. Cases III: Matrixes W_1 and W_2 are selected such that the resulting model (1) is not a convex problem, a new method needs to be explored.

For any $\lambda \geq 0$, we define a Bayesian information criterion (BIC) as

$$\mathrm{BIC}_\lambda = \log\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)^2\right) + \hat{\mathrm{df}}_\lambda \cdot \frac{\log(n)}{n}, \tag{11}$$

where $\hat{B}(\lambda)$ is the solution of (1) and $\hat{\mathrm{df}}_\lambda$ is given in Theorem 1. Then, the optimal tuning parameter is

$$\lambda^* = \arg\min_{\lambda\geq 0} BIC_\lambda = \arg\min_{\lambda\geq 0}\left\{\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)^2\right) + \hat{\mathrm{df}}_\lambda \cdot \frac{\log(n)}{n}\right\}.$$

To prove the selection result of this new defined BIC, we introduce some technique conditions and new notations as follows.

- (C1) $\{X_i\}_{i=1}^n$ are sampled i.i.d. from X and $\{y_i\}_{i=1}^n$ are sampled i.i.d. from y. The predictor X is standardized with $\|X\|_F = 1$. X and y have finite fourth order moments, i.e.,

$$\max_{j,k}\left\{\mathrm{E}\left(X_{j,k}^4\right)\right\} < \infty \; and \; E\left(y^4\right) < \infty.$$

- (C2) $\exists B^*$ such that $B^* \neq 0$, $\mathrm{rank}(B^*) = r^* < \min_{\{p_1,p_2\}}$ and

$$E\left(y_i | X_1, X_2, \cdots, X_n\right) = \langle X_i, B^*\rangle, \quad i = 1, 2, \cdots, n.$$

$$Var\left(y_i|X_1, X_2, \cdots, X_n\right) = \sigma^2, \quad i = 1, 2, \cdots, n.$$

- (C3) $\Sigma = \mathrm{E}\left(\mathrm{vec}(X)\mathrm{vec}(X)^{\top}\right)$ is positive definite, which means there is a $\kappa > 0$ such that the smallest eigenvalue of Σ is larger than κ.
- (C4) X and ϵ are independent with $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

Remark 3 Condition (C2) states that there exists a low rank matrix B^* such that the relationship of X and y is linear, which is a common assumption and ensures that the true solution has low rank. Conditions (C1) and (C3) are nature sampling assumptions. Condition (C4) guarantees that the random error is independent with the predictor.

Under conditions (C1)-(C4), (Bach, 2008, Theorem 15) proved the consistency and rank consistency of the solution of (1). That is, if $\gamma \in (0, 1]$, $\lambda_n n^{1/2+\gamma/2} \to \infty$ and $\lambda_n n^{1/2} \to 0$ with $n \to \infty$, $\lim_{n\to\infty} \mathrm{P}\left(\hat{B}(\lambda_n) = B^*\right) = 1$ and $\lim_{n\to\infty} \mathrm{P}(\mathrm{r}_{\lambda_n} = r^*) = 1$.

Let Ω_- denote the underfitted case, which means $\Omega_- = \{\lambda : 0 < r_\lambda < r^*\}$. Let Ω_+ denote the overfitted case, which means $\Omega_+ = \{\lambda : r_\lambda > r^*\}$. Next, we prove the rank selection consistency of BIC_λ.

Theorem 2 *Assume technical conditions* (*C*1)-(*C*4). *Let* $\gamma \in (0, 1]$ *and the tuning parameter satisfy that* $\lambda_n n^{1/2+\gamma/2} \to \infty$ *and* $\lambda_n n^{1/2} \to 0$ *with* $n \to \infty$. *For any* $\lambda \in \Omega_+ \bigcup \Omega_-$,

$$\lim_{n\to\infty} P\left(\mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\right) = 1,$$

Proof For any $\lambda \in \Omega_+ \bigcup \Omega_-$,

$$\mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} = \underbrace{\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)^2\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left\langle X_i, \hat{B}(\lambda_n)\right\rangle\right)^2\right)}_{a}$$
$$\underbrace{+ \hat{\mathrm{df}}_\lambda \cdot \frac{\log(n)}{n} - \hat{\mathrm{df}}_{\lambda_n} \cdot \frac{\log(n)}{n}}_{b}.$$

