

Group exponential penalized estimation in high-dimensional multivariate linear regression

Takumi Sekino* , Ryoya Oda* and Hirofumi Wakaki

Graduate School of Advanced Science and Engineering, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan

(Last Modified: October 25, 2025)

Abstract

This paper introduces a penalized least squares method using the group exponential (GEXP) penalty to simultaneously perform variable selection and estimation of the coefficient regression matrix in multivariate linear regression models intended to show the relationship between multiple response variables and multiple explanatory variables. The GEXP penalty is a continuous approximation of the ℓ_0 penalty, which, despite its simple form, avoids the drawbacks of ℓ_0 penalized estimation methods that arise from the discontinuity of the ℓ_0 penalty and thereby provides more stable estimation. We show that the estimator possesses the oracle properties under a high-dimensional asymptotic framework in which the sample size goes to infinity, and both the number of response variables and the number of explanatory variables can go to infinity. We also apply the EGCV criterion as a model selection criterion to choose the two tuning parameters required for the GEXP penalized estimation and show that the EGCV criterion maintains variable selection consistency under certain assumptions. Through numerical experiments, we confirm that the performance of the proposed method is comparable to existing methods.

Keywords: Multivariate linear regression; High-dimensional asymptotic framework; Penalized least squares method; Oracle properties; Variable selection; EGCV criterion; Coordinate descent method

1 Introduction

The multivariate linear regression model is a fundamental multivariate model describing the relationship between multiple response variables and multiple explanatory variables. As such, it is a topic covered in many books (e.g., [29, 31]). To define the model, let \mathbf{Y} be the $n \times p$ response matrix consisting of p variables, \mathbf{X} be the $n \times k$ explanatory matrix consisting of k variables, and n be the sample size. The multivariate linear regression model can then be defined as follows:

$$\mathbf{Y} = \mathbf{X}\Theta^* + \boldsymbol{\varepsilon}\boldsymbol{\Sigma}^{1/2}, \quad (1)$$

where Θ^* is a $k \times p$ true regression coefficients matrix, $\boldsymbol{\Sigma}$ is a $p \times p$ true unknown positive definite covariance matrix, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ is an $n \times p$ error matrix, for which $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed as an error term ε that satisfies $E[\varepsilon] = \mathbf{0}_p$, $\text{Cov}[\varepsilon] = \mathbf{I}_p$. Without loss of generality, we assume that \mathbf{Y} and \mathbf{X} are centered (i.e., $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}_p$ and $\mathbf{X}'\mathbf{1}_n = \mathbf{0}_k$ hold), where $\mathbf{0}_p$ is the p -dimensional vector of zeros, $\mathbf{1}_n$ is the n -dimensional vector of ones, and \mathbf{I}_p is the p -dimensional identity matrix. In multivariate linear regression models, it is often desirable to select the explanatory variables affecting the response variables from k available ones. This is determined by checking whether the regression coefficient vector corresponding to the j -th column of the explanatory matrix \mathbf{X} , that is, the j -th row of Θ^* , is the zero vector. Consequently, this type of variable selection in multivariate linear regression models is regarded as a variable selection problem with a group structure.

In an empirical context, it is not just selecting the explanatory variables for the model that is important; of equal importance is estimating the regression coefficients. To accomplish both tasks, penalized estimation methods offer an attractive approach. Among such methods, the most basic are penalized estimation methods using the ℓ_0 penalty, such as Mallows' C_p [21], AIC [1], and BIC [26]. These methods enable variable selection by imposing a penalty based on the number of non-zero regression coefficients, regardless of the magnitude of the coefficient values. However, penalized

*Corresponding authors.

E-mail address: hepokouki2go@gmail.com (Sekino, T.), ryoya-oda@hiroshima-u.ac.jp (Oda, R.)

estimation methods using the ℓ_0 penalty are known to have two major drawbacks. First, the discontinuity of the ℓ_0 penalty at the origin leads to instability in the resulting estimators [6]. Second, ℓ_0 penalized estimation is NP-hard, requiring a computation time that increases exponentially with the number of explanatory variables to find the optimal variable set [33]. In recent years, algorithms for penalized estimation methods that directly employ the ℓ_0 penalty under a given number of nonzero coefficients have been studied. In [3], the estimation methods have been formulated as a mixed integer optimization (MIO) problem, and an algorithm for deriving the approximate solution. In [36], an efficient iterative algorithm for the methods has been proposed. The proposed algorithm can produce solutions within a reasonable computational time. Furthermore, [8, 23] have proposed efficient algorithms for penalized estimation methods that directly employ the $\ell_{2,0}$ penalty, the ℓ_0 penalty for matrices.

As a way to alleviate the traditional drawbacks of using the ℓ_0 penalty, penalized estimation methods based on continuous penalties have also been proposed. A representative example is the ℓ_1 penalized estimation method, LASSO [30]. Due to the singularity of the ℓ_1 penalty at the origin, LASSO can estimate regression coefficients to be exactly zero. In [20], it is shown that under certain conditions, the LASSO estimator has asymptotic normality. However, the LASSO estimator is known not to simultaneously possess variable selection consistency, which is the property that the probability of selecting the true set of explanatory variables converges to 1. As mentioned in [13], it is one of the desired properties in sparse estimation methods, which can perform both variable selection and regression coefficients estimation simultaneously.

To address the shortcomings of LASSO, penalized estimation methods with improved penalties such as SCAD [13], adaptive LASSO [37], and MCP [35] have been proposed. The SCAD penalty and MCP are generally non-convex penalties constructed to reduce the bias of the estimator while maintaining the properties of the ℓ_1 penalty near the origin. Under certain conditions, the estimators based on these methods in (1) where $p = 1$ have been shown to possess the oracle properties [13], meaning the simultaneous achievement of both variable selection consistency and asymptotic normality [13, 35, 37]. While these penalties were applied to each element of the regression coefficients, penalized estimation methods using penalties that incorporate a group structure for explanatory variables have also been proposed. In [32], an algorithm was developed for the estimation method using the group SCAD (GSCAD) penalty, which employs the SCAD penalty with a group structure for the explanatory variables. In [18], an algorithm was constructed for the estimation method using the group MCP (GMCP), derived in a similar fashion using the MCP.

In recent years, as direct relaxations of the ℓ_0 penalized estimation approach, penalized estimation methods using the SELO [10] and exponential (EXP) penalties [33] have also been proposed. The SELO and EXP penalties are known as continuous approximations of the ℓ_0 penalty and have a simpler form than the SCAD penalty and MCP. For a different EXP-type penalty from [33], the group exponential LASSO was proposed in [4], with a group structure for the explanatory variables. In [10, 33], it was shown that for the model (1) with $p = 1$, the SELO and EXP estimators possess the oracle properties under the constraint $k/n \rightarrow 0$. Furthermore, a new criterion similar to the BIC criterion was proposed as a model selection criterion for selecting tuning parameters while maintaining the variable selection consistency of the estimators. [12] proposed an EXP-type penalty with a different form from that in [33], which is not singular at the origin. Although it is unable to directly estimate regression coefficients to be exactly zero, a two-step process was proposed to achieve this result, thereby avoiding the problem of the ℓ_0 penalty discontinuity. Importantly, its estimator was shown to possess the oracle properties.

For multivariate linear regression models with $p \geq 1$, penalized estimation methods that apply a penalty to each element of the regression coefficient matrix were addressed in [9, 28], and the asymptotic properties of their estimators were shown. In [28], it was shown that the estimators by several specific non-convex penalized methods satisfy the oracle properties under the constraints $p < n$ and $\log pk/n \rightarrow 0$. Similarly, in [9], it was shown that the estimators have the oracle properties under the constraint $p^3 k^3/n \rightarrow 0$. However, these studies do not address penalties with a group structure for the explanatory variables, nor do they cover a high-dimensional asymptotic framework where p exceeds n . Additionally, while the group exponential (GEXP) penalties have been addressed for linear and generalized linear regression models when $p = 1$ in [4], they have not been applied to multivariate linear regression models.

To fill this gap, this paper introduces the GEXP penalized estimation method based on the EXP penalty from [33] for multivariate linear regression models. We show that the GEXP estimator has the oracle properties under a high-dimensional asymptotic framework where both the number of response variables p and the number of explanatory variables k can tend to infinity. We also apply the EGCV criterion [24] for selecting the tuning parameters included in the GEXP penalty. We then show that the GEXP estimator using the tuning parameters selected by the EGCV criterion maintains variable selection consistency under certain conditions. Furthermore, the group coordinate descent algorithm is used as an estimation algorithm. Our paper makes two main contributions to the field:

- The GEXP penalized estimation method for multivariate linear regression models is firmly established through the provision of theoretical guarantees and the construction of an appropriate estimation algorithm. The GEXP penalty has a simpler form than the GSCAD penalty and GMCP but possesses comparable asymptotic properties. In a series of numerical experiments, we confirm that the GEXP penalized estimation method performs on par with methods that use the GSCAD penalty and GMCP. Thus, the GEXP penalized estimation approach can be positioned as an effective estimation method.
- The variable selection consistency of the GEXP estimator is guaranteed for the situation where p may diverge faster than n (i.e., $p/n \rightarrow \tau \in [0, \infty)$). This is a unique ultra-high-dimensional setting for multivariate linear regression models. In [9, 28], p is assumed to be less than n ; it is also assumed that the maximum eigenvalue of Σ is bounded in order to guarantee the oracle properties. However, this paper does not impose this boundedness

assumption, thus allowing the maximum eigenvalue of Σ to diverge (see condition (E)).

The remainder of the paper is organized as follows: In section 2, we introduce the penalized estimation method using the GEXP penalty in multivariate linear regression models. In section 3, we show that the GEXP estimator possesses the oracle properties. It is also shown that the GEXP estimator using the tuning parameters selected by the EGCV criterion ensures variable selection consistency under certain conditions. In section 4, we derive the update formulas when applying the coordinate descent method as the estimation algorithm. In section 5, we validate the estimation accuracy of the proposed method through numerical simulation results and a real data example and compare the accuracy of the method with that of other penalized estimation methods.

2 Estimation method with group exponential penalty

2.1 Group penalized estimation methods

We first introduce several group penalized estimation methods in multivariate linear regression models. In model (1), a penalized estimator for the regression coefficient matrix based on the residual sum of squares is generally defined by the $k \times p$ regression coefficient matrix Θ that minimizes the following objective function:

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\Theta\|^2 + \sum_{j=1}^k p_{\lambda,a}(\|\theta_j\|), \quad (2)$$

where $\|\mathbf{M}\|$ is the Frobenius norm of matrix \mathbf{M} , $\lambda, a (\geq 0)$ are tuning parameters, θ_j is the j -th row vector of Θ , and $p_{\lambda,a}(\|\theta\|)$ is a penalty function. In this paper, we address the problem of identifying the explanatory variables necessary for the model. Therefore, the penalty term is structured to account for a group structure related to the explanatory variables. Specifically, the sum of the penalty terms is a sum over the regression coefficient vectors corresponding to each explanatory variable rather than a sum for each individual regression coefficient. As examples of penalties in group penalized estimation methods in model (1) with $p = 1$, [18] applied the group SCAD penalty and group MCP defined as follows:

$$\text{GSCAD} : p_{\lambda,a_1}(\|\theta\|) = p_{\text{GS}}(\|\theta\|) = \begin{cases} \lambda\|\theta\| & \text{if } \|\theta\| \leq \lambda \\ -\frac{\|\theta\|^2 - 2a_1\lambda\|\theta\| + \lambda^2}{2(a_1 - 1)} & \text{if } \lambda < \|\theta\| \leq a_1\lambda, \\ \frac{(a_1 + 1)\lambda^2}{2} & \text{if } a_1\lambda < \|\theta\| \end{cases}, \quad (3)$$

$$\text{GMCP} : p_{\lambda,a_2}(\|\theta\|) = p_{\text{GM}}(\|\theta\|) = \begin{cases} \lambda\|\theta\| - \frac{\|\theta\|^2}{2a_2} & \text{if } \|\theta\| \leq a_2\lambda, \\ \frac{a_2\lambda^2}{2} & \text{if } \|\theta\| > a_2\lambda \end{cases}, \quad (4)$$

where $a_1 (> 2)$ and $a_2 (> 1)$ are tuning parameters. Both the GSCAD penalty (3) and the GMCP (4) have characteristics similar to the group LASSO (GLASSO) penalty [34]: $p_{\lambda,a}(\|\theta\|) = p_{\text{GL}}(\|\theta\|) = \lambda\|\theta\|$ near the origin. The penalties (3) and (4) also satisfy the three properties required for sparse estimation mentioned in [13]: unbiasedness, sparsity, and continuity. Furthermore, according to [5], the objective function (2) is strictly convex with respect to the row vectors of Θ when using either the GSCAD penalty (3), provided that $a_1 > 2$, or GMCP (4), provided that $a_2 > 1$.

2.2 GEXP penalty

The GEXP estimator for multivariate linear regression models is defined as the matrix Θ that minimizes the following objective function $Q_n(\Theta)$:

$$Q_n(\Theta) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\Theta\|^2 + \sum_{j=1}^k p_{\lambda,a}(\|\theta_j\|), \quad (5)$$

where $p_{\lambda,a}(\|\theta\|)$ is the GEXP penalty function defined by

$$p_{\lambda,a}(\|\theta\|) = p_{\text{GEXP}}(\|\theta\|) = \lambda \left\{ 1 - \exp\left(-\frac{\|\theta\|}{a}\right) \right\}, \quad (6)$$

in which $\lambda (\geq 0)$ and $a (> 0)$ are tuning parameters; λ controls the strength of the penalty term in (5), while a controls the shape of the GEXP penalty function, which has the same form as the EXP penalty in [33]. Figure 1 plots the shapes of five penalties for $\lambda = 0.5$: ℓ_0 , GEXP ($a = 0.02$), GSCAD ($a_1 = 3.7$), GMCP ($a_2 = 3$), and GLASSO. The ℓ_0 penalty is

defined as $\lambda I(\|\boldsymbol{\theta}\| \neq 0)$ using the indicator function $I(\cdot)$. Similar to the GSCAD penalty and GMCP, the GEXP penalty is a non-convex function that possesses the properties of (near) unbiasedness, sparsity, and continuity.

On the other hand, the GEXP penalty has several distinct features from the GSCAD penalty and GMCP. The GSCAD penalty $p_{\text{GS}}(\|\boldsymbol{\theta}\|)$ and the GMCP $p_{\text{GM}}(\|\boldsymbol{\theta}\|)$ can be considered approximations of the GLASSO penalty since they converge to $\lambda\|\boldsymbol{\theta}\|$ as $a_1 \rightarrow \infty$ and $a_2 \rightarrow \infty$, respectively, and can therefore be seen as approximations of the GLASSO penalty with respect to the behavior of a single parameter. However, the GEXP penalty $p_{\text{GEXP}}(\|\boldsymbol{\theta}\|)$ can be regarded as a direct approximation of the ℓ_0 penalty because in addition to approaching the value of the ℓ_0 penalty as $\|\boldsymbol{\theta}\|$ increases, it also satisfies the following property:

$$\lim_{a \rightarrow 0} p_{\text{GEXP}}(\|\boldsymbol{\theta}\|) = \lambda I(\|\boldsymbol{\theta}\| \neq 0).$$

Therefore, the tuning parameter a controls the degree of the ℓ_0 approximation. Furthermore, the GEXP penalty is sufficiently smooth everywhere except at the origin, as it is a C^∞ class function on $\mathbb{R}_{>0}$ with respect to $\|\boldsymbol{\theta}\|$. Therefore, penalized estimation methods using the GEXP penalty are expected to provide stable estimation as a continuous approximation of the ℓ_0 penalty. Here, (6) is similar to the following EXP penalty proposed in [4]:

$$p_{\lambda,a}(\|\boldsymbol{\theta}\|) = \frac{\lambda^2}{a} \left\{ 1 - \exp\left(-\frac{a\|\boldsymbol{\theta}\|}{\lambda}\right) \right\}. \quad (7)$$

Note that the roles of the respective tuning parameters in (6) and (7) are different. (7) converges to the ℓ_0 penalty when $\lambda a^{-1} \rightarrow 0$ and $\lambda^2 a^{-1} \rightarrow \tau_0 \in [0, \infty)$, but it converges to $\lambda\|\boldsymbol{\theta}\|$ when $a \rightarrow 0$. Therefore, similar to the GSCAD penalty and GMCP, (7) can be considered an approximation of the GLASSO penalty. In this paper, we use (6), which directly approximates the ℓ_0 penalty. However, by considering a parameter transformation, our results can also be applied to (7).

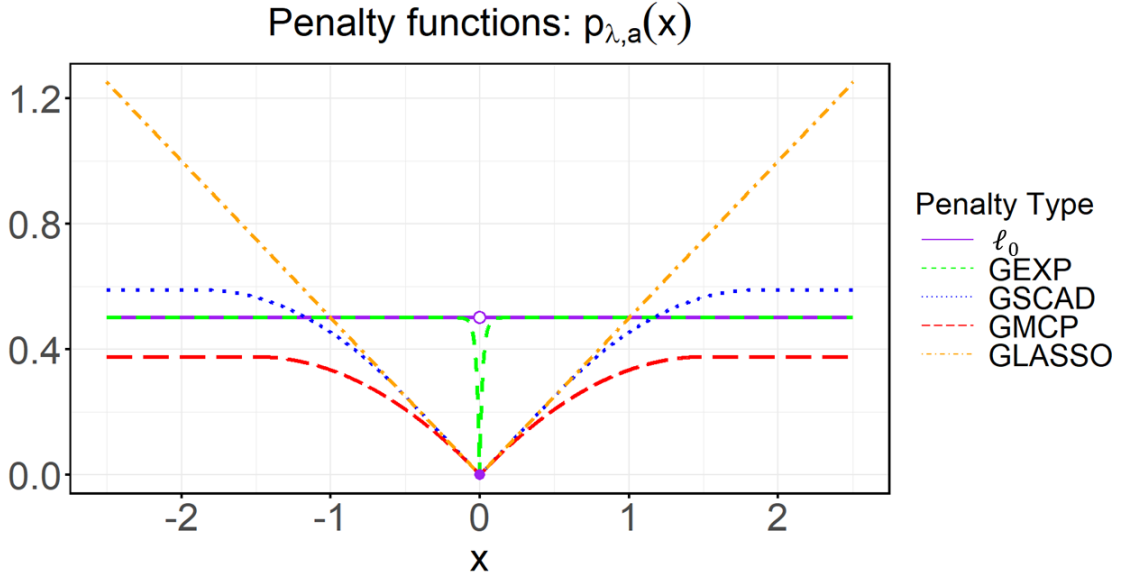


Figure 1: Shapes of various penalties for $\lambda = 0.5$: ℓ_0 , GEXP ($a = 0.02$), GSCAD ($a_1 = 3.7$), GMCP ($a_2 = 3$), and GLASSO

3 Theoretical properties of the GEXP estimator

In this section, we derive the asymptotic properties of the GEXP estimator under the high-dimensional asymptotic framework where both the number of explanatory variables k and the number of response variables p are allowed to tend to infinity as the sample size n increases. For the discussion below, we assume that k and p are sequences depending on n and introduce some notation. Let $\hat{\boldsymbol{\Theta}}$ be a local minimizer of $Q_n(\boldsymbol{\Theta})$, $\hat{\boldsymbol{\theta}}_j$ ($j = 1, \dots, k$) be the j -th row vector of $\hat{\boldsymbol{\Theta}}$, and $\boldsymbol{\theta}_j^*$ be the j -th row vector of the true regression coefficient matrix $\boldsymbol{\Theta}^*$. We define the set of indices of the true explanatory variables as $A = \{j \in \{1, \dots, k\} \mid \boldsymbol{\theta}_j^* \neq \mathbf{0}_p\}$ and its estimator as $\hat{A} = \{j \in \{1, \dots, k\} \mid \hat{\boldsymbol{\theta}}_j \neq \mathbf{0}_p\}$. The number of true explanatory variables is denoted by $k_0 = \#(A)$, where k_0 is a sequence depending on n , and $\hat{k}_0 = \#(\hat{A})$. For a subset $S \subseteq \{1, \dots, k\}$, the matrices \mathbf{X}_S , $\hat{\boldsymbol{\Theta}}'_S$, $\boldsymbol{\Theta}'_S$, $\boldsymbol{\Theta}'_S$ are defined as the submatrices of \mathbf{X} , $\hat{\boldsymbol{\Theta}}'$, $\boldsymbol{\Theta}'^*$, and $\boldsymbol{\Theta}'$, respectively, formed by selecting the columns corresponding to the indices in S . Also, $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the minimum and maximum eigenvalues of a square matrix \mathbf{M} , respectively. For a $p \times k$ matrix \mathbf{M} , its vectorization is $\text{vec}(\mathbf{M}) = (\mathbf{m}'_1, \dots, \mathbf{m}'_k)'$, where \mathbf{m}_j is the j -th column of \mathbf{M} .

3.1 Oracle properties

To show the variable selection consistency, one of the oracle properties of the GEXP estimator, we first present the following conditions (A)–(E):

- (A) $n \rightarrow \infty$, $p/n \rightarrow \tau \in [0, \infty]$, $k/n \rightarrow \nu \in [0, 1]$.
- (B) $n^{1/2}p^{-1/2}k^{-1/2}\rho \rightarrow \infty$, where $\rho = \min_{j \in A} \|\boldsymbol{\theta}_j^*\|$.
- (C) There exists a constant $c \in (0, 1)$ such that $\lambda = o(n^{-1/2}p^{1/2}k^{1/2}k_0^{-1}ae^{c\rho/a})$.
- (D) There exist constants $C_1, C_2 > 0$ such that $C_1 < \lambda_{\min}(n^{-1}\mathbf{X}'\mathbf{X}) \leq \lambda_{\max}(n^{-1}\mathbf{X}'\mathbf{X}) < C_2$.
- (E) $\text{tr}(\boldsymbol{\Sigma}) = O(p)$, $\liminf_{n \rightarrow \infty} \lambda_{\min}(\boldsymbol{\Sigma}) > 0$.

Condition (A) describes the high-dimensional asymptotic framework, allowing for p and k to be either fixed or to diverge to infinity. Specifically, it permits p to diverge faster than n and allows k_0 to diverge to infinity. Condition (B) is an assumption regarding the true regression coefficient vector corresponding to A and is a natural assumption in sparse estimation and variable selection problems. It allows ρ to converge to 0 depending on the divergence rates of p and k . Condition (C) is an assumption regarding the tuning parameters λ and a within the GEXP penalty. It controls the GEXP penalty to estimate some regression coefficients as zero while keeping the influence of the penalty term on $Q_n(\boldsymbol{\Theta})$ small. Condition (D) is a common assumption regarding the explanatory matrix in the asymptotic theory of regression analysis. Condition (E) is an assumption regarding the true covariance matrix $\boldsymbol{\Sigma}$. It allows the order of a finite number of variances to be $O(p)$ and ensures that $\boldsymbol{\Sigma}$ is positive definite asymptotically. Furthermore, condition (E) holds even when the maximum eigenvalue of $\boldsymbol{\Sigma}$ is of order $O(p)$, such as in a uniform correlation structure. This condition is somewhat similar to “the strongly spiked eigenvalue model” proposed in [2]. Under these conditions, as a prerequisite for deriving the variable selection consistency, the following Theorem 1 holds (the proof is given in Appendix 1).

Theorem 1. *Suppose that conditions (A)–(E) hold. Then, there exists a local minimizer $\widehat{\boldsymbol{\Theta}}$ of $Q_n(\boldsymbol{\Theta})$ that satisfies*

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\| = O_p(\sqrt{pk/n}).$$

Theorem 1 states the convergence rate of the GEXP estimator, which is the same as the least squares estimator, $\widehat{\boldsymbol{\Theta}}_{\text{LSE}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. In particular, it indicates that the GEXP estimator is consistent under the constraint $pk/n \rightarrow 0$. As an assumption necessary for deriving the variable selection consistency, we add condition (F) for the tuning parameters λ and a .

- (F) $n^{-1/2}p^{1/2}k^{1/2}a^{-1} = O(1)$, $n^{1/2}p^{-1/2}k^{-1/2}\lambda a^{-1} \rightarrow \infty$.

Condition (F) is a technical assumption for adjusting the degree of sparsity. The following Lemma 1 concerning sparsity holds (the proof is given in Appendix 2).

Lemma 1. *Suppose that conditions (A), and (D)–(F) hold. Then, for any constant $C > 0$, the following holds:*

$$\lim_{n \rightarrow \infty} P \left(\arg \min_{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\| \leq C\sqrt{pk/n}} Q_n(\boldsymbol{\Theta}) \subseteq \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{k \times p} \mid \boldsymbol{\Theta}_{A^c} = \mathbf{O}_{k-k_0, p} \right\} \right) = 1,$$

where $\mathbf{O}_{k-k_0, p}$ is the $(k - k_0) \times p$ zero matrix.

Lemma 1 means that the GEXP estimator has the property of correctly excluding truly unnecessary explanatory variables. By using Theorem 1 and Lemma 1, the variable selection consistency can be derived (the proof is given in Appendix 3).

Theorem 2. *Suppose that conditions (A)–(F) hold. Then, there exists a local minimizer $\widehat{\boldsymbol{\Theta}}$ of $Q_n(\boldsymbol{\Theta})$ such that $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\| = O_p(\sqrt{pk/n})$ and the following holds:*

$$\lim_{n \rightarrow \infty} P(\widehat{A} = A) = 1.$$

Theorem 2 indicates that the GEXP estimator not only correctly selects truly necessary explanatory variables but also correctly excludes truly unnecessary ones. In other words, the GEXP estimator possesses variable selection consistency.

Next, we derive the asymptotic normality of the non-zero regression coefficient matrix for the GEXP estimator. For this purpose, we add the following conditions (G)–(J):

- (G) $\lambda = o(n^{-1/2}k_0^{-1/2}ae^{\rho/a})$.
- (H) $\max_{1 \leq i \leq n} n^{-1} \sum_{j \in A} x_{ij}^2 = O(k_0 n^{-1})$, where x_{ij} is the (i, j) -th element of \mathbf{X} .
- (I) There exists a constant $\delta > 0$ such that $E[\|\boldsymbol{\varepsilon}\|^{2+\delta}] = O(p^{1+\delta/2})$.
- (J) For the δ in condition (I), $n^{-1}p^{1+2/\delta}k_0 = o(1)$.

Condition (G) relates to the tuning parameters λ and a . In the proof of asymptotic normality, it is used to suppress the influence of the penalty term so that the estimator converges to a multivariate normal distribution. Condition (I) concerns the distributional form of $\boldsymbol{\varepsilon}$, which is a relaxation of the assumption of a multivariate normal distribution. Conditions (H) and (J) are necessary assumptions for using the Lindeberg-Feller central limit theorem in the proof of asymptotic normality. Condition (H) is similar to the assumption in [10, 33], but our assumption involves only the true explanatory variables. Condition (J) controls the divergence speeds of p and k_0 , and it is worth noting that our asymptotic normality requires at least $pk_0/n = o(1)$. Under these conditions, Theorem 3 holds (the proof is given in Appendix 4).

Theorem 3. *Suppose that conditions (A)–(J) hold. Then, there exists a local minimizer $\widehat{\boldsymbol{\Theta}}$ of $Q_n(\boldsymbol{\Theta})$ such that $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\| = O_p(\sqrt{pk/n})$ and the following holds:*

$$\sqrt{n}\mathbf{B}_n \left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma}^{-1} \right)^{1/2} \left\{ \text{vec}(\widehat{\boldsymbol{\Theta}}'_A) - \text{vec}(\boldsymbol{\Theta}^*'_A) \right\} \xrightarrow{d} \mathcal{N}_q(\mathbf{0}_q, \mathbf{G}),$$

where \mathbf{B}_n and \mathbf{G} are any $q \times k_0$ and $q \times q$ positive definite matrices, respectively, satisfying $\mathbf{B}_n \mathbf{B}'_n \rightarrow \mathbf{G}$.

From Theorem 2 and Theorem 3, it is evident that under conditions (A)–(J), the GEXP estimator simultaneously possesses both variable selection consistency and asymptotic normality. This means it has the oracle properties.

3.2 Tuning parameter selection

The behavior of the GEXP estimator depends on the tuning parameters λ and a . Additionally, since there are countless pairs of tuning parameters (λ, a) that satisfy the oracle properties, it is necessary to select an appropriate pair from a set of candidate pairs. As our model selection criterion for choosing the tuning parameters, we employ the Extended Generalized Cross-Validation (EGCV) criterion [24], defined as follows:

$$\text{EGCV}(\widehat{\boldsymbol{\Theta}}(\lambda, a)) = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}(\lambda, a)\|^2}{n(1 - \widehat{k}_0 n^{-1})^\alpha},$$

where α is a positive constant. In this section, the GEXP estimator is denoted as $\widehat{\boldsymbol{\Theta}}(\lambda, a)$ to emphasize its dependence on the tuning parameter pair (λ, a) . The EGCV criterion is equivalent to the GCV criterion if $\alpha = 2$. Whereas a BIC-type criterion was used as the model selection criterion in [10, 33], we employ the EGCV criterion since we are dealing with the objective function $Q_n(\boldsymbol{\Theta})$ based on the residual sum of squares.

Variable selection consistency is a characteristic of the GEXP estimator's oracle properties that should be preserved. Therefore, we derive the condition for α under which the EXP estimator, selected by minimizing the EGCV criterion, maintains consistency in variable selection. To do so, we replace conditions (A) and (B) with conditions (A') and (B').

(A') $n \rightarrow \infty$, $p/n \rightarrow \tau \in [0, \infty]$, $k/n \rightarrow 0$.

(B') $\liminf_{n \rightarrow \infty} p^{-1/2} \rho > 0$.

Condition (A') is stricter regarding the rate of divergence of k than Condition (A), but it has also been assumed in [10, 33]. Condition (B') means that ρ diverges at a rate of at least \sqrt{p} as $p \rightarrow \infty$, which is a natural assumption since the row vectors of $\boldsymbol{\Theta}^*$ are p -dimensional. Under condition (A'), condition (B') is stricter than condition (B). Moreover, there is a trade-off between the leniency of condition (B') and the strictness of the condition on α required for consistency. However, condition (B') is necessary in this paper to ensure that the condition on α does not depend on unknown parameters. In this case, Theorem 4 holds (the proof is given in Appendix 5).

Theorem 4. *Suppose that conditions (A'), (B'), and (C)–(G) hold. Then, the following are true.*

- (i) *Suppose that conditions (I) and $\limsup_{n \rightarrow \infty} \lambda_{\max}(\boldsymbol{\Sigma}) < \infty$ hold. If, for the δ in condition (I), $p^{-2}k^{-2/(2+\delta)}\alpha \rightarrow \infty$ and $kn^{-1}\alpha \rightarrow 0$, then the following also holds:*

$$\lim_{n \rightarrow \infty} P \left(\text{EGCV}\{\widehat{\boldsymbol{\Theta}}(\lambda_n^*, a_n^*)\} < \text{EGCV}^- \right) = 1, \quad (8)$$

where $\text{EGCV}^- = \inf_{(\lambda, a) \in \Omega} \{\text{EGCV}\{\widehat{\boldsymbol{\Theta}}(\lambda, a)\} \mid \widehat{A} \neq A\}$, $\Omega \subseteq \mathbb{R}^2$ is a finite subset containing the pair (λ_n^*, a_n^*) that satisfies conditions (C), (F), and (G), and $\widehat{\boldsymbol{\Theta}}(\lambda_n^*, a_n^*)$ is a local minimizer of $Q_n(\boldsymbol{\Theta})$ that possesses variable selection consistency.

- (ii) *Suppose that $\limsup_{n \rightarrow \infty} \kappa_4 \text{tr}(\boldsymbol{\Sigma})^{-2} < \infty$ holds, where $\kappa_4 = \mathbb{E}[\|\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}\|^4] - \text{tr}(\boldsymbol{\Sigma})^2 - 2\text{tr}(\boldsymbol{\Sigma}^2)$. If $k^{-1}\alpha \rightarrow \infty$ and $kn^{-1}\alpha \rightarrow 0$, then (8) also holds.*

The boundedness of the maximum eigenvalue of $\boldsymbol{\Sigma}$ in (i) of Theorem 4 is a somewhat restrictive condition in high-dimensional settings where p diverges. In contrast, the assumption $\limsup_{n \rightarrow \infty} \kappa_4 \text{tr}(\boldsymbol{\Sigma})^{-2} < \infty$ in (ii) is a condition on the distribution of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\Sigma}$ similar to that used in [22]. Like condition (I), it is a relaxation of the multivariate normal distribution assumption. Specifically, this assumption holds for many elliptical distributions where the 4th moment of

ε exists, and it also allows the maximum eigenvalue of Σ to diverge. Comparing the conditions on α in (i) and (ii) of Theorem 4, the condition in (ii) is stricter than the one in (i) when p is fixed. However, when p goes to infinity, the condition in (i) becomes stricter than that in (ii) as p diverges more rapidly. The condition in (ii) is valid for any rate of divergence of p .