For a, we know that

$$
\begin{aligned}
a &= \log\left(\frac{1}{n}\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda)\right)\right\|^2\right)-\log\left(\frac{1}{n}\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)\right)\right\|^2\right)\\
&= \log\left(\frac{\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}(\hat{B}(\lambda))\right\|^2/n}{\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)\right)\right\|^2/n}\right)\\
&= \log\left(\frac{\|\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)\right)+\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)-\hat{B}(\lambda)\right)\|^2/n}{\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)\right)\right\|^2/n}\right)\\
&= \log(1+a1+a2),
\end{aligned}
$$

where

$$
a1=\frac{\mathrm{vec}\left(\hat{B}(\lambda_n)-\hat{B}(\lambda)\right)^{\top}\frac{\mathscr{X}^{\top}\mathscr{X}}{n}\mathrm{vec}\left(\hat{B}(\lambda_n)-\hat{B}(\lambda)\right)}{\frac{1}{n}\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}(\hat{B}(\lambda_n))\right\|^2}
$$

and

$$
a2=\frac{2\left(\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)\right)\right)^{\top}\frac{\mathscr{X}}{n}\mathrm{vec}\left(\hat{B}(\lambda_n)-\hat{B}(\lambda)\right)}{\frac{1}{n}\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}(\hat{B}(\lambda_n))\right\|^2}.
$$

For any x, we have $\log(1+x)\geq\min\{\log(2),0.5x\}$. So, $a\geq\min\{\log(2),0.5(a1+a2)\}$. Next, we compute these values. Remark 3 illustrates that $P\left(\hat{B}(\lambda_n)=B^*\right)\to 1$ when $n\to\infty$, which leads to

$$
\mathrm{vec}\left(\hat{B}(\lambda_n)-\hat{B}(\lambda)\right)^{\top}\frac{\mathscr{X}^{\top}\mathscr{X}}{n}\mathrm{vec}\left(\hat{B}(\lambda_n)-\hat{B}(\lambda)\right)\to vec\left(B^*-\hat{B}(\lambda)\right)^{\top}E\left(\mathrm{vec}(X)\mathrm{vec}(X)^{\top}\right)vec\left(B^*-\hat{B}(\lambda)\right).
$$

Based on the fact that $\mathbf{y}=\mathscr{X}\mathrm{vec}(B^*)+\varepsilon$ and the condition (C4),

$$
\frac{\left\|\mathbf{y}-\mathscr{X}\mathrm{vec}\left(\hat{B}(\lambda_n)\right)\right\|^2}{n}=\frac{\left\|\mathscr{X}\mathrm{vec}\left(B^*-\hat{B}(\lambda_n)\right)\right\|^2}{n}+\frac{2}{n}\varepsilon^{\top}\mathscr{X}\mathrm{vec}\left(B^*-\hat{B}(\lambda_n)\right)+\frac{\varepsilon^{\top}\varepsilon}{n}\to\sigma^2.
$$

Combing these facts and (C3), we have $P\left(a1>\kappa\left\|B^*-\hat{B}(\lambda)\right\|_F^2/\sigma^2\right)\to 1$ and $a2\to 0$ in probability, which leads to

$$
P\left(a\geq\min\left\{\log(2),\frac{\kappa}{2\sigma^2}\left\|B^*-\hat{B}(\lambda)\right\|_F^2\right\}\right)\to 1\ with\ n\to\infty.
$$

For b, we consider the bound of $\hat{\mathrm{df}}_{\lambda}$. By Cauchy-Schwartz inequality, it is sure that

$$\hat{\mathrm{df}}_\lambda = \frac{1}{n}\sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)^\top M_r^+ \mathrm{vec}(X_k) = \left\langle \frac{1}{n}\sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)\mathrm{vec}\big(X_k\big)^\top, M_r^+ \right\rangle$$
$$\leq \left\| \frac{1}{n}\sum_{k=1}^{n} \mathrm{vec}\big(X_k\big)\mathrm{vec}\big(X_k\big)^\top \right\|_F \cdot \left\| M_r^+ \right\|_2 \leq \frac{1}{n}\sum_{k=1}^{n} \left\| X_k \right\|_F^2 \cdot \left\| M_r^+ \right\|_2.$$