Finally, we confirm the existence of (λ_n^*, a_n^*) in Theorem 4 under conditions (A') and (B'). Conditions (C), (F), and (G) might seem complex at first glance, but it is easy to provide examples of λ and a that satisfy them under Conditions (A') and (B'). Assuming that under these conditions, p and k are, at most, of polynomial order in n , we can see that $\lambda = pkn^{-1} \log n$ and $a = \sqrt{pk/n}(\log n)^{3/4}$ satisfy conditions (C), (F), and (G).

4 Estimation algorithm

Because the GEXP estimator cannot be obtained in closed form, an estimation algorithm must be used. In [4], coordinate descent (e.g., [15, 16]) is applied to each regression coefficient. However, since the penalty term here has a group structure as in (6), we apply the group coordinate descent (GCD) algorithm (e.g., [18]), which updates the regression coefficient vector for one explanatory variable at a time. As can be seen from Figure 1, the GEXP penalty is generally a non-convex function, so the objective function $Q_n(\Theta)$ may have multiple local minimizers. Therefore, in this paper, we consider the estimation algorithm under the assumption that the objective function $Q_n(\Theta)$ is strictly convex. First, we derive the strict convexity condition as follows (the proof is given in Appendix 6).

Proposition 1. *If $\lambda_{\min}(n^{-1}\mathbf{X}'\mathbf{X}) > \lambda a^{-2}$ holds, then $Q_n(\Theta)$ is strictly convex with respect to Θ . Moreover, suppose the following inequality holds for some $j \in \{1, \dots, k\}$:*

$$n^{-1}\|\mathbf{x}_{(j)}\|^2 > \lambda a^{-2}, \quad (9)$$

where $\mathbf{x}_{(j)}$ is the j -th column vector of \mathbf{X} . Then, $Q_n(\Theta)$ is strictly convex in the θ_j direction.

Under the strict convexity condition for Θ from Proposition 1, the theorems and lemmas derived in section 3 are all applicable to the global minimum. We derive the update rule for the GCD algorithm to obtain the GEXP estimator, assuming that (9) holds for all $j \in \{1, \dots, k\}$. The update rule for $\hat{\theta}_\ell$ given the current parameters $\hat{\theta}_j$ ($j \neq \ell$) in the GCD algorithm can be obtained as follows (the derivation is given in Appendix 7).

$$\hat{\theta}_\ell = \begin{cases} \mathbf{0}_p & \text{if } \|\mathbf{c}_\ell\| \leq \frac{n\lambda}{a} \\ \frac{x_0}{\|\mathbf{c}_\ell\|} \mathbf{c}_\ell & \text{if } \|\mathbf{c}_\ell\| > \frac{n\lambda}{a} \end{cases}, \quad (10)$$

where $\mathbf{c}_\ell = \mathbf{Y}'\mathbf{x}_{(\ell)} - \sum_{j \neq \ell} (\mathbf{x}'_{(\ell)}\mathbf{x}_{(j)})\hat{\theta}_j$, and x_0 is the unique positive solution to the following equation for x :

$$\|\mathbf{x}_{(\ell)}\|^2 x - \|\mathbf{c}_\ell\| + n\lambda a^{-1} e^{-x/a} = 0. \quad (11)$$

The update rule (10) makes it possible to accurately estimate the regression coefficient vector as zero. Although a numerical solution is required to find x_0 in (10) because (11) is a non-linear equation, there is only one positive solution, so the computational cost is low. Using (10), the estimation algorithm for GEXP penalized least squares with the GCD method is as shown in Algorithm 1.

The λ_L in Algorithm 1 is determined as follows. If we set $\tilde{\theta}_\ell = \mathbf{0}_p$ for all $\ell \in \{1, \dots, k\}$, then $\mathbf{c}_\ell = \mathbf{Y}'\mathbf{x}_{(\ell)}$. Therefore, the λ that estimates all regression coefficient vectors to be zero satisfies, from the update rule (10),

$$\|\mathbf{c}_\ell\| \leq \frac{n\lambda}{a} \Leftrightarrow \lambda \geq \frac{a\|\mathbf{Y}'\mathbf{x}_{(\ell)}\|}{n}.$$

Consequently, we can set

$$\lambda_L = \frac{a}{n} \max_{1 \leq \ell \leq k} \|\mathbf{Y}'\mathbf{x}_{(\ell)}\|.$$

5 Numerical experiments

In this section, we confirm the effectiveness of the GEXP estimator by using both simulations and real data. Algorithm 1 is applied as the estimation algorithm for finding the GEXP estimator, and the EGCV criterion is used for selecting the tuning parameters λ and a . As comparators for the GEXP estimator, we use the least squares estimator $\hat{\Theta}_{\text{LSE}}$ and the estimators produced by minimizing (2) with penalties, including the GLASSO penalty, the GSCAD penalty (3), and the GMCP (4). For ease of reference, the above estimators are referred to as GEXP, LSE, GLASSO, GSCAD, and GMCP, respectively. For the GSCAD and GMCP estimators, the tuning parameters are fixed at $a_1 = 3.7$ and $a_2 = 3$, respectively. The group coordinate descent algorithm by [18] is applied as the estimation algorithm for these comparators, and each tuning parameter λ is selected using the EGCV criterion.

Algorithm 1 Group coordinate descent algorithm for GEXP estimator

1. Input a grid of size R of increasing a values $\Gamma = \{a_1, \dots, a_R\}$, and a grid of size L of increasing λ values $\Lambda = \{\lambda_1, \lambda_2(a_r), \dots, \lambda_L(a_r)\}$. Define λ_L , such that $\widehat{\Theta}(\lambda_L, a_r) = \mathbf{O}_{k,p}$ and $\lambda_1 = 0$.
 2. For each value of $r \in \{1, \dots, R-1, R\}$ repeat the following:
 - (a) Initialize $\widetilde{\Theta} = (\widetilde{\theta}_1, \dots, \widetilde{\theta}_k)' = \widehat{\Theta}(\lambda_1, a_r)$, which means that $\widehat{\Theta}(\lambda_1, a_r)$ is the ordinary least squares estimator.
 - (b) For each value of $\ell \in \{2, \dots, L-1, L\}$ repeat the following:
 - (i) Cyclic coordinate descent, for $j = 1, 2, \dots, k$, calculate $\widehat{\theta}_j$ using (10).
 - (ii) Set $\widetilde{\theta}_j = \widehat{\theta}_j$.
 - (iii) Repeat steps (i) and (ii) until $\|\widehat{\Theta} - \widetilde{\Theta}\|/\|\widetilde{\Theta}\|$ is sufficiently small.
 - (iv) Obtain $\widehat{\Theta}(\lambda_\ell, a_r)$. Set $\widetilde{\Theta} = \widehat{\Theta}(\lambda_\ell, a_r)$.
 3. Return the two-dimensional solution surface $\widehat{\Theta}(\lambda, a)$, $(\lambda, a) \in \Lambda \times \Gamma$.
-

5.1 Simulation studies

In this subsection, we confirm the effectiveness of GEXP through simulation, with 10000 iterations for 15 combinations of (n, p) : $n = 50, 100, 300, 500, 1000$ and $p = 5, 100, 1000$. We set $k = \lceil 4n^{1/4} \rceil$ and $k_0 = \lceil k/2 \rceil$. The response matrix \mathbf{Y} was generated from $\mathcal{N}_{n \times p}(\mathbf{X}\Theta^*, \Sigma \otimes \mathbf{I}_n)$, and the matrices \mathbf{X} , Θ^* , and Σ were determined as follows. The explanatory matrix \mathbf{X} was the $n \times k$ matrix whose columns were scaled to have length \sqrt{n} , after being generated from $\mathcal{N}_{n \times k}(\mathbf{O}_{n,k}, \Phi \otimes \mathbf{I}_n)$, where the (i, j) -th element of Φ is $0.5^{|i-j|}$. The regression coefficient matrix was $\Theta^* = (\Theta_1^{*'}, \mathbf{O}'_{k-k_0, p})'$, where $\Theta_1^* \sim \mathcal{N}_{k_0 \times p}(\mathbf{O}_{k_0, p}, \mathbf{I}_p \otimes \mathbf{I}_{k_0})$. The covariance matrix Σ was given as $\Sigma = 0.4 \times (0.2\mathbf{I}_p + 0.8\mathbf{1}_p\mathbf{1}_p')$. Two values were set for α used in the EGCV criterion: $\alpha = \sqrt{n}$ and $\alpha = \sqrt{n}/\log \log n$. When $k = \lceil 4n^{1/4} \rceil$, $\alpha = \sqrt{n}$ represents a balanced case in which the condition on α in Theorem 4 (ii) holds, while $\alpha = \sqrt{n}/\log \log n$ is a slightly smaller value for which the condition on α also holds. The following three metrics are used to evaluate variable selection and estimation accuracy.

- The probability for selecting true explanatory variables: $P(\hat{A} = A)$.
- Mean Squared Error: $\text{MSE}(\widehat{\Theta}) = \text{E}[\|\widehat{\Theta} - \Theta^*\|^2]$.
- Prediction Mean Squared Error: $\text{PMSE}(\widehat{\Theta}) = \text{E}[\|\mathbf{X}\widehat{\Theta} - \mathbf{X}\Theta^*\|^2]/n$.

Values of $P(\hat{A} = A)$ close to 1 are desirable; smaller values of MSE and PMSE indicate better estimation and prediction accuracy, respectively.

The simulation results for the three metrics are shown in Tables 1, 2, and 3, respectively. For all combinations of p , n , and α , GEXP yielded better metric values than LSE and GLASSO. Furthermore, it can be observed that even when p is large, as the sample size increases, GEXP showed a tendency for $P(\hat{A} = A)$ to increase toward 1 and its MSE and PMSE values to decrease. A comparison with the existing methods, GSCAD and GMCP, is presented below. For all combinations of p and α , EXP exhibited behavior similar to GSCAD and GMCP: the probability $P(\hat{A} = A)$ increased toward 1, while both MSE and PMSE decreased. These results indicate that GEXP achieves a level of variable selection and estimation accuracy comparable to the existing methods.

Table 1: Probabilities for selecting true explanatory variables

p	n	$\alpha = \sqrt{n}$				$\alpha = \sqrt{n}/\log \log n$			
		GLASSO	GSCAD	GMCP	GEXP	GLASSO	GSCAD	GMCP	GEXP
5	50	0.4984	0.9483	0.9651	0.9797	0.3314	0.8538	0.9315	0.9271
	100	0.5661	0.9946	0.9974	0.9959	0.4582	0.9827	0.9754	0.9661
	300	0.6134	1.0000	1.0000	0.9999	0.2393	0.9987	0.9964	0.9954
	500	0.4751	0.9996	1.0000	1.0000	0.2684	0.9998	0.7304	0.9987
	1000	0.5297	1.0000	1.0000	1.0000	0.4871	1.0000	0.9999	1.0000
100	50	0.5991	0.9918	0.9852	0.9849	0.4084	0.9655	0.9554	0.9385
	100	0.5973	0.9990	0.9973	0.9980	0.3936	0.9882	0.9795	0.9711
	300	0.6780	1.0000	0.9999	1.0000	0.3114	0.9992	0.9982	0.9964
	500	0.5549	1.0000	1.0000	1.0000	0.3660	0.9996	0.9996	0.9981
	1000	0.5541	1.0000	1.0000	1.0000	0.3305	1.0000	1.0000	1.0000
1000	50	0.5582	0.9906	0.9862	0.9826	0.4409	0.9723	0.9492	0.9432
	100	0.5760	0.9987	0.9974	0.9974	0.3855	0.9879	0.9812	0.9727
	300	0.6683	0.9998	1.0000	0.9999	0.2882	0.9990	0.9971	0.9951
	500	0.5820	1.0000	1.0000	1.0000	0.3120	0.9997	0.9987	0.9991
	1000	0.5036	1.0000	1.0000	1.0000	0.2998	0.9999	0.9999	1.0000

Table 2: MSEs

p	n	LSE	$\alpha = \sqrt{n}$				$\alpha = \sqrt{n}/\log \log n$			
			GLASSO	GSCAD	GMCP	GEXP	GLASSO	GSCAD	GMCP	GEXP
5	50	0.8220	1.0816	0.3008	0.3491	0.3464	0.9445	0.4094	0.3966	0.4047
	100	0.4075	0.5676	0.2043	0.2188	0.1923	0.5166	0.2020	0.2025	0.2028
	300	0.1859	0.4699	0.0878	0.0836	0.0849	0.3291	0.0874	0.0889	0.0851
	500	0.1239	0.3146	0.0583	0.0568	0.0600	0.2467	0.0582	0.0941	0.0586
	1000	0.0730	0.2554	0.0375	0.0366	0.0357	0.1530	0.0353	0.0349	0.0359
100	50	14.5367	16.8955	5.9843	7.2755	6.0151	19.5368	7.7082	6.3961	7.9730
	100	8.6675	13.6134	4.5003	4.2043	3.8301	13.3858	4.2474	3.6057	4.3297
	300	3.5954	7.9655	1.6631	1.7717	1.6702	6.3307	1.6693	1.6943	1.6749
	500	2.4602	6.9696	1.1701	1.1439	1.1758	5.4946	1.1882	1.1304	1.1903
	1000	1.4915	5.0688	0.7216	0.7175	0.7310	3.7175	0.7143	0.7183	0.7143
1000	50	154.1758	223.0908	73.7176	67.9849	70.5429	149.3529	61.4005	71.4812	81.5849
	100	85.3633	161.7851	36.5428	43.2434	41.5250	111.2761	38.0660	36.3489	42.2788
	300	36.2551	95.1596	17.7791	17.1163	16.9260	59.5057	16.4614	17.0615	18.7825
	500	23.9648	72.8213	12.1102	11.0490	11.0490	49.5587	11.8191	11.7123	11.2966
	1000	14.6871	42.7924	6.9571	6.9571	6.9571	33.7853	7.3065	7.0330	6.9248

Table 3: PMSEs

p	n	LSE	$\alpha = \sqrt{n}$				$\alpha = \sqrt{n}/\log \log n$			
			GLASSO	GSCAD	GMCP	GEXP	GLASSO	GSCAD	GMCP	GEXP
5	50	0.4008	0.5431	0.2141	0.2114	0.2081	0.4471	0.2283	0.2172	0.2191
	100	0.2390	0.3574	0.1211	0.1203	0.1205	0.3015	0.1220	0.1225	0.1254
	300	0.1065	0.2248	0.0536	0.0537	0.0532	0.1693	0.0532	0.0539	0.0534
	500	0.0721	0.1630	0.0361	0.0360	0.0361	0.1249	0.0362	0.0522	0.0361
	1000	0.0441	0.1258	0.0221	0.0220	0.0220	0.0877	0.0220	0.0219	0.0220
100	50	7.9737	10.2147	4.0126	4.1020	4.0790	9.2601	4.1528	4.2072	4.3168
	100	4.7852	7.2838	2.4022	2.4129	2.3869	6.4900	2.4168	2.4950	2.4815
	300	2.1271	4.3339	1.0678	1.0675	1.0662	3.3472	1.0613	1.0678	1.0660
	500	1.4404	3.3877	0.7182	0.7153	0.7161	2.7063	0.7140	0.7182	0.7159
	1000	0.8809	2.4291	0.4411	0.4390	0.4340	1.8550	0.4399	0.4406	0.4399
1000	50	79.8671	109.2933	40.5890	40.7126	41.2022	89.8746	41.2091	42.5112	42.5894
	100	48.1503	76.5458	23.9724	24.1460	24.0891	62.4551	24.2020	24.5596	24.7606
	300	21.2890	47.1531	10.6526	10.6508	10.6417	32.8384	10.7201	10.7104	10.7715
	500	14.3591	34.2655	7.1740	7.1887	7.1887	24.9247	7.1745	7.1891	7.1678
	1000	8.7955	22.0416	4.3970	4.3970	4.3970	17.6975	4.3951	4.3673	4.3878

5.2 Real data analysis

In this subsection, the variable selection behavior and predictive accuracy of GEXP are examined using the yeast cell cycle gene expression data in [7, 9]. The response variables include 18 measurements of messenger ribonucleic acid (mRNA) levels sampled every 7 minutes over two cell cycles, spanning 119 minutes. The explanatory variables record binding information for 106 transcription factors (TFs). Genes with missing values in both experiments were excluded, resulting in a subset of 542 cell cycle-related genes. This yielded $p = 18$, $k = 106$, and $n = 542$. Both the response and explanatory variables were standardized. For the EGCV criterion, we set two values for α : $\alpha = \sqrt{n}$ and $\alpha = \sqrt{n}/\log \log n$. We calculated the number of selected explanatory variables, \hat{k}_0 , using all the data, and the PMSE values obtained by 5-fold cross-validation. The results are shown in Table 4.