When $n \to \infty$, $\frac{1}{n}\sum_{k=1}^{n} \left\| X_k \right\|_F^2 \to E(\|X\|_F^2)$, which is bounded because of (C1). In addition, $\left\| M_r^+ \right\|_2 = \frac{1}{\sigma_{\tilde{r}}(M_r)}$ with $\tilde{r}$ being the rank of M_r, which is bounded when $n \to \infty$. Combining the fact that $\frac{\log(n)}{n} \to 0$, a directly result is $P(\log(2) + b > 0) \to 1$ and $P\left(\frac{\kappa}{2\sigma^2} \left\| B^* - \hat{B}(\lambda) \right\|_F^2 + b > 0 \right) \to 1$.

Therefore, the result about $P\big(\mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} > 0\big) \to 1$ holds, when $n \to \infty$. □

According to this theorem, the information criterion BIC_λ will not select the underfitted or overfitted models when the sample size increasing, which means BIC_λ can select the true model consistently. As one can see, this result holds on the condition that the parameter p_1 and p_2 are fixed, because the rank consistency of (1) proved in Bach (2008) and our proof process require this assumption.

Remark 4 Following the same proof pattern, the Akaike information criterion (AIC) defined as $\mathrm{AIC}_\lambda = \log\left(\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \left\langle X_i, \hat{B}(\lambda) \right\rangle \right)^2 \right) + \hat{\mathrm{df}}_\lambda \cdot \frac{2}{n}$ can also be proved the rank selection consistency. Therefore, both AIC and BIC can be used to selecting the tuning parameter of (1). This paper just use BIC as the main point to illustrate the contribution of the work.

4 Numerical experiments

This section will show some numerical results of our proposed Bayesian information criterion (BIC) for tuning parameter selection. To compare our method with the popular tuning parameter selection methods, we extend the corresponding expressions for the model (1) without further proofs, which includes Akaike Information Criterion (AIC), Akaike Information Criterion corrected (AICc) and cross validation. The difference between BIC, AIC and AICc is on the second term of their expressions, which are showed as follows.

AIC is defined as

$$\mathrm{AIC}_\lambda = \log\left(\sum_{i=1}^{n} \frac{\left(y_i - \left\langle X_i, \hat{B}(\lambda) \right\rangle \right)^2}{n} \right) + \hat{\mathrm{df}}_\lambda \cdot \frac{2}{n}.$$

AICc is defined as

$$\mathrm{AICc}_\lambda = \log\left(\sum_{i=1}^{n} \frac{\left(y_i - \left\langle X_i, \hat{B}(\lambda)\right\rangle\right)^2}{n}\right) + \hat{\mathrm{df}}_\lambda \cdot \frac{2}{n} + \frac{2\hat{\mathrm{df}}_\lambda \cdot (\hat{\mathrm{df}}_\lambda + 1)}{n - \hat{\mathrm{df}}_\lambda - 1}.$$

K-fold cross-validation is a popular method used to evaluate and compare the generalization ability to predict new data. This approach works by splitting the dataset into K smaller subsets (folds). One of these folds is used as a test set to validate the performance of the model, while the other $K-1$ subsets are used as training sets for the model. This process is repeated K times, and each time selects a different fold as the test set and the rest as the training set. Finally, the K test results are averaged or otherwise integrated to obtain an overall model performance evaluation. The usual choice of K is 5 and 10, which are 5-fold cross validation and 10-fold cross validation, respectively.

In the following numerical experiments, we compare our BIC with AIC, AICc, 5-fold cross validation and 10-fold cross validation. The comparison indexes include the selected best tuning parameter λ^*, the mean square error (MSE) on the predictor y, the rank of the selected solution r and the computational time. Here, the MSE is defined as $\|\hat{y} - y\|^2$, where $\hat{y}$ is the predicted response variable under the selected tuning parameter. Our numerical experiments are all implemented on MATLAB R2023b with AMD Ryzen 5 5600 H with Radeon Graphics 3.30 GHz and 16 G RAM.