Table 4: Number of selected explanatory variables and PMSEs from yeast cell cycle gene expression data

		LSE	GLASSO	GSCAD	GMCP	GEXP
$\alpha = \sqrt{n}$	\hat{k}_0	–	0	0	0	3
	PMSE	17.9072	18.1814	18.1814	17.7969	15.8244
$\alpha = \sqrt{n}/\log \log n$	\hat{k}_0	–	3	3	4	4
	PMSE	17.9072	17.9393	17.9393	15.9322	15.7296

As shown in Table 4, the PMSE for GEXP was smaller than the PMSEs of the LSE and the existing group penalized least squares estimators for both values of α . However, when $\alpha = \sqrt{n}$, the influence of the number of selected explanatory variables in the denominator of the EGCV criterion was excessive compared to the residual sum of squares in the numerator, so none of the three penalized estimators other than GEXP selected any variables. On the other hand, when $\alpha = \sqrt{n}/\log \log n$, the PMSE values for all penalties were smaller than in the case where $\alpha = \sqrt{n}$, and 3 or 4 variables were selected by each estimation method. Furthermore, at this time, GEXP had the smallest PMSE value. From these results, GEXP was confirmed to have a performance level equal to or better than the other estimators. Here, the appropriate selection of the parameter α in the EGCV criterion under a finite sample size is important; however, addressing this problem is beyond the scope of our current paper.

6 Conclusion

As detailed above, we applied a penalized estimation method using the group exponential penalty, a continuous approximation of the ℓ_0 penalty, to multivariate linear regression models. We derived the asymptotic properties of the estimator for the regression coefficient matrix and constructed an estimation algorithm using the group coordinate descent method. It was shown that the proposed estimator possesses the oracle properties, meaning it simultaneously achieves variable selection consistency and asymptotic normality. The EGCV criterion was adopted for selecting the tuning parameters necessary for estimation, and under certain conditions, it was proved that the GEXP estimator, with the parameters selected by the EGCV criterion, maintains variable selection consistency. Notably, the high-dimensional asymptotic framework used in this paper allows for p to diverge faster than n . Through simulation studies and real data analysis, we confirm that the GEXP estimator, despite its simple penalty function form, achieves performance comparable to penalized estimation methods using the GSCAD penalty and GMCP.

One area for future study is the effect of relaxing the strict convexity condition on the objective function in the GEXP penalized estimation method. Specifically, by considering a mixed penalty that combines the GEXP with the ridge penalty—similar to the Elastic Net [38] and adaptive elastic net [39]—it may be possible to further relax the strict convexity condition. Even when a mixed penalty with a ridge penalty is used, it has been shown that the estimator can still have variable selection consistency (Mnet [19]) and even the oracle properties (SCAD-Ridge [11]). Therefore, it is expected that our proposed method will also maintain the oracle properties when the penalty is replaced with a mixture that includes a ridge penalty. Another area for future research is deriving the asymptotic properties of the GEXP estimator in cases where $k > n$, a situation not addressed in this paper. In such high-dimensional settings, the objective function $Q_n(\Theta)$ might have multiple local minima, necessitating the development of alternative estimation algorithms. The coupling of the concave convex procedure (CCCP) algorithm (e.g., [9]) would likely be one effective option. Finally, while this paper focuses on applying the group exponential penalty to multivariate linear regression models, this penalty function could also be applied to various other multivariate models, such as the GMANOVA model [25] and discriminant analysis models [17].

Appendix 1: Proof of Theorem 1

Let $\beta_n = \sqrt{pk/n}$. First, we show that

$$\forall \varepsilon > 0, \exists C > 0 \text{ s.t. } \exists N \in \mathbb{N} \text{ s.t. } \forall n \geq N, P\left(Q_n(\Theta^*) < \inf_{\|\mathbf{M}\|=C} Q_n(\Theta^* + \beta_n \mathbf{M})\right) > 1 - \varepsilon. \quad (12)$$

Let $\varepsilon > 0$ be arbitrary. If we let $A_n = O_p(1)$, then

$$\exists K > 0 \text{ s.t. } \exists N_1 \in \mathbb{N} \text{ s.t. } \forall n \geq N_1, P(|A_n| < K) > 1 - \varepsilon \quad (13)$$

holds. Similarly, if we let $B_n = o(1)$, then for the same K as in (13),

$$\exists N_2 \in \mathbb{N} \text{ s.t. } \forall n \geq N_2, |B_n| < K$$

holds. Let $C = 6K/C_1 (> 0)$, and let $\mathbf{M} \in \mathbb{R}^{k \times p}$ be any matrix satisfying $\|\mathbf{M}\| = C$. Let $\boldsymbol{\mu}_j$ be the j -th row vector of \mathbf{M} , and let $S(\mathbf{M}) = \{j \in \{1, \dots, k\} \mid p_{\lambda,a}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|) - p_{\lambda,a}(\|\boldsymbol{\theta}_j^*\|) < 0\}$. In this case, for the c in condition (C), condition (B) implies that

$$\exists N_3 \in \mathbb{N} \text{ s.t. } \forall n \geq N_3, \forall j \in S(\mathbf{M}), \|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\| \geq cp$$

holds. Let $N = \max\{N_1, N_2, N_3\}$. Hereafter, we consider $n \geq N$. In this case,

$$\begin{aligned} D_n(\mathbf{M}) &= Q_n(\Theta^* + \beta_n \mathbf{M}) - Q_n(\Theta^*) \\ &= \frac{1}{2n} \left\{ \|\mathbf{Y} - \mathbf{X}(\Theta^* + \beta_n \mathbf{M})\|^2 - \|\mathbf{Y} - \mathbf{X}\Theta^*\|^2 \right\} + \sum_{j=1}^k \{p_{\lambda,a}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|) - p_{\lambda,a}(\|\boldsymbol{\theta}_j^*\|)\} \\ &= \frac{1}{2n} \left\{ \beta_n^2 \|\mathbf{X}\mathbf{M}\|^2 - 2\beta_n \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) \right\} + \sum_{j=1}^k \{p_{\lambda,a}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|) - p_{\lambda,a}(\|\boldsymbol{\theta}_j^*\|)\} \\ &\geq \frac{1}{2n} \left\{ \beta_n^2 \|\mathbf{X}\mathbf{M}\|^2 - 2\beta_n \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) \right\} + \sum_{j \in S(\mathbf{M})} \{p_{\lambda,a}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|) - p_{\lambda,a}(\|\boldsymbol{\theta}_j^*\|)\} \end{aligned} \quad (14)$$

holds. Since $p_{\lambda,a}(\|\boldsymbol{\theta}\|)$ is a concave and monotonically increasing function on $[0, \infty)$, for any $j \in S(\mathbf{M})$,

$$\begin{aligned} p_{\lambda,a}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|) - p_{\lambda,a}(\|\boldsymbol{\theta}_j^*\|) &\geq p_{\lambda,a}^{(1)}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|)(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\| - \|\boldsymbol{\theta}_j^*\|) \\ &\geq p_{\lambda,a}^{(1)}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|)(-\beta_n \|\boldsymbol{\mu}_j\|) \end{aligned}$$

$$= -\frac{\lambda\beta_n\|\boldsymbol{\mu}_j\|}{a}e^{-\|\boldsymbol{\theta}_j^*+\beta_n\boldsymbol{\mu}_j\|/a} \quad (15)$$

holds. Here, $p_{\lambda,a}^{(1)}(\|\boldsymbol{\theta}\|)$ is the first derivative of $p_{\lambda,a}(\|\boldsymbol{\theta}\|)$ with respect to $\|\boldsymbol{\theta}\|$. Therefore, from (14) and (15),

$$\begin{aligned} D_n(\mathbf{M}) &\geq \frac{1}{2n} \left\{ \beta_n^2 \|\mathbf{X}\mathbf{M}\|^2 - 2\beta_n \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) \right\} + \sum_{j \in S(\mathbf{M})} \left\{ p_{\lambda,a}(\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|) - p_{\lambda,a}(\|\boldsymbol{\theta}_j^*\|) \right\} \\ &\geq \frac{1}{2n} \left\{ \beta_n^2 \|\mathbf{X}\mathbf{M}\|^2 - 2\beta_n \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) \right\} - \lambda \sum_{j \in S(\mathbf{M})} \frac{\beta_n \|\boldsymbol{\mu}_j\|}{a} e^{-\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|/a} \\ &\geq \frac{1}{2n} \left\{ \beta_n^2 \|\mathbf{X}\mathbf{M}\|^2 - 2\beta_n \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) \right\} - \frac{\lambda C \beta_n}{a} \sum_{j \in S(\mathbf{M})} e^{-\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|/a} \\ &= \frac{\beta_n^2}{2n} \|\mathbf{X}\mathbf{M}\|^2 - \frac{\beta_n}{n} \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) - \frac{\lambda C \beta_n}{a} \sum_{j \in S(\mathbf{M})} e^{-\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|/a} \end{aligned} \quad (16)$$

is obtained. Here, from condition (D),

$$\frac{\beta_n^2 \|\mathbf{X}\mathbf{M}\|^2}{2n} = \frac{\beta_n^2 \text{tr}(\mathbf{M}' \mathbf{X}' \mathbf{X} \mathbf{M})}{2n} \geq \frac{C^2 \beta_n^2 \lambda_{\min}(n^{-1} \mathbf{X}' \mathbf{X})}{2} \geq \frac{C_1 C^2 \beta_n^2}{2} \quad (17)$$

holds. Also,

$$\frac{\beta_n}{n} \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) \leq \frac{\beta_n}{n} \|\mathbf{M}\| \cdot \|\mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2}\| = \frac{C \beta_n}{n} \|\mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2}\|$$

and from conditions (D) and (E),

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2}\|^2 \right] &= \mathbb{E} \left[\text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X} \mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2}) \right] \\ &= \text{tr}(\mathbf{X} \mathbf{X}' \mathbb{E}[\boldsymbol{\mathcal{E}} \boldsymbol{\Sigma} \boldsymbol{\mathcal{E}}']) \\ &= \text{tr}(\mathbf{X} \mathbf{X}' \text{tr}(\boldsymbol{\Sigma}) \mathbf{I}_n) \\ &= \text{tr}(\boldsymbol{\Sigma}) \text{tr}(\mathbf{X} \mathbf{X}') \\ &\leq \text{tr}(\boldsymbol{\Sigma}) \cdot nk \lambda_{\max} \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right) \\ &\leq \text{tr}(\boldsymbol{\Sigma}) \cdot nk C_2 = O(npk) = O(n^2 \beta_n^2). \end{aligned}$$

Since this is the case, by Markov's inequality, $\|\mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2}\|^2 = O(n^2 \beta_n^2)$. Therefore,

$$\frac{\beta_n}{n} \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) = O_p(C \beta_n^2) \quad (18)$$

holds. Thus, using the fact that $\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\| \geq c\rho$ as well as conditions (C) and (E), from (16), (17), and (18),

$$\begin{aligned} D_n(\mathbf{M}) &\geq \frac{\beta_n^2}{2n} \|\mathbf{X}\mathbf{M}\|^2 - \frac{\beta_n}{n} \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X}\mathbf{M}) - \frac{\lambda C \beta_n}{a} \sum_{j \in S(\mathbf{M})} e^{-\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|/a} \\ &\geq \frac{C_1 C^2 \beta_n^2}{2} + O_p(C \beta_n^2) - \frac{\lambda C \beta_n}{a} \sum_{j \in S(\mathbf{M})} e^{-\|\boldsymbol{\theta}_j^* + \beta_n \boldsymbol{\mu}_j\|/a} \\ &\geq \frac{C_1 C^2 \beta_n^2}{2} + O_p(C \beta_n^2) - \frac{\lambda C \beta_n}{a} \sum_{j \in S(\mathbf{M})} e^{-c\rho/a} \\ &\geq \frac{C_1 C^2 \beta_n^2}{2} + O_p(C \beta_n^2) - \frac{k_0 \lambda C \beta_n}{a} e^{-c\rho/a} \\ &= C \beta_n^2 \left\{ \frac{C_1 C}{2} + O_p(1) - \frac{k_0 \lambda}{a \beta_n} e^{-c\rho/a} \right\} \\ &= C \beta_n^2 \left\{ \frac{C_1 C}{2} + O_p(1) + o(1) \right\} \\ &\geq C \beta_n^2 \left(\frac{C_1 C}{2} - |A_n| - |B_n| \right) \end{aligned}$$

holds. Hence, if we let $T = \{D_n(\mathbf{M}) \geq C \beta_n^2 (2^{-1} C_1 C - |A_n| - |B_n|)\}$, we have that

$$\begin{aligned} 1 - \varepsilon &< P(\{|A_n| < K\} \cap \{|B_n| < K\} \cap (T \cup T^c)) \\ &\leq P\left(D_n(\mathbf{M}) \geq C \beta_n^2 \left(\frac{C_1 C}{2} - 2K\right)\right) \end{aligned}$$

$$\begin{aligned}
&= P\left(D_n(\mathbf{M}) \geq \frac{6K^2\beta_n^2}{C_1}\right) \\
&= P\left(\inf_{\|\mathbf{M}\|=C} D_n(\mathbf{M}) \geq \frac{6K^2\beta_n^2}{C_1}\right) \\
&\leq P\left(Q_n(\boldsymbol{\Theta}^*) < \inf_{\|\mathbf{M}\|=C} Q_n(\boldsymbol{\Theta}^* + \beta_n\mathbf{M})\right)
\end{aligned}$$

which proves (12).

Now, since a minimizer of $Q_n(\boldsymbol{\Theta})$ on the closed set $\{\boldsymbol{\Theta}^* + \beta_n\mathbf{M} \mid \|\mathbf{M}\| \leq C\}$ always exists, a local minimizer $\widehat{\boldsymbol{\Theta}}$ also exists. Therefore, if $Q_n(\boldsymbol{\Theta}^*) < \inf_{\|\mathbf{M}\|=C} Q_n(\boldsymbol{\Theta}^* + \beta_n\mathbf{M})$, then $\widehat{\boldsymbol{\Theta}}$ exists in the interior of the closed set, and thus

$$1 - \varepsilon < P\left(Q_n(\boldsymbol{\Theta}^*) < \inf_{\|\mathbf{M}\|=C} Q_n(\boldsymbol{\Theta}^* + \beta_n\mathbf{M})\right) \leq P\left(\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|}{\beta_n} < C\right)$$

is satisfied. As a result, $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\| = O_p(\beta_n)$ is satisfied, and Theorem 1 is proved. \square

Appendix 2: Proof of Lemma 1

Let $\beta_n = \sqrt{pk/n}$. Let any $\varepsilon > 0$ and $C > 0$ be given. If we set $A_n = O_p(1)$, then

$$\exists N_1 \in \mathbb{N} \text{ s.t. } \exists K > 0 \text{ s.t. } \forall n \geq N_1, P(|A_n| < K) > 1 - \varepsilon \quad (19)$$

holds. Furthermore, for the K in (19), by condition (F),

$$\exists N_2 \in \mathbb{N} \text{ s.t. } \forall n \geq N_2, \frac{\lambda}{a\beta_n} e^{-C\beta_n/a} - K > 0$$

holds. Let $N = \max\{N_1, N_2\}$. Consider any $\boldsymbol{\Theta} \in \mathbb{R}^{k \times p}$ that satisfies $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\| \leq C\beta_n$ and $\boldsymbol{\Theta}_{A^c} \neq \mathbf{O}_{k-k_0, p}$. We define $\tilde{\boldsymbol{\Theta}} \in \mathbb{R}^{k \times p}$ such that $\tilde{\boldsymbol{\Theta}}_{A^c} = \mathbf{O}_{k-k_0, p}$ and $\tilde{\boldsymbol{\Theta}}_A = \boldsymbol{\Theta}_A$. Then,

$$\begin{aligned}
D_n(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) &= Q_n(\boldsymbol{\Theta}) - Q_n(\tilde{\boldsymbol{\Theta}}) \\
&= \frac{1}{2n} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2 - \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}}\|^2 \right\} + \sum_{j=1}^k \left\{ p_{\lambda, a}(\|\boldsymbol{\theta}_j\|) - p_{\lambda, a}(\|\tilde{\boldsymbol{\theta}}_j\|) \right\} \\
&= \frac{1}{2n} \left\{ \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}} - \mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\|^2 - \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}}\|^2 \right\} + \sum_{j \in A^c} p_{\lambda, a}(\|\boldsymbol{\theta}_j\|) \\
&= \frac{1}{2n} \|\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\|^2 - \frac{1}{n} \text{tr} \left((\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}})' \mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}) \right) + \sum_{j \in A^c} p_{\lambda, a}(\|\boldsymbol{\theta}_j\|) \quad (20)
\end{aligned}$$

is obtained. Here, using condition (D) and the fact that $\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\| \leq \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|$, we have

$$\begin{aligned}
\frac{\|\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\|^2}{2n} &= \frac{\text{tr}((\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})' n^{-1} \mathbf{X}' \mathbf{X} (\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}))}{2} \\
&\leq \frac{\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|^2 \lambda_{\max}(n^{-1} \mathbf{X}' \mathbf{X})}{2} \\
&\leq \frac{\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\| \cdot \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\| C_2}{2} \\
&\leq \frac{C_2 C \beta_n \|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|}{2}.
\end{aligned}$$

Therefore,

$$\frac{1}{2n} \|\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\|^2 = O(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\| \beta_n) \quad (21)$$

holds. Also,

$$\begin{aligned}
\frac{1}{n} \text{tr}((\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}})' \mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) &= \frac{1}{n} \text{tr}(\{\mathbf{X}(\boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}}) + \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2}\}' \mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) \\
&= \frac{1}{n} \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) + \frac{1}{n} \text{tr}((\boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) \quad (22)
\end{aligned}$$

is obtained. Here, using conditions (D) and (E) as well as Markov's theorem, for any $D > 0$,

$$\begin{aligned}
P\left(\frac{\{\text{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}))\}^2}{n^2} \geq D\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|^2\beta_n^2\right) &\leq \frac{\mathbb{E}[\{\text{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}))\}^2]}{D\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|^2n^2\beta_n^2} \\
&= \frac{\text{tr}(\boldsymbol{\Sigma}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}))}{D\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|^2n^2\beta_n^2} \\
&\leq \frac{\lambda_{\max}(\boldsymbol{\Sigma})\lambda_{\max}(n^{-1}\mathbf{X}'\mathbf{X})}{Dn\beta_n^2} \\
&\leq \frac{C_2\lambda_{\max}(\boldsymbol{\Sigma})}{Dn\beta_n^2} \\
&= \frac{O(1)}{D}.
\end{aligned}$$

Thus, $n^{-1}\text{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) = O_p(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n)$. Furthermore, by condition (D),

$$\begin{aligned}
\frac{1}{n}\text{tr}((\boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) &\leq \frac{1}{\sqrt{n}}\|\mathbf{X}(\boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}})\| \cdot \frac{1}{\sqrt{n}}\|\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\| \\
&\leq \|\boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}}\|\lambda_{\max}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{1/2} \cdot \|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\lambda_{\max}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{1/2} \\
&\leq CC_2\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n.
\end{aligned}$$

Thus, $n^{-1}\text{tr}((\boldsymbol{\Theta}^* - \tilde{\boldsymbol{\Theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) = O(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n)$. Therefore, from (22),

$$\frac{1}{n}\text{tr}((\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}})'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})) = O_p(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n) + O(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n) = O_p(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n) \quad (23)$$

holds. Hence, from (20), (21), and (23),

$$\begin{aligned}
D_n(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) &= \frac{1}{2n}\|\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\|^2 - \frac{1}{n}\text{tr}\left((\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}})'\mathbf{X}(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})\right) + \sum_{j \in A^c} p_{\lambda,a}(\|\boldsymbol{\theta}_j\|) \\
&= O_p(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n) + \sum_{j \in A^c} p_{\lambda,a}(\|\boldsymbol{\theta}_j\|)
\end{aligned} \quad (24)$$

holds. Here, since $p_{\lambda,a}(\|\boldsymbol{\theta}\|)$ is a monotonically increasing and concave function on $[0, \infty)$,

$$\begin{aligned}
\sum_{j \in A^c} p_{\lambda,a}(\|\boldsymbol{\theta}_j\|) &\geq \sum_{j \in A^c} p_{\lambda,a}^{(1)}(\|\boldsymbol{\theta}_j\|)\|\boldsymbol{\theta}_j\| \\
&= \sum_{j \in A^c} \frac{\lambda}{a}e^{-\|\boldsymbol{\theta}_j\|/a}\|\boldsymbol{\theta}_j\| \\
&\geq \sum_{j \in A^c} \frac{\lambda}{a}e^{-C\beta_n/a}\|\boldsymbol{\theta}_j\| \\
&\geq \frac{\lambda}{a}e^{-C\beta_n/a}\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|.
\end{aligned} \quad (25)$$

Therefore, from (24) and (25),

$$\begin{aligned}
D_n(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) &= O_p(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n) + \sum_{j \in A^c} p_{\lambda,a}(\|\boldsymbol{\theta}_j\|) \\
&\geq O_p(\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n) + \frac{\lambda}{a}e^{-C\beta_n/a}\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\| \\
&= \|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n \left\{ O_p(1) + \frac{\lambda}{a\beta_n}e^{-C\beta_n/a} \right\} \\
&\geq \|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n \left(-|A_n| + \frac{\lambda}{a\beta_n}e^{-C\beta_n/a} \right)
\end{aligned}$$

holds. Let $T = \{D_n(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) \geq \|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n\{-|A_n| + \lambda(a\beta_n)^{-1}e^{-C\beta_n/a}\}\}$, and noting that $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\| \leq C\beta_n$, we have

$$\begin{aligned}
1 - \varepsilon &< P(\{|A_n| < K\} \cap (T \cup T^c)) \\
&\leq P\left(D_n(\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}}) \geq \|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|\beta_n \left(-K + \frac{\lambda}{a\beta_n}e^{-C\beta_n/a}\right)\right) \\
&\leq P(Q_n(\tilde{\boldsymbol{\Theta}}) < Q_n(\boldsymbol{\Theta})) \\
&\leq P\left(\arg \min_{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\| \leq C\beta_n} Q_n(\boldsymbol{\Theta}) \subseteq \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{k \times p} \mid \boldsymbol{\Theta}_{A^c} = \mathbf{O}_{k-k_0, p} \right\}\right).
\end{aligned}$$

Thus, Lemma 1 is proved. \square

Appendix 3: Proof of Theorem 2

Let $\beta_n = \sqrt{pk/n}$. Let any $\varepsilon > 0$ be given. From Theorem 1, there exists a local minimizer $\hat{\Theta}$ of $Q_n(\Theta)$ such that $\|\hat{\Theta} - \Theta^*\| = O_p(\beta_n)$, which satisfies the following:

$$\exists C > 0 \text{ s.t. } \exists N_1 \in \mathbb{N} \text{ s.t. } \forall n \geq N_1, P\left(\hat{\Theta} \in \arg \min_{\|\Theta - \Theta^*\| \leq C\beta_n} Q_n(\Theta)\right) > 1 - \frac{\varepsilon}{2}. \quad (26)$$

Furthermore, for the C in (26), the following holds from Lemma 1:

$$\exists N_2 \in \mathbb{N} \text{ s.t. } \forall n \geq N_2, P\left(\arg \min_{\|\Theta - \Theta^*\| \leq C\beta_n} Q_n(\Theta) \subseteq \left\{\Theta \in \mathbb{R}^{k \times p} \mid \Theta_{A^c} = \mathbf{O}_{k-k_0, p}\right\}\right) > 1 - \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$ and assume $n \geq N$ henceforth. Since

$$\begin{aligned} 1 - \varepsilon &< P\left(\left\{\hat{\Theta} \in \arg \min_{\|\Theta - \Theta^*\| \leq C\beta_n} Q_n(\Theta)\right\} \cap \left\{\arg \min_{\|\Theta - \Theta^*\| \leq C\beta_n} Q_n(\Theta) \subseteq \left\{\Theta \in \mathbb{R}^{k \times p} \mid \Theta_{A^c} = \mathbf{O}_{k-k_0, p}\right\}\right\}\right) \\ &\leq P\left(\hat{\Theta} \in \left\{\Theta \in \mathbb{R}^{k \times p} \mid \Theta_{A^c} = \mathbf{O}_{k-k_0, p}\right\}\right) \\ &= P\left(\hat{\Theta}_{A^c} = \mathbf{O}_{k-k_0, p}\right), \end{aligned}$$

it follows that

$$\lim_{n \rightarrow \infty} P\left(\hat{\Theta}_{A^c} = \mathbf{O}_{k-k_0, p}\right) = 1.$$

Next, we show that $\lim_{n \rightarrow \infty} P(\cap_{j \in A} \{\hat{\theta}_j \neq \mathbf{0}_p\}) = 1$. For this, it is sufficient to show that $\lim_{n \rightarrow \infty} P(\|\hat{\Theta}_A - \Theta_A^*\|^2 < \rho^2) = 1$. From $\|\hat{\Theta} - \Theta^*\| = O_p(\beta_n)$ and condition (B), we have

$$\begin{aligned} P\left(\|\hat{\Theta}_A - \Theta_A^*\|^2 < \rho^2\right) &= P\left(\frac{\|\hat{\Theta}_A - \Theta_A^*\|^2}{\beta_n^2} < \frac{\rho^2}{\beta_n^2}\right) \\ &\geq P\left(\left\{\frac{\|\hat{\Theta}_A - \Theta_A^*\|^2}{\beta_n^2} < C^2\right\} \cap \left\{C^2 \leq \frac{\rho^2}{\beta_n^2}\right\}\right) \\ &\geq P\left(\left\{\frac{\|\hat{\Theta} - \Theta^*\|^2}{\beta_n^2} < C^2\right\} \cap \left\{C^2 \leq \frac{\rho^2}{\beta_n^2}\right\}\right) \\ &\rightarrow 1. \end{aligned}$$

Therefore,

$$P(\hat{A} = A) \geq P\left(\left\{\hat{\Theta}_{A^c} = \mathbf{O}_{k-k_0, p}\right\} \cap \bigcap_{j \in A} \left\{\hat{\theta}_j \neq \mathbf{0}_p\right\}\right) \rightarrow 1,$$

which completes the proof of Theorem 2. □

Appendix 4: Proof of Theorem 3

Assume $\hat{A} = A$, the local minimizer $\hat{\Theta}$ of $Q_n(\Theta)$ satisfies the following equation:

$$\begin{aligned} \nabla_{\text{vec}(\Theta_A^*)} Q_n(\hat{\Theta}_A) = \mathbf{0}_{k_0 p} &\Leftrightarrow -\frac{1}{n}(\mathbf{X}_A \otimes \mathbf{I}_p)' \{\text{vec}(\mathbf{Y}') - (\mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\hat{\Theta}'_A)\} + \mathbf{p}_A^{(1)}(\hat{\Theta}_A) = \mathbf{0}_{k_0 p} \\ &\Leftrightarrow (\mathbf{X}_A \otimes \mathbf{I}_p)' \text{vec}(\mathbf{Y}') = (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\hat{\Theta}'_A) + n \mathbf{p}_A^{(1)}(\hat{\Theta}_A), \end{aligned} \quad (27)$$

where $\mathbf{p}_A^{(1)}(\Theta_A) = \lambda a^{-1}(e^{-\|\theta_j\|^a} \theta_j / \|\theta_j\|)_{j \in A}$. Therefore, from (27), we have

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}_A \Theta_A^* + \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2} \\ \Leftrightarrow \text{vec}(\mathbf{Y}') &= \text{vec}(\Theta_A^{*'} \mathbf{X}'_A) + \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}') \\ \Leftrightarrow \text{vec}(\mathbf{Y}') &= (\mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\Theta_A^{*'}) + \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}') \\ \Rightarrow (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\mathbf{Y}') &= (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\Theta_A^{*'}) + (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}') \\ \Leftrightarrow (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\mathbf{Y}') &= \text{vec}(\Theta_A^{*'}) + (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}') \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1} \{ (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\widehat{\boldsymbol{\Theta}}'_A) + n \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A) \} = \text{vec}(\boldsymbol{\Theta}_A^*) + (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}') \\
&\Leftrightarrow \text{vec}(\widehat{\boldsymbol{\Theta}}'_A) - \text{vec}(\boldsymbol{\Theta}_A^*) = (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}') - \left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p \right)^{-1} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sqrt{n} \mathbf{B}_n \left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma}^{-1} \right)^{1/2} \left\{ \text{vec}(\widehat{\boldsymbol{\Theta}}'_A) - \text{vec}(\boldsymbol{\Theta}_A^*) \right\} \\
&= \sqrt{n} \mathbf{B}_n \left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma}^{-1} \right)^{1/2} \left\{ (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1} (\mathbf{X}'_A \otimes \mathbf{I}_p) (\mathbf{I}_p \otimes \boldsymbol{\Sigma}^{1/2}) \text{vec}(\boldsymbol{\mathcal{E}}') - \left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p \right)^{-1} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A) \right\} \\
&= \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\mathcal{E}}') - n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1/2} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A) \tag{28}
\end{aligned}$$

holds. From condition (F), we have $(\rho - \min_{j \in A} \|\hat{\boldsymbol{\theta}}_j\|)/a = O_p(1)$. Using conditions (D) and (E), we have

$$\begin{aligned}
\frac{\|n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1/2} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A)\|^2}{nk_0 \lambda^2 a^{-2} e^{-2\rho/a}} &= \frac{n \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A)' (\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1/2} \mathbf{B}'_n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1/2} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A)}{k_0 \lambda^2 a^{-2} e^{-2\rho/a}} \\
&\leq \frac{n \text{tr}(\mathbf{B}_n \mathbf{B}'_n) \lambda_{\max}((\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1}) \|\mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A)\|^2}{k_0 \lambda^2 a^{-2} e^{-2\rho/a}} \\
&\leq \frac{n \text{tr}(\mathbf{B}_n \mathbf{B}'_n) \lambda_{\max}((\mathbf{X}'_A \mathbf{X}_A)^{-1}) \lambda_{\max}(\boldsymbol{\Sigma}^{-1}) \sum_{j \in A} e^{-2\|\hat{\boldsymbol{\theta}}_j\|/a}}{k_0 e^{-2\rho/a}} \\
&\leq \frac{\text{tr}(\mathbf{B}_n \mathbf{B}'_n)}{C_1} O(1) e^{2(\rho - \min_{j \in A} \|\hat{\boldsymbol{\theta}}_j\|)/a} \\
&= O_p(1)
\end{aligned}$$

which implies $\|n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1/2} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A)\| = O_p(n^{1/2} k_0^{1/2} \lambda a^{-1} e^{-\rho/a})$. Hence, by condition (G), we have

$$\left\| n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma})^{-1/2} \mathbf{p}_A^{(1)}(\widehat{\boldsymbol{\Theta}}_A) \right\| = o_p(1). \tag{29}$$

Therefore, from (28) and (29), we have

$$\sqrt{n} \mathbf{B}_n \left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \otimes \boldsymbol{\Sigma}^{-1} \right)^{1/2} \left\{ \text{vec}(\widehat{\boldsymbol{\Theta}}'_A) - \text{vec}(\boldsymbol{\Theta}_A^*) \right\} = \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\mathcal{E}}') + o_p(1).$$

Thus, it suffices to use the Lindeberg-Feller central limit theorem to show that

$$\mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\mathcal{E}}') \xrightarrow{d} \mathcal{N}_q(\mathbf{0}_q, \mathbf{G}).$$