4.1 Simulation data

As in Bach (2008), we simulate a lots of data sets with different values p_1, p_2 and n, where prediction matrixes follow the standard normal distribution/Gaussian distribution. Here, we select $p_1 \in \{15, 25, 35\}$, $p_2 \in \{30, 45\}$ and $n \in \{10^3, 10^2, 50\}$. We generate random i.i.d. data matrixes $P_i \in \mathbb{R}^{p_1}$ and $Q_i \in \mathbb{R}^{p_2}$, and we set a true solution $B^* \in \mathbb{R}^{p \times q}$ with rank 2. All elements of these matrix distribute with the standard normal distribution. Then, $y_i = \langle P_i Q_i^\top, B^* \rangle + \epsilon_i, i = 1, 2, \cdots, n$, where ϵ_i have i.i.d components with normal distributions with zero mean and 0.1 standard variance. To select the best tuning parameter, same as the setting in Bach (2008), we set the tuning parameter sequence as $\lambda_k = \mathrm{e}^{\log(\lambda_{\max}) + (k-1) \times \frac{\log(\lambda_{\min}) - \log(\lambda_{\max})}{100}}$ with $k = 1, 2, \cdots, 100$. The numerical results are reported in Tables 1 to 3.

Based on these results in Tables 1 to 3, there are some conclusions. (i) Comparing to the other tuning parameter methods, our proposed BIC and AIC will select the best tuning parameter with the smallest solution rank. Because the true rank of the solution is 2, BIC and AIC will tend to select the best tuning parameter with the same true solution rank. (ii) The computational time of BIC, AIC and AICc are same, and obviously smaller than that of 5-fold CV and 10-fold CV. Because BIC, AIC and AICc are different only on the second term, their computational time are same. Comparing to cross validation, our BIC and AIC surely reduce the computational time, because that our method does not need to solve the model many times as cross validation. (iii) BIC and AIC generally performs same with the other methods on MSE, and the MSE under BIC and AIC

Table 1 Simulation consistent results on prediction matrix dimension as $p_1 = 15$ and $p_2 = 45$. The sample size includes $10^3, 10^2$ and 50, respectively

sample size	method	λ^*	MSE	r	time (s)
$n = 10^3$	BIC	22.860	7.2380	**2**	**101.69**
	AIC	0.0600	0.0808	6	**101.69**
	AIC_c	0.0003	150.23	15	**101.69**
	5-fold CV	0.1361	0.0259	**2**	178.12
	10-fold CV	0.1361	0.0259	**2**	672.54
$n = 10^2$	BIC	0.8335	0.0154	**4**	**121.22**
	AIC	0.8335	0.0154	**4**	**121.22**
	AICc	0.0004	0.0002	15	**121.22**
	5-fold CV	0.0006	1.22×10^{-8}	14	344.23
	10-fold CV	0.0006	1.22×10^{-8}	14	5112.1
$n = 50$	BIC	6.8011	0.0154	**4**	**110.38**
	AIC	6.8011	0.0154	**4**	**110.38**
	AICc	0.0906	0.0002	**4**	**110.38**
	5-fold CV	0.0026	1.22×10^{-8}	10	344.38
	10-fold CV	0.0026	1.22×10^{-8}	10	484.55

are larger than cross validation in some special cases, which is a trade off of the computational effect and computational cost.Obviously, the product of p_1 and p_2 in Table 1 is smaller than that in Tables 2 and 3. It is well known that least square estimator is a solution of a equation system, where the product of p_1 and p_2 is the unknown variable number and the sample size is the equation number. If the unknown variable number is much larger than the equation number, the solution

Table 2 Simulation results on prediction matrix dimension as $p_1 = 25$ and $p_2 = 30$. The sample size includes $10^3, 10^2$ and 50, respectively

sample size	method	λ^*	MSE	r	time (s)
$n = 10^3$	BIC	0.1092	0.1210	**3**	**56.76**
	AIC	0.0218	0.0487	7	**56.76**
	AICc	0.0489	0.0569	4	**56.76**
	5-fold CV	0.0218	0.0487	7	128.60
	10-fold CV	0.0218	0.0487	7	266.08
$n = 10^2$	BIC	0.0176	0.0001	14	**89.55**
	AIC	0.0176	0.0001	14	**89.55**
	AICc	0.0596	0.0003	**10**	**89.55**
	5-fold CV	0.0176	0.0001	14	301.95
	10-fold CV	0.0176	0.0001	14	802.78
$n = 50$	BIC	0.6608	0.0629	**6**	**112.48**
	AIC	0.6608	0.0629	**6**	**112.48**
	AICc	0.0277	0.3611	12	**112.48**
	5-fold CV	0.0910	0.0292	7	259.31
	10-fold CV	0.0612	0.0279	9	498.99