Let $\mathbf{x}_{i,A}$ be the i -th row vector of \mathbf{X}_A , and let $\mathbf{w}_i = \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{x}_{i,A} \otimes \mathbf{I}_p) \boldsymbol{\varepsilon}_i$. Then,

$$\mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{X}'_A \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\mathcal{E}}') = \sum_{i=1}^n \mathbf{w}_i$$

and

$$\sum_{i=1}^n \text{Cov}[\mathbf{w}_i] = \mathbf{B}_n \mathbf{B}'_n$$

holds. Let $\mathbf{H}_i = (\mathbf{x}'_{i,A} \otimes \mathbf{I}_p) (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} \mathbf{B}'_n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{x}_{i,A} \otimes \mathbf{I}_p)$. Then,

$$\|\mathbf{w}_i\|^2 = \boldsymbol{\varepsilon}'_i (\mathbf{x}'_{i,A} \otimes \mathbf{I}_p) (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} \mathbf{B}'_n \mathbf{B}_n (\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1/2} (\mathbf{x}_{i,A} \otimes \mathbf{I}_p) \boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}'_i \mathbf{H}_i \boldsymbol{\varepsilon}_i$$

holds. Using Hölder's inequality and Markov's inequality, for any $\delta_0 > 0$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_i\|^2 I(\|\mathbf{w}_i\|^2 > \delta_0)] &= \mathbb{E}[\boldsymbol{\varepsilon}'_i \mathbf{H}_i \boldsymbol{\varepsilon}_i I(\boldsymbol{\varepsilon}'_i \mathbf{H}_i \boldsymbol{\varepsilon}_i > \delta_0)] \\
&\leq \lambda_{\max}(\mathbf{H}_i) \mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^2 I(\boldsymbol{\varepsilon}'_i \mathbf{H}_i \boldsymbol{\varepsilon}_i > \delta_0)] \\
&\leq \lambda_{\max}(\mathbf{H}_i) \mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^{2+\delta}]^{2/(2+\delta)} P(\boldsymbol{\varepsilon}'_i \mathbf{H}_i \boldsymbol{\varepsilon}_i > \delta_0)^{\delta/(2+\delta)} \\
&\leq \lambda_{\max}(\mathbf{H}_i) \mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^{2+\delta}]^{2/(2+\delta)} \left\{ \frac{q \lambda_{\max}(\mathbf{H}_i)}{\delta_0} \right\}^{\delta/(2+\delta)} \\
&= q^{\delta/(2+\delta)} \delta_0^{-\delta/(2+\delta)} \lambda_{\max}(\mathbf{H}_i)^{1+\delta/(2+\delta)} \mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^{2+\delta}]^{2/(2+\delta)}. \tag{30}
\end{aligned}$$

Also, from conditions (D) and (H), we have

$$\begin{aligned}
\max_{1 \leq i \leq n} \lambda_{\max}(\mathbf{H}_i) &\leq \lambda_{\max}(\mathbf{B}'_n \mathbf{B}_n) \max_{1 \leq i \leq n} \lambda_{\max}((\mathbf{x}'_{i,A} \otimes \mathbf{I}_p)(\mathbf{X}'_A \mathbf{X}_A \otimes \mathbf{I}_p)^{-1}(\mathbf{x}_{i,A} \otimes \mathbf{I}_p)) \\
&= \lambda_{\max}(\mathbf{B}'_n \mathbf{B}_n) \max_{1 \leq i \leq n} \lambda_{\max}(\mathbf{x}'_{i,A}(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{x}_{i,A} \otimes \mathbf{I}_p) \\
&\leq \lambda_{\max}(\mathbf{B}'_n \mathbf{B}_n) \lambda_{\max} \left(\left(\frac{1}{n} \mathbf{X}'_A \mathbf{X}_A \right)^{-1} \right) \frac{1}{n} \max_{1 \leq i \leq n} \mathbf{x}'_{i,A} \mathbf{x}_{i,A} \\
&\leq \lambda_{\max}(\mathbf{B}'_n \mathbf{B}_n) \frac{1}{C_1} \max_{1 \leq i \leq n} \frac{1}{n} \sum_{j \in A} x_{ij}^2 \\
&= O\left(\frac{k_0}{n}\right). \tag{31}
\end{aligned}$$

Therefore, noting that $\sum_{i=1}^n \text{tr}(\mathbf{H}_i) = \text{tr}(\mathbf{B}_n \mathbf{B}'_n) = O(1)$, and using (30) and (31), from conditions (I) and (J), we have

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}[\|\mathbf{w}_i\|^2 I(\|\mathbf{w}_i\|^2 > \delta_0)] &\leq q^{\delta/(2+\delta)} \delta_0^{-\delta/(2+\delta)} \sum_{i=1}^n \lambda_{\max}(\mathbf{H}_i)^{1+\delta/(2+\delta)} \mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^{2+\delta}]^{2/(2+\delta)} \\
&\leq q^{\delta/(2+\delta)} \delta_0^{-\delta/(2+\delta)} O(p) \max_{1 \leq i \leq n} \lambda_{\max}(\mathbf{H}_i)^{\delta/(2+\delta)} \sum_{i=1}^n \text{tr}(\mathbf{H}_i) \\
&= O(p) \cdot O\left(\frac{k_0}{n}\right) \\
&= o(1).
\end{aligned}$$

Thus, by the Lindeberg-Feller central limit theorem, Theorem 3 is proved. \square

Appendix 5: Proof of Theorem 4

Based on the monotonicity of $\log x$, it suffices to show the following:

$$\lim_{n \rightarrow \infty} P\left(\log \text{EGCV}\{\widehat{\boldsymbol{\Theta}}(\lambda_n^*, a_n^*)\} < \inf_{(\lambda, a) \in \Omega} \left\{ \log \text{EGCV}\{\widehat{\boldsymbol{\Theta}}(\lambda, a)\} \mid \hat{A} \neq A \right\}\right) = 1.$$

Let $(\lambda_0, a_0) \in \Omega$ ($\lambda_0 = 0$). Let $\widehat{\boldsymbol{\Theta}}_0 = \widehat{\boldsymbol{\Theta}}(\lambda_0, a_0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and let $\widehat{\boldsymbol{\Theta}}$ be a local minimizer of the objective function $Q_n(\boldsymbol{\Theta})$. Also, let $\widetilde{\boldsymbol{\Theta}}$ be the least squares estimator corresponding to \hat{A} , i.e., $\widetilde{\boldsymbol{\Theta}}_{A \setminus \hat{A}} = \mathbf{O}_{k_0 - \hat{k}_0, p}$ and $\widetilde{\boldsymbol{\Theta}}_{\hat{A}} = (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}})^{-1} \mathbf{X}'_{\hat{A}} \mathbf{Y}$. Since $\widehat{\boldsymbol{\Theta}}$ is a local minimizer of $Q_n(\boldsymbol{\Theta})$, using the fact that $\text{vec}(\widehat{\boldsymbol{\Theta}}'_{\hat{A}}) = (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p)^{-1} (\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\mathbf{Y}') - n(\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p)^{-1} \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}})$, we have

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}\|^2 &= \left\{ \text{vec}(\mathbf{Y}') - (\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\widehat{\boldsymbol{\Theta}}'_{\hat{A}}) \right\}' \left\{ \text{vec}(\mathbf{Y}') - (\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\widehat{\boldsymbol{\Theta}}'_{\hat{A}}) \right\} \\
&= \text{vec}(\mathbf{Y}')' \text{vec}(\mathbf{Y}') - 2 \text{vec}(\mathbf{Y}')' (\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\widehat{\boldsymbol{\Theta}}'_{\hat{A}}) + \text{vec}(\widehat{\boldsymbol{\Theta}}'_{\hat{A}})' (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\widehat{\boldsymbol{\Theta}}'_{\hat{A}}) \\
&= \|\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\Theta}}\|^2 + n^2 \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}})' \{ (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}})^{-1} \otimes \mathbf{I}_p \} \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}}) \\
&= \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^2 + \|\mathbf{X}(\widehat{\boldsymbol{\Theta}}_0 - \widetilde{\boldsymbol{\Theta}})\|^2 + n^2 \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}})' \{ (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}})^{-1} \otimes \mathbf{I}_p \} \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}}). \tag{32}
\end{aligned}$$

Now, we consider the case where $A \setminus \hat{A} \neq \emptyset$. From (32) and condition (D), we have

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^2 &= \|\mathbf{X}(\widehat{\boldsymbol{\Theta}}_0 - \widetilde{\boldsymbol{\Theta}})\|^2 + n^2 \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}})' \{ (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}})^{-1} \otimes \mathbf{I}_p \} \mathbf{p}_{\hat{A}}^{(1)}(\widehat{\boldsymbol{\Theta}}_{\hat{A}}) \\
&\geq nC_1 \|\widehat{\boldsymbol{\Theta}}_0 - \widetilde{\boldsymbol{\Theta}}\|^2 \\
&= nC_1 \left\{ \|\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*\|^2 - 2 \text{tr}((\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*)'(\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)) + \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|^2 \right\} \\
&= nC_1 \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|^2 \left\{ 1 - \frac{2 \text{tr}((\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*)'(\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*))}{\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|^2} + \frac{\|\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*\|^2}{\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|^2} \right\} \\
&\geq nC_1 \rho^2 \left\{ 1 - \frac{2 \text{tr}((\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*)'(\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*))}{\|\widetilde{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*\|^2} \right\} \\
&\geq nC_1 \rho^2 \left(1 - \frac{2\|\widehat{\boldsymbol{\Theta}}_0 - \boldsymbol{\Theta}^*\|}{\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|} \right)
\end{aligned}$$

$$\geq nC_1\rho^2 \left(1 - \frac{2\|\widehat{\Theta}_0 - \Theta^*\|}{\rho}\right). \quad (33)$$

Here, from condition (D), we have

$$\begin{aligned} \mathbb{E} \left[\|\widehat{\Theta}_0 - \Theta^*\|^2 \right] &= \mathbb{E} \left[\|(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\Sigma}^{1/2}\|^2 \right] \\ &= \mathbb{E} \left[\text{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\Sigma}^{1/2}) \right] \\ &= \text{tr}(\boldsymbol{\Sigma})\text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-2}\mathbf{X}') \\ &= \frac{\text{tr}(\boldsymbol{\Sigma})}{n} \text{tr} \left(\left(\frac{1}{n}\mathbf{X}'\mathbf{X} \right)^{-1} \right) \\ &\leq \frac{k\text{tr}(\boldsymbol{\Sigma})}{C_1n} \\ &= O \left(\frac{k\text{tr}(\boldsymbol{\Sigma})}{n} \right) \end{aligned}$$

and by Markov's inequality,

$$\|\widehat{\Theta}_0 - \Theta^*\| = O_p(n^{-1/2}k^{1/2}\text{tr}(\boldsymbol{\Sigma})^{1/2}). \quad (34)$$

Also, since

$$\mathbb{E} \left[\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2 \right] = \mathbb{E} \left[\text{tr}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}'\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\boldsymbol{\varepsilon}\boldsymbol{\Sigma}^{1/2}) \right] = (n-k)\text{tr}(\boldsymbol{\Sigma})$$

and by Markov's inequality,

$$\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2 = O_p(n\text{tr}(\boldsymbol{\Sigma})). \quad (35)$$

Therefore, for any $x > 0$, using $\log x \geq 1 - x^{-1}$ and (33), (34), and (35), along with conditions (A'), (B'), (E) and $n^{-1}k\alpha \rightarrow 0$, we have

$$\begin{aligned} &\log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2}{n(1 - \hat{k}_0n^{-1})^\alpha} - \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2}{n(1 - kn^{-1})^\alpha} \\ &= \log \frac{(1 - kn^{-1})^\alpha \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2}{(1 - \hat{k}_0n^{-1})^\alpha \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2} \\ &\geq 1 - \frac{(1 - \hat{k}_0n^{-1})^\alpha \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2}{(1 - kn^{-1})^\alpha \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2} \\ &= \frac{1}{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2} \left\{ \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2 - \left(\frac{n - \hat{k}_0}{n - k} \right)^\alpha \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2 \right\} \\ &= \frac{1}{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2} \left[nC_1\rho^2 \left(1 - \frac{2\|\widehat{\Theta}_0 - \Theta^*\|}{\rho} \right) + \left\{ 1 - \left(\frac{n - \hat{k}_0}{n - k} \right)^\alpha \right\} \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2 \right] \\ &= \frac{nC_1\rho^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2} \left[1 - \frac{2\|\widehat{\Theta}_0 - \Theta^*\|}{\rho} + \left\{ 1 - \left(\frac{n - \hat{k}_0}{n - k} \right)^\alpha \right\} \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2}{nC_1\rho^2} \right] \\ &= \frac{nC_1\rho^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2} \left[1 + O_p \left(\frac{k^{1/2}\text{tr}(\boldsymbol{\Sigma})^{1/2}}{n^{1/2}\rho} \right) + \left\{ 1 - \left(\frac{n - \hat{k}_0}{n - k} \right)^\alpha \right\} O_p \left(\frac{\text{tr}(\boldsymbol{\Sigma})}{\rho^2} \right) \right] \\ &= \frac{nC_1\rho^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2} \{1 + o_p(1)\}. \end{aligned}$$

Thus, in both cases (i) and (ii) of Theorem 4, we have

$$\lim_{n \rightarrow \infty} P \left(\inf\{\text{EGCV}(\widehat{\Theta}) \mid A \setminus \hat{A} \neq \emptyset\} > \text{EGCV}(\widehat{\Theta}_0) \right) = 1.$$

Therefore, to prove Theorem 4, it is sufficient to consider the case of an overfitted model and show

$$\lim_{n \rightarrow \infty} P \left(\inf\{\text{EGCV}(\widehat{\Theta}) \mid A \subsetneq \hat{A}\} > \text{EGCV}(\widehat{\Theta}^*) \right) = 1, \quad (36)$$

where $\widehat{\Theta}^*$ is a local minimizer of the objective function $Q_n(\boldsymbol{\Theta})$ that possesses variable selection consistency. Assume that $\{j \in \{1, \dots, k\} \mid \hat{\boldsymbol{\theta}}_j^* \neq \mathbf{0}_p\} = A$ and $A \subsetneq \hat{A}$. Noting that $\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2 \geq \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}^*\|^2 \geq \|\mathbf{Y} - \mathbf{X}\widehat{\Theta}_0\|^2$, we have

$$\log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2}{n(1 - \hat{k}_0n^{-1})^\alpha} - \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}^*\|^2}{n(1 - k_0n^{-1})^\alpha} = \log \frac{(1 - k_0n^{-1})^\alpha}{(1 - \hat{k}_0n^{-1})^\alpha} + \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}\|^2}{\|\mathbf{Y} - \mathbf{X}\widehat{\Theta}^*\|^2}$$

$$\begin{aligned}
&\geq \alpha \log \frac{n - k_0}{n - \hat{k}_0} + \log \frac{\|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2}{\|\mathbf{Y} - \mathbf{X}\hat{\Theta}^*\|^2} \\
&\geq \alpha \left(1 - \frac{n - \hat{k}_0}{n - k_0}\right) + \frac{\|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\Theta}^*\|^2}{\|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2} \\
&\geq \frac{\alpha(\hat{k}_0 - k_0)}{n - k_0} - \frac{\left|\|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\Theta}^*\|^2\right|}{\|\mathbf{Y} - \mathbf{X}\hat{\Theta}_0\|^2}.
\end{aligned} \tag{37}$$

From here, we first prove (36) for case (i) of Theorem 4.

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2 &= \left\{ \text{vec}(\mathbf{Y}') - (\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\tilde{\Theta}'_{\hat{A}}) \right\}' \left\{ \text{vec}(\mathbf{Y}') - (\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\tilde{\Theta}'_{\hat{A}}) \right\} \\
&= \left\{ \text{vec}(\mathbf{Y}') - \{\mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p\} \text{vec}(\mathbf{Y}') \right\}' \left\{ \text{vec}(\mathbf{Y}') - \{\mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p\} \text{vec}(\mathbf{Y}') \right\} \\
&= \text{vec}(\mathbf{Y}')' \{ \mathbf{I}_{np} - \mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p \} \text{vec}(\mathbf{Y}') \\
&= \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')' \{ \mathbf{I}_{np} - \mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')
\end{aligned} \tag{38}$$

holds. Since $\hat{\Theta}^*$ is a local minimizer of $Q_n(\Theta)$, by using the fact that $\text{vec}(\hat{\Theta}^*_{\hat{A}}) = (\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p)^{-1}(\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p) \text{vec}(\mathbf{Y}') - n(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p)^{-1}\mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}})$, we have

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\hat{\Theta}^*\|^2 &= \left\{ \text{vec}(\mathbf{Y}') - (\mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\hat{\Theta}^*_{\hat{A}}) \right\}' \left\{ \text{vec}(\mathbf{Y}') - (\mathbf{X}_A \otimes \mathbf{I}_p) \text{vec}(\hat{\Theta}^*_{\hat{A}}) \right\} \\
&= \text{vec}(\mathbf{Y}')' \{ \mathbf{I}_{np} - \mathbf{X}_A(\mathbf{X}'_A\mathbf{X}_A)^{-1}\mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\mathbf{Y}') + n^2 \mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}})' \{ (\mathbf{X}'_A\mathbf{X}_A)^{-1} \otimes \mathbf{I}_p \} \mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}}) \\
&= \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')' \{ \mathbf{I}_{np} - \mathbf{X}_A(\mathbf{X}'_A\mathbf{X}_A)^{-1}\mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}') + n^2 \mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}})' \{ (\mathbf{X}'_A\mathbf{X}_A)^{-1} \otimes \mathbf{I}_p \} \mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}}).
\end{aligned} \tag{39}$$

Therefore, using (38) and (39) and from conditions (D), (F), and (G), we have

$$\begin{aligned}
\left| \|\mathbf{Y} - \mathbf{X}\tilde{\Theta}\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\Theta}^*\|^2 \right| &\leq \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')' \{ \mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p - \mathbf{X}_A(\mathbf{X}'_A\mathbf{X}_A)^{-1}\mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}') \\
&\quad + n^2 \mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}})' \{ (\mathbf{X}'_A\mathbf{X}_A)^{-1} \otimes \mathbf{I}_p \} \mathbf{p}_A^{(1)}(\hat{\Theta}^*_{\hat{A}}) \\
&= \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')' \{ \mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p - \mathbf{X}_A(\mathbf{X}'_A\mathbf{X}_A)^{-1}\mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}') + o_p(1).
\end{aligned} \tag{40}$$

Here, let $\mathbf{Z} = \mathbf{X}_{\hat{A}} \otimes \mathbf{I}_p$, $\mathbf{Z}_1 = \mathbf{X}_A \otimes \mathbf{I}_p$, and $\mathbf{Z}_2 = \mathbf{X}_{\hat{A} \setminus A} \otimes \mathbf{I}_p$. Let $\mathbf{S} = \mathbf{Z}'_2 \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \mathbf{Z}_2$ be the Schur complement of $\mathbf{Z}'\mathbf{Z}$ with respect to $\mathbf{Z}'_1\mathbf{Z}_1$. Since \mathbf{S} is nonsingular,

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{pmatrix} (\mathbf{Z}'_1\mathbf{Z}_1)^{-1} + (\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1\mathbf{Z}_2\mathbf{S}^{-1}\mathbf{Z}'_2\mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1} & -(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1\mathbf{Z}_2\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{Z}'_2\mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1} & \mathbf{S}^{-1} \end{pmatrix}$$

can be represented. On the other hand,

$$\begin{aligned}
\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' &= \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 + \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1\mathbf{Z}_2\mathbf{S}^{-1}\mathbf{Z}'_2\mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \\
&\quad - \mathbf{Z}_2\mathbf{S}^{-1}\mathbf{Z}'_2\mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1\mathbf{Z}_2\mathbf{S}^{-1}\mathbf{Z}'_2 + \mathbf{Z}_2\mathbf{S}^{-1}\mathbf{Z}'_2.
\end{aligned}$$

Therefore,

$$\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 = \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \mathbf{Z}_2\mathbf{S}^{-1}\mathbf{Z}'_2 \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \}$$

holds. Thus, from condition (D),

$$\begin{aligned}
&\text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')' \{ \mathbf{X}_{\hat{A}}(\mathbf{X}'_{\hat{A}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p - \mathbf{X}_A(\mathbf{X}'_A\mathbf{X}_A)^{-1}\mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}') \\
&= \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')' \{ \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}') \\
&\leq \|\mathbf{Z}'_2 \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')\|^2 \lambda_{\max}(\mathbf{S}^{-1}) \\
&\leq \|\mathbf{Z}'_2 \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')\|^2 \lambda_{\max}((\mathbf{Z}'\mathbf{Z})^{-1}) \\
&\leq \|\mathbf{Z}'_2 \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')\|^2 \frac{1}{\lambda_{\min}(\mathbf{X}'\mathbf{X})} \\
&\leq \|\mathbf{Z}'_2 \{ \mathbf{I}_{np} - \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1 \} \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')\|^2 \frac{1}{nC_1} \\
&\leq \frac{1}{nC_1} \|(\mathbf{X}'_{\hat{A} \setminus A} \otimes \mathbf{I}_p) \mathbf{P}_A \text{vec}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\mathcal{E}}')\|^2
\end{aligned} \tag{41}$$

holds, where $\mathbf{P}_A = \mathbf{I}_{np} - \mathbf{X}_A(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \otimes \mathbf{I}_p$. Furthermore, from condition (D), using $\max_{1 \leq j \leq k} n^{-1} \|\mathbf{x}_{(j)}\|^2 < C_2$, we have

$$\begin{aligned}
& \frac{1}{n} \|(\mathbf{X}'_{\hat{A} \setminus A} \otimes \mathbf{I}_p) \mathbf{P}_A \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}')\|^2 \\
&= \frac{1}{n} \|(\mathbf{X}'_{\hat{A} \setminus A} \otimes \mathbf{I}_p) \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \text{vec}(\boldsymbol{\mathcal{E}}')\|^2 \\
&= \frac{1}{n} \sum_{j \in \hat{A} \setminus A} \sum_{p(j-1)+1 \leq \ell \leq pj} \left[\mathbf{h}'_{(\ell)} \{(\mathbf{I}_n - \mathbf{X}_A(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A) \otimes \boldsymbol{\Sigma}^{1/2}\} \text{vec}(\boldsymbol{\mathcal{E}}') \right]^2 \\
&= \frac{1}{n} \sum_{j \in \hat{A} \setminus A} \sum_{p(j-1)+1 \leq \ell \leq pj} \left\{ \text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)} \right\}^2 \\
&\leq \frac{1}{n} \sum_{j \in \hat{A} \setminus A} \sum_{p(j-1)+1 \leq \ell \leq pj} \frac{\|(\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\|^2}{\|\mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\|^2} \left\{ \text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)} \right\}^2 \\
&\leq \frac{1}{n} \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \|(\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\|^2 \sum_{j \in \hat{A} \setminus A} \sum_{p(j-1)+1 \leq \ell \leq pj} \left\{ \frac{\text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}}{\|\mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\|} \right\}^2 \\
&\leq \lambda_{\max}(\boldsymbol{\Sigma}) \max_{1 \leq \ell \leq k} \frac{1}{n} \|\mathbf{x}_{(\ell)}\|^2 \sum_{j \in \hat{A} \setminus A} \sum_{p(j-1)+1 \leq \ell \leq pj} \left\{ \frac{\text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}}{\|\mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\|} \right\}^2 \\
&\leq C_2 p(\hat{k}_0 - k_0) \lambda_{\max}(\boldsymbol{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \left\{ \frac{\text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}}{\|\mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\|} \right\}^2 \\
&= C_2 p(\hat{k}_0 - k_0) \lambda_{\max}(\boldsymbol{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \left\{ \text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{u}_\ell \right\}^2. \tag{42}
\end{aligned}$$

Here, $\mathbf{h}_{(\ell)}$ is the ℓ -column vector of $\mathbf{X} \otimes \mathbf{I}_p$, and $\mathbf{u}_\ell \in \mathbb{R}^{np}$ is defined by $\|\mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}\| \mathbf{u}_\ell = \mathbf{P}_A (\mathbf{I}_n \otimes \boldsymbol{\Sigma}^{1/2}) \mathbf{h}_{(\ell)}$. Consequently, from (40), (41), and (42),

$$\begin{aligned}
& \left| \|\mathbf{Y} - \mathbf{X} \tilde{\boldsymbol{\Theta}}\|^2 - \|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}^*\|^2 \right| \\
&\leq \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}')' \{ \mathbf{X}'_{\hat{A}} (\mathbf{X}'_{\hat{A}} \mathbf{X}_{\hat{A}})^{-1} \mathbf{X}'_{\hat{A}} \otimes \mathbf{I}_p - \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}') + o_p(1) \\
&\leq \frac{1}{nC_1} \|(\mathbf{X}'_{\hat{A} \setminus A} \otimes \mathbf{I}_p) \mathbf{P}_A \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}')\|^2 + o_p(1) \\
&\leq \frac{C_2}{C_1} p(\hat{k}_0 - k_0) \lambda_{\max}(\boldsymbol{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \left\{ \text{vec}(\boldsymbol{\mathcal{E}}')' \mathbf{u}_\ell \right\}^2 + o_p(1) \tag{43}
\end{aligned}$$

holds. Also,

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}_0\|^2 &= \text{tr} \left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \{ \mathbf{I}_n - \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \} \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2} \right) \\
&= \text{tr}(\boldsymbol{\mathcal{E}} \boldsymbol{\Sigma} \boldsymbol{\mathcal{E}}') - \text{tr} \left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2} \right)
\end{aligned}$$

and since $\text{E}[\text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2})] = k \text{tr}(\boldsymbol{\Sigma})$, from condition (A') and Markov's inequality, we have

$$\text{tr} \left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\mathcal{E}}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}^{1/2} \right) = \text{tr}(\boldsymbol{\Sigma})(n - k) o_p(1). \tag{44}$$

On the other hand, for any $\varepsilon > 0$, we represent $\boldsymbol{\Sigma} = \mathbf{H}' \boldsymbol{\Lambda} \mathbf{H}$ using an orthogonal matrix \mathbf{H} and a diagonal matrix $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ (λ_j is the eigenvalues of $\boldsymbol{\Sigma}$), and let $\mathbf{u}_i = (u_{i1}, \dots, u_{ip})' = \mathbf{H} \boldsymbol{\varepsilon}_i$. Then,

$$\begin{aligned}
P \left(\left| \frac{1}{n \text{tr}(\boldsymbol{\Sigma})} \sum_{i=1}^n \{ \boldsymbol{\varepsilon}'_i \boldsymbol{\Sigma} \boldsymbol{\varepsilon}_i - \text{tr}(\boldsymbol{\Sigma}) \} \right| > \varepsilon \right) &= P \left(\left| \frac{1}{n \text{tr}(\boldsymbol{\Sigma})} \sum_{i=1}^n \{ \boldsymbol{\varepsilon}'_i \mathbf{H}' \boldsymbol{\Lambda} \mathbf{H} \boldsymbol{\varepsilon}_i - \text{tr}(\boldsymbol{\Sigma}) \} \right| > \varepsilon \right) \\
&= P \left(\left| \frac{1}{n \text{tr}(\boldsymbol{\Sigma})} \sum_{i=1}^n \sum_{j=1}^p (\lambda_j u_{ij}^2 - \lambda_j) \right| > \varepsilon \right) \\
&= P \left(\left| \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p \frac{\lambda_j}{\text{tr}(\boldsymbol{\Sigma})} (u_{ij}^2 - 1) \right| > \varepsilon \right) \\
&\leq \frac{1}{(n\varepsilon)^{1+\delta_1/2}} \text{E} \left[\left| \sum_{i=1}^n \sum_{j=1}^p \frac{\lambda_j}{\text{tr}(\boldsymbol{\Sigma})} (u_{ij}^2 - 1) \right|^{1+\delta_1/2} \right],
\end{aligned}$$

holds, where δ_1 is a positive number such that if $\delta < 2$, then $\delta_1 < \delta$, and if $\delta \geq 2$, then $\delta_1 \in (0, 2)$. Here, we use Lemma A.1.

Lemma A.1 ([27]). Let X_1, \dots, X_n be independent random variables with $E[X_i] = 0$ ($i = 1, \dots, n$) and $E[|X_i|^\eta] < \infty$ ($i = 1, \dots, n$) for some $\eta \in (1, 2)$. There exists a constant $K_\eta > 0$ depending only on η such that

$$E \left[\left| \sum_{i=1}^n X_i \right|^\eta \right] \leq K_\eta \sum_{i=1}^n E[|X_i|^\eta].$$

Using Lemma A.1, from conditions (E) and (I), for a sufficiently large constant $M > 0$, we have

$$\begin{aligned} P \left(\left| \frac{1}{n \text{tr}(\mathbf{\Sigma})} \sum_{i=1}^n \{\boldsymbol{\varepsilon}'_i \mathbf{\Sigma} \boldsymbol{\varepsilon}_i - \text{tr}(\mathbf{\Sigma})\} \right| > \varepsilon \right) &\leq \frac{1}{(n\varepsilon)^{1+\delta_1/2}} E \left[\left| \sum_{i=1}^n \sum_{j=1}^p \frac{\lambda_j}{\text{tr}(\mathbf{\Sigma})} (u_{ij}^2 - 1) \right|^{1+\delta_1/2} \right] \\ &\leq \frac{K_{\delta_1}}{(n\varepsilon)^{1+\delta_1/2}} \sum_{i=1}^n E \left[\left| \sum_{j=1}^p \frac{\lambda_j}{\text{tr}(\mathbf{\Sigma})} (u_{ij}^2 - 1) \right|^{1+\delta_1/2} \right] \\ &\leq \frac{K_{\delta_1}}{(n\varepsilon)^{1+\delta_1/2}} \sum_{i=1}^n E \left[\left(\sum_{j=1}^p \frac{\lambda_j}{\text{tr}(\mathbf{\Sigma})} u_{ij}^2 + 1 \right)^{1+\delta_1/2} \right] \\ &\leq \frac{K_{\delta_1}}{(n\varepsilon)^{1+\delta_1/2}} \sum_{i=1}^n E \left[2^{1+\delta_1/2} \left\{ \left(\sum_{j=1}^p \frac{\lambda_j}{\text{tr}(\mathbf{\Sigma})} u_{ij}^2 \right)^{1+\delta_1/2} + 1 \right\} \right] \\ &\leq K_{\delta_1} \left(\frac{2}{n\varepsilon} \right)^{1+\delta_1/2} \sum_{i=1}^n \left\{ \left(\frac{\lambda_{\max}(\mathbf{\Sigma})}{\text{tr}(\mathbf{\Sigma})} \right)^{1+\delta_1/2} E \left[\|\mathbf{u}_i\|^{2+\delta_1} \right] + 1 \right\} \\ &\leq K_{\delta_1} \left(\frac{2}{n\varepsilon} \right)^{1+\delta_1/2} \sum_{i=1}^n \left\{ \left(\frac{M}{p\lambda_{\min}(\mathbf{\Sigma})} \right)^{1+\delta_1/2} E \left[\|\boldsymbol{\varepsilon}_i\|^{2+\delta_1} \right] + 1 \right\} \\ &= K_{\delta_1} \left(\frac{2}{n\varepsilon} \right)^{1+\delta_1/2} n \{O(1) + 1\} \\ &= o(1). \end{aligned}$$

Here, we used the fact that by the Lyapunov inequality, $E[\|\boldsymbol{\varepsilon}\|^{2+\delta_1}] \leq E[\|\boldsymbol{\varepsilon}\|^{2+\delta}]^{(2+\delta_1)/(2+\delta)} = O(p^{1+\delta_1/2})$. In this case, from condition (A'),

$$\begin{aligned} \frac{1}{\text{tr}(\mathbf{\Sigma})(n-k)} \sum_{i=1}^n \{\boldsymbol{\varepsilon}'_i \mathbf{\Sigma} \boldsymbol{\varepsilon}_i - \text{tr}(\mathbf{\Sigma})\} + \frac{k}{n-k} &= o_p(1) \Leftrightarrow \frac{1}{\text{tr}(\mathbf{\Sigma})(n-k)} \text{tr}(\boldsymbol{\varepsilon} \mathbf{\Sigma} \boldsymbol{\varepsilon}') - 1 = o_p(1) \\ &\Leftrightarrow \text{tr}(\boldsymbol{\varepsilon} \mathbf{\Sigma} \boldsymbol{\varepsilon}') = \text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}. \end{aligned} \quad (45)$$

Therefore, using (44) and (45), which gives $\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}_0\|^2 = \text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}$, and from (37) and (43), we have

$$\begin{aligned} &\log \frac{\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}\|^2}{n(1 - \hat{k}_0 n^{-1})^\alpha} - \log \frac{\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}^*\|^2}{n(1 - k_0 n^{-1})^\alpha} \\ &\geq \frac{\alpha(\hat{k}_0 - k_0)}{n - k_0} - \frac{\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}\|^2 - \|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}^*\|^2}{\|\mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\Theta}}_0\|^2} \\ &\geq \frac{\alpha(\hat{k}_0 - k_0)}{n - k_0} - \frac{C_2 C_1^{-1} p(\hat{k}_0 - k_0) \lambda_{\max}(\mathbf{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2 + o_p(1)}{\text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}} \\ &= \frac{\hat{k}_0 - k_0}{\text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}} \left[\frac{\alpha \text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}}{n - k_0} - \frac{C_2 p \lambda_{\max}(\mathbf{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1} \right] \\ &\quad - \frac{o_p(1)}{\text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\frac{\alpha \text{tr}(\mathbf{\Sigma})(n-k) \{1 + o_p(1)\}}{n - k_0} - \frac{C_2 p \lambda_{\max}(\mathbf{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1} \\ &= \alpha \text{tr}(\mathbf{\Sigma}) \left[1 + o_p(1) - \frac{C_2 p \lambda_{\max}(\mathbf{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1 \alpha \text{tr}(\mathbf{\Sigma})} \right]. \end{aligned}$$

Thus, it is sufficient to show that there exists a constant $d \in (0, 1)$ such that

$$\lim_{n \rightarrow \infty} P \left(\frac{C_2 p \lambda_{\max}(\mathbf{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1 \alpha \text{tr}(\mathbf{\Sigma})} \geq 1 - d \right) = 0.$$

Here, we note that the following Lemma A.2 holds (the proof is given in Appendix 8).