Table 3 Simulation results on prediction matrix dimension as $p_1 = 35$ and $p_2 = 30$. The sample size includes 10^3, 10^2 and 50, respectively

sample size	method	λ^*	MSE	r	time (s)
$n = 10^3$	BIC	0.0089	0.0134	**13**	**190.15**
	AIC	0.0026	0.0050	22	**190.15**
	AICc	0.0059	0.0088	16	**190.15**
	5-fold CV	0.0007	0.0025	30	566.70
	10-fold CV	0.0007	0.0025	30	1113.4
$n = 10^2$	BIC	0.4744	0.2561	**10**	**178.11**
	AIC	0.4744	0.2561	**10**	**178.11**
	AICc	0.3883	0.3430	**10**	**178.11**
	5-fold CV	0.2602	0.1944	11	655.69
	10-fold CV	0.2602	0.1944	11	1125.49
$n = 50$	BIC	0.1312	0.4607	**7**	**90.17**
	AIC	0.1312	0.4607	**7**	**90.17**
	AICc	0.0153	1.3717	11	**90.17**
	5-fold CV	1.0588	58.339	**7**	260.10
	10-fold CV	1.0588	58.338	**7**	521.09

will be unstable or imprecise. Because that W_1 and W_2 are constructed based on the least square estimator, values of p_1 and p_2 affects their results and the result of degrees of freedom in Theorem 1. Therefore, our tuning parameter selection results rely on values of p_1 and p_2.

To verify the optimal tuning parameter satisfying the requirements of the Theorem 2, we set $\gamma = 1$ and simulated a dataset with $p_1 = 25, p_2 = 30$, where the sample size ranges in the set $\{1 \times 10^4, 1.5 \times 10^4, 2 \times 10^4, 2.5 \times 10^4, 3 \times 10^4, 3.5 \times 10^4\}$. In this dataset, we set the candidate sequence as $\lambda_k = p_1 p_2 \times n^{-0.1k+0.05}$, where $k = 0, 2, 4, \cdots, 20$. Here, we omit other comparisons and only report the selected optimal tuning parameters in Table 4. Actually, AIC and BIC select the same optimal tuning parameters under different n in this data set.

From results in Table 4, it is known that the sequence $\lambda^* n$ has an increasing tread, and $\lambda^* n^{1/2}$ has a decreasing tread with sample size n increasing. So, these results may meet requirements of Theorem 2.

Table 4 The optimal tuning parameter selection results of AIC and BIC, with $p_1 = 25, p_2 = 30$

n	λ^*	$\lambda^* n$	$\lambda^* n^{1/2}$
1.0×10^4	0.1189	1189	11.8900
1.5×10^4	0.0809	1214	9.9082
2.0×10^4	0.0615	1230	8.6974
2.5×10^4	0.0498	1245	7.8741
3.0×10^4	0.0419	1257	7.2573
3.5×10^4	0.0362	1267	6.7724

Table 5 COVID data set results with $p_1 = 41$, $p_2 = 30$ and $n = 138$

Method	λ^*	MSE	rank	time
BIC	0.1346	0.0231	**1**	**19.088**
AIC	0.1346	0.0231	**1**	**19.088**
AICc	0.1346	0.0231	**1**	**19.088**
5-fold CV	0.1346	0.0231	**1**	37.855
10-fold CV	0.1346	0.0231	**1**	74.117

Table 6 Bike sharing data set results with $p_1 = 24$, $p_2 = 6$ and $n = 731$.

Method	λ^*	MSE	rank	time
BIC	1.5074	0.5173	**1**	**1.6880**
AIC	0.3557	0.4911	2	**1.6880**
AICc	0.0011	0.4029	6	**1.6880**
5-fold CV	0.0011	0.4029	6	7.1090
10-fold CV	0.0011	0.4029	6	9.2680

4.2 Real data

This section applies the tuning parameter selection process of the model (1) on the COVID-19 data set and Bike data set, and compares BIC with AIC, AICc, 5-fold CV and 10-fold CV in Table 5 and Table 6. For every data set, we standardize the data set by columns, i.e., every number is subtracted by the column mean and divided by the column standard deviation. To select the best tuning parameter, we set the tuning parameter sequence as $\lambda_k = 0.618^k \lambda_{\max}$ with $k = 1, 2, \cdots, 20$ and $\lambda_{\max} = \left\| W_1^\top \cdots \left(\sum_{i=1}^n y_i X_i \right) W_2^\top \right\|_2$. Here, $\lambda_{\max}$ is the smallest tuning parameter such that the solution of (1) is zero.

The COVID-19 dataset (Li et al., 2021; Wahltinez et al., 2020) consists of daily measurements related to COVID-19 for 138 countries around the world. This data set records the newly confirmed case in the period June 13, 2020 to July 12, 2020. In addition, this data also includes the 41 COVID-19 related government policies in each day, i.e., school-closing, restrictions on gathering, stay-at-home requirement, income support and so on. Each of these policies may have several levels, for example, school closing includes no closing, recommend closing, require some closing (e.g. just high school) or require all closing, which varied during the 30-day period. Therefore, for every sample, its prediction matrix is $X \in \mathbb{R}^{41 \times 30}$ and response is the newly confirmed case. The sample size is $n = 138$.

The Bike sharing data set (Fanaee-T and Gama, 2014) includes matrixes data from January 1st, 2011 to December 31th, 2012. For every data during this period, the recorded data includes 6 weather conditions (including sunny, mist or others, temperature, apparent temperature, humidity and wind speed) every hour,

which makes the prediction matrix dimension as $X \in \mathbb{R}^{24\times 6}$ and the sample size $n = 731$. The response variable is the daily aggregated count of rented bikes.

On these real data sets, our proposed BIC and AIC perform better than AICc because that the rank under λ^* with them are smaller and their MSE values are similar. In addition, BIC and AIC outperforms 5-fold CV and 10-fold CV on computational time.

5 Conclusion

To select the best tuning parameter of the adaptive nuclear norm regularized trace regression model, we propose a Bayesian information criterion (BIC) in this paper. We first compute the unbiased estimator of the degrees of the freedom of the model, which is the basic element of the BIC. Under this estimator, we build up the BIC and prove its rank selection consistency, i.e., BIC will select the tuning parameter achieving the true solution and true solution rank in probability with the sample size increasing. To evaluate the performance of BIC, we compare it with AIC, AICc, 5-fold CV and 10-fold CV on some simulation data and real data sets. The numerical results show that BIC and AIC perform better than others.

References

Abbruzzo, A., Vujačić, I., Mineo, A. M., Wit, E. C. (2019). Selecting the tuning parameter in penalized gaussian graphical models. *Statistics and Computing, 29*(3), 559–569.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics, 22*(1), 203–217.

Bach, F. R. (2008). Consistency of trace norm minimization. *Journal of Machine Learning Research, 9*, 1019–1048.

Beck, A. (2017). *First-Order Methods in Optimization*. Philadelphia: Society for Industrial and Applied Mathematics.

Chetverikov, D., Liao, Z., Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics, 49*(3), 1300–1317.

Datta, A., Zou, H. (2020). A note on cross-validation for lasso under measurement errors. *Technometrics, 62*(4), 549–556.

Davenport, M. A., Romberg, J. (2016). An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing, 10*(4), 608–622.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association, 99*(467), 619–632.

Elsener, A., van de Geer, S. (2018). Robust low-rank matrix estimation. *The Annals of Statistics, 46*(6), 3481–3509.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Fan, J., Gong, W., Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics, 212*(1), 177–202.

Fan, Y., Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(3), 531–552.

Fanaee-T, H., Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence, 2*, 113–127.

Hamidi, N., Bayati, M. (2022). On low-rank trace regression under general sampling distribution. *Journal of Machine Learning Research, 23*(321), 1–49.

Hirose, K., Tateishi, S., Konishi, S. (2013). Tuning parameter selection in sparse regression modeling. *Computational Statistics and Data Analysis, 59*, 28–40.

Homrighausen D., McDonald D. J. (2013) The lasso, persistence, and cross-validation. *In: the International Conference on Machine Learning, 28*, 1031–1039.

Homrighausen, D., McDonald, D. J. (2017). Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica, 27*, 1017–1036.

Horn, R. A., Johnson, C. R. (2012). *Matrix Analysis*. New York: Cambridge University Press.

Kim, Y., Kwon, S., Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research, 13*(1), 1037–1057.

Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association, 115*(532), 1978–1997.

Li, M., Kong, L., Su, Z. (2021). Double fused lasso regularized regression with both matrix and vector valued predictors. *Electronic Journal of Statistics, 15*, 1909–1950.

Liu, H., Wang, L., Zhao, T. (2014). Multivariate regression with calibration. *Advances in Neural Information Processing Systems, 27*, 127–135.

Liu, Z., Vandenberghe, L. (2010). Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications, 31*(3), 1235–1256.

Lu, Z., Monteiro, R. D., Yuan, M. (2012). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming, 131*(1), 163–194.

Magnus, J. R., Neudecker, H. (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Oxford: John Wiley and Sons.

Mazumder, R., Hastie, T., Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research, 11*(11), 2287–2322.

Negahban, S., Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics, 39*(2), 1069–1097.

Noorossana, R., Eyvazian, M., Amiri, A., Mahmoud, M. A. (2010). Statistical monitoring of multivariate multiple linear regression profiles in phase i with calibration application. *Quality and Reliability Engineering International, 26*(3), 291–303.

Rockafellar, R. T. (2015). *Convex Analysis*. Princeton: Princeton University Press.

Rothman, A. J., Levina, E., Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics, 19*(4), 947–962.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics, 9*(6), 1135–1151.

Sun, W., Wang, J., Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research, 14*(1), 3419–3440.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58*(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(1), 91–108.

Vladimir, K., Karim, L., Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics, 39*(5), 2302–2329.

Wahltinez O., Murphy K., Brenner M., Lee M., Erlinger A., Daswani M., Yawalkar P., Ontiveros Z., Alcantara R., Cheung A., Nath C., Le P., Navarro P. O. (2020) COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2.

Wang, H., Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association, 102*(479), 1039–1048.

Wang, H., Li, R., Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika, 94*(3), 553–568.

Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*(3), 671–683.

Wang, T., Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis, 102*(7), 1141–1151.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications, 170*, 33–45.

Wei, Z., Lee, T. C. (2022). High-dimensional multi-task learning using multivariate regression and generalized fiducial inference. *Journal of Computational and Graphical Statistics, 31*, 1–15.
Wu, Y., Wang, L. (2020). A survey of tuning parameter selection for high-dimensional regression. *Annual Review of Statistics and Its Application, 7*, 209–226.
Yaguang, L., Yaohua, W., Baisuo, J. (2019). Consistent tuning parameter selection in high-dimensional group-penalized regression. *Science China Mathematics, 62*(4), 751–770.
Yuan, M. (2016). Degrees of freedom in low rank matrix estimation. *Science China Mathematics, 59*(12), 2485–2502.
Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49–67.
Yuan, M., Ekici, A., Lu, Z., Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(3), 329–346.
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.
Zhao B., Haldar J. P., Brinegar C., Liang Z. P. (2010) Low rank matrix recovery for real-time cardiac mri. *In: 2010 IEEE International Symposium on Biomedical Imaging: from Nano to Macro*, IEEE, 996–999
Zhao, J., Niu, L., Zhan, S. (2017). Trace regression model with simultaneously low rank and row (column) sparse parameter. *Computational Statistics and Data Analysis, 116*, 1–18.
Zhou, H., Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(2), 463–483.
Zhu, Y. (2020). A convex optimization formulation for multivariate regression. *Advances in Neural Information Processing Systems, 33*, 17652–17661.
Zou, C., Ke, Y., Zhang, W. (2022). Estimation of low rank high-dimensional multivariate linear models for multi-response data. *Journal of the American Statistical Association, 117*(538), 693–703.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.
Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.
Zou, H., Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics, 37*(4), 1733–1751.