Lemma A.2. Let the p -dimensional random vectors $\varepsilon_1, \dots, \varepsilon_n$ be independent. Also, let $E[\varepsilon_i] = \mathbf{0}_p$ and for some constant $\delta > 0$, $E[\|\varepsilon_i\|^{2+\delta}] = O(p^{1+\delta/2})$. Then, the following holds:

$$\sup_{\|\mathbf{t}\|=1} E[|\boldsymbol{\varepsilon}'\mathbf{t}|^{2+\delta}] = O(p^{1+\delta/2}),$$

where $\boldsymbol{\varepsilon} = (\varepsilon'_1, \dots, \varepsilon'_n)'$.

Therefore, using Lemma A.2 from condition (I), and from Markov's inequality, condition (E), $p^{-2}k^{-2/(2+\delta)}\alpha \rightarrow \infty$ and $\limsup_{n \rightarrow \infty} \lambda_{\max}(\boldsymbol{\Sigma}) < \infty$, for any $d \in (0, 1)$, we have

$$\begin{aligned} & P\left(\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma}) \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})} \geq 1-d\right) \\ & \leq \sum_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} P\left(\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma}) \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})} \geq 1-d\right) \\ & = \sum_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} P\left(\left[\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma}) \{\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell\}^2}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})}\right]^{1+\delta/2} \geq (1-d)^{1+\delta/2}\right) \\ & \leq \sum_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} \left\{\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma})}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})(1-d)}\right\}^{1+\delta/2} E[|\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell|^{2+\delta}] \\ & \leq p(\hat{k}_0 - k_0) \left\{\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma})}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})(1-d)}\right\}^{1+\delta/2} \max_{p(j-1)+1 \leq \ell \leq pj, j \in \hat{A} \setminus A} E[|\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{u}_\ell|^{2+\delta}] \\ & \leq p(\hat{k}_0 - k_0) \left\{\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma})}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})(1-d)}\right\}^{1+\delta/2} \sup_{\|\mathbf{t}\|=1} E[|\text{vec}(\boldsymbol{\varepsilon}')' \mathbf{t}|^{2+\delta}] \\ & = p(\hat{k}_0 - k_0) \left\{\frac{C_2 p \lambda_{\max}(\boldsymbol{\Sigma})}{C_1 \alpha \text{tr}(\boldsymbol{\Sigma})(1-d)}\right\}^{1+\delta/2} O(p^{1+\delta/2}) \\ & \leq p(\hat{k}_0 - k_0) \left\{\frac{C_2 M}{C_1 \alpha \lambda_{\min}(\boldsymbol{\Sigma})(1-d)}\right\}^{1+\delta/2} O(p^{1+\delta/2}) \\ & \leq \left\{\frac{C_2 M}{C_1 (1-d)}\right\}^{1+\delta/2} O\left(p(\hat{k}_0 - k_0) \left(\frac{p}{\alpha}\right)^{1+\delta/2}\right) \\ & = \left\{\frac{C_2 M}{C_1 (1-d)}\right\}^{1+\delta/2} O\left(\frac{p^{2+\delta} k}{\alpha^{1+\delta/2}}\right) \\ & = o(1) \end{aligned}$$

and (36) is proved.

Next, we prove (36) for case (ii) of Theorem 4. From conditions (D), (F), and (G), we have

$$\begin{aligned} & \left| \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}}\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}}^*\|^2 \right| \\ & \leq \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}')' \{ \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \otimes \mathbf{I}_p - \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}') + o_p(1) \\ & \leq \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}')' \{ \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \otimes \mathbf{I}_p - \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}') + o_p(1) \\ & = \|\{ \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}')\|^2 - \|\{ \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \otimes \mathbf{I}_p \} \text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}')\|^2 + o_p(1) \\ & = \|\text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}')\|^2 - \|\text{vec}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A)\|^2 + o_p(1) \\ & = \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2}) - \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2}) + o_p(1) \\ & = \text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \{ \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' - \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \} \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2}) + o_p(1) \end{aligned}$$

and since

$$\begin{aligned} E[\text{tr}(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' \{ \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' - \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \} \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2})] & = \text{tr}(\{ \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' - \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \} E[\boldsymbol{\varepsilon} \boldsymbol{\Sigma} \boldsymbol{\varepsilon}']) \\ & = \text{tr}(\boldsymbol{\Sigma}) \{ \text{tr}(\mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}') - \text{tr}(\mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A) \} \\ & = \text{tr}(\boldsymbol{\Sigma})(k - k_0), \end{aligned}$$

it follows from Markov's inequality that

$$\left| \|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}}\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}}^*\|^2 \right| = O_p(\text{tr}(\boldsymbol{\Sigma})(k - k_0)) + o_p(1) \quad (46)$$

holds. Also, if we set $\mathbf{P}_\omega = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^4 \right] &= \mathbb{E} \left[\text{tr} \left(\boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}' (\mathbf{I}_n - \mathbf{P}_\omega) \boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1/2} \right)^2 \right] \\ &= \left\{ \sum_{i=1}^n \{(\mathbf{I}_n - \mathbf{P}_\omega)_{ii}\}^2 \right\} \kappa_4 + (n-k)^2 \text{tr}(\boldsymbol{\Sigma})^2 + 2(n-k) \text{tr}(\boldsymbol{\Sigma}^2) \end{aligned}$$

and from $\mathbb{E} \left[\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^2 \right] = (n-k) \text{tr}(\boldsymbol{\Sigma})$,

$$\begin{aligned} \text{Var} \left[\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^2 \right] &= \left\{ \sum_{i=1}^n \{(\mathbf{I}_n - \mathbf{P}_\omega)_{ii}\}^2 \right\} \kappa_4 + 2(n-k) \text{tr}(\boldsymbol{\Sigma}^2) \\ &\leq (n-k) \max\{0, \kappa_4\} + 2(n-k) \text{tr}(\boldsymbol{\Sigma}^2) \\ &= O \left(n \max\{\kappa_4, \text{tr}(\boldsymbol{\Sigma}^2)\} \right). \end{aligned}$$

Therefore, from $\limsup_{n \rightarrow \infty} \kappa_4 \text{tr}(\boldsymbol{\Sigma})^{-2} < \infty$,

$$\begin{aligned} \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^2 &= (n-k) \text{tr}(\boldsymbol{\Sigma}) + O_p \left(n^{1/2} \max\{\kappa_4, \text{tr}(\boldsymbol{\Sigma}^2)\}^{1/2} \right) \\ &= (n-k) \text{tr}(\boldsymbol{\Sigma}) \left\{ 1 + O_p \left(\frac{1}{n^{1/2}} \max \left\{ \frac{\kappa_4}{\text{tr}(\boldsymbol{\Sigma})^2}, \frac{\text{tr}(\boldsymbol{\Sigma}^2)}{\text{tr}(\boldsymbol{\Sigma})^2} \right\} \right) \right\} \\ &= (n-k) \text{tr}(\boldsymbol{\Sigma}) \left\{ 1 + O_p \left(\frac{1}{n^{1/2}} \right) \right\} \\ &= (n-k) \text{tr}(\boldsymbol{\Sigma}) \{1 + o_p(1)\} \end{aligned} \tag{47}$$

holds. Using (37), (46), and (47), we have

$$\begin{aligned} \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}\|^2}{n(1 - \hat{k}_0 n^{-1})^\alpha} - \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}^*\|^2}{n(1 - k_0 n^{-1})^\alpha} &\geq \frac{\alpha(\hat{k}_0 - k_0)}{n - k_0} - \frac{\left| \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}\|^2 - \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}^*\|^2 \right|}{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}_0\|^2} \\ &= \frac{\alpha(\hat{k}_0 - k_0)}{n - k_0} - \frac{O_p(\text{tr}(\boldsymbol{\Sigma})(k - k_0)) + o_p(1)}{\text{tr}(\boldsymbol{\Sigma})(n-k)\{1 + o_p(1)\}} \\ &\geq \frac{\alpha}{n - k_0} - O_p \left(\frac{k - k_0}{n - k} \right) \\ &= \frac{k}{n - k_0} \left\{ \frac{\alpha}{k} - O_p(1) \right\}. \end{aligned}$$

We take an arbitrary $\varepsilon > 0$ and let $A_n = O_p(1)$. The following holds:

$$\exists K > 0 \text{ s.t. } \exists N_1 \in \mathbb{N} \text{ s.t. } \forall n \geq N_1, P(|A_n| < K) > 1 - \varepsilon. \tag{48}$$

Also, for the K in (48), from $k^{-1}\alpha \rightarrow \infty$,

$$\exists N_2 \in \mathbb{N} \text{ s.t. } \forall n \geq N_2, \frac{\alpha}{k} > K$$

holds. Thus, if we set $N = \max\{N_1, N_2\}$, for $n \geq N$,

$$\begin{aligned} 1 - \varepsilon &< P \left(\left\{ \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}\|^2}{n(1 - \hat{k}_0 n^{-1})^\alpha} - \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}^*\|^2}{n(1 - k_0 n^{-1})^\alpha} \geq \frac{k}{n - k_0} \left(\frac{\alpha}{k} - A_n \right) \right\} \cap \{|A_n| < K\} \cap \left\{ \frac{\alpha}{k} > K \right\} \right) \\ &\leq P \left(\log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}\|^2}{n(1 - \hat{k}_0 n^{-1})^\alpha} - \log \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Theta}}^*\|^2}{n(1 - k_0 n^{-1})^\alpha} > 0 \right) \end{aligned}$$

and (36) is proved. □

Appendix 6: Proof of Proposition 1

First, we will show that the objective function $Q_n(\boldsymbol{\Theta})$ is strictly convex if $\lambda_{\min}(n^{-1}\mathbf{X}'\mathbf{X}) > \lambda a^{-2}$.

$$\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2 = \frac{1}{2n} \text{tr} \left((\mathbf{Y}' - \boldsymbol{\Theta}'\mathbf{X}')' (\mathbf{Y}' - \boldsymbol{\Theta}'\mathbf{X}') \right)$$

$$\begin{aligned}
&= \frac{1}{2n} \text{vec}(\mathbf{Y}' - \boldsymbol{\Theta}' \mathbf{X}')' \text{vec}(\mathbf{Y}' - \boldsymbol{\Theta}' \mathbf{X}') \\
&= \frac{1}{2n} \{ \text{vec}(\mathbf{Y}') - \text{vec}(\boldsymbol{\Theta}' \mathbf{X}') \}' \{ \text{vec}(\mathbf{Y}') - \text{vec}(\boldsymbol{\Theta}' \mathbf{X}') \} \\
&= \frac{1}{2n} \{ \text{vec}(\mathbf{Y}') - (\mathbf{X} \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Theta}') \}' \{ \text{vec}(\mathbf{Y}') - (\mathbf{X} \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Theta}') \}
\end{aligned}$$

holds. Therefore, if $\text{vec}(\boldsymbol{\Theta}')$ is a differentiable point,

$$\nabla_{\text{vec}(\boldsymbol{\Theta}')} \left(\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2 \right) = -\frac{1}{n} (\mathbf{X} \otimes \mathbf{I}_p)' \{ \text{vec}(\mathbf{Y}') - (\mathbf{X} \otimes \mathbf{I}_p) \text{vec}(\boldsymbol{\Theta}') \},$$

so

$$\nabla_{\text{vec}(\boldsymbol{\Theta}')}^2 \left(\frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2 \right) = \frac{1}{n} (\mathbf{X}' \mathbf{X} \otimes \mathbf{I}_p)$$

holds. For the penalty term,

$$\nabla_{\boldsymbol{\theta}_j} \left\{ \lambda \sum_{\ell=1}^k \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a} \right) \right\} = \frac{\lambda}{a} e^{-\|\boldsymbol{\theta}_j\|/a} \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|},$$

which gives

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_j}^2 \left\{ \lambda \sum_{\ell=1}^k \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a} \right) \right\} &= \frac{\lambda}{a} \left(-\frac{\boldsymbol{\theta}_j \boldsymbol{\theta}_j'}{a \|\boldsymbol{\theta}_j\|^2} e^{-\|\boldsymbol{\theta}_j\|/a} - \frac{\boldsymbol{\theta}_j \boldsymbol{\theta}_j'}{\|\boldsymbol{\theta}_j\|^3} e^{-\|\boldsymbol{\theta}_j\|/a} + \frac{1}{\|\boldsymbol{\theta}_j\|} e^{-\|\boldsymbol{\theta}_j\|/a} \mathbf{I}_p \right) \\
&= \frac{\lambda}{a \|\boldsymbol{\theta}_j\|} e^{-\|\boldsymbol{\theta}_j\|/a} \left(\mathbf{I}_p - \frac{\boldsymbol{\theta}_j \boldsymbol{\theta}_j'}{\|\boldsymbol{\theta}_j\|^2} \right) - e^{-\|\boldsymbol{\theta}_j\|/a} \frac{\lambda \boldsymbol{\theta}_j \boldsymbol{\theta}_j'}{a^2 \|\boldsymbol{\theta}_j\|^2}. \tag{49}
\end{aligned}$$

The first term on the right-hand side of (49) is a positive semidefinite matrix because the eigenvalue corresponding to the eigenvector $\boldsymbol{\theta}_j$ is zero. On the other hand, let \mathbf{T} be a matrix with $\mathbf{T}_j = -e^{-\|\boldsymbol{\theta}_j\|/a} \lambda (a \|\boldsymbol{\theta}_j\|)^{-2} \boldsymbol{\theta}_j \boldsymbol{\theta}_j'$ as the j -th diagonal block. Then,

$$\frac{1}{n} (\mathbf{X}' \mathbf{X} \otimes \mathbf{I}_p) + \mathbf{T} \geq \frac{1}{n} \{ \lambda_{\min}(\mathbf{X}' \mathbf{X}) \mathbf{I}_k \otimes \mathbf{I}_p \} + \mathbf{T}$$

holds. Here, for matrices \mathbf{A} and \mathbf{B} , $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix. Also, since $e^{-\|\boldsymbol{\theta}_j\|/a} \leq 1$,

$$\frac{1}{n} \lambda_{\min}(\mathbf{X}' \mathbf{X}) \mathbf{I}_p - \mathbf{T}_j \geq \frac{1}{n} \lambda_{\min}(\mathbf{X}' \mathbf{X}) \mathbf{I}_p - \frac{\lambda \boldsymbol{\theta}_j \boldsymbol{\theta}_j'}{a^2 \|\boldsymbol{\theta}_j\|^2}$$

holds. Furthermore,

$$\left(\frac{1}{n} \lambda_{\min}(\mathbf{X}' \mathbf{X}) \mathbf{I}_p - \frac{\lambda \boldsymbol{\theta}_j \boldsymbol{\theta}_j'}{a^2 \|\boldsymbol{\theta}_j\|^2} \right) \boldsymbol{\theta}_j = \left\{ \lambda_{\min} \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right) - \frac{\lambda}{a^2} \right\} \boldsymbol{\theta}_j.$$

Therefore, a sufficient condition for $n^{-1} \lambda_{\min}(\mathbf{X}' \mathbf{X}) \mathbf{I}_p - \lambda (a \|\boldsymbol{\theta}_j\|)^{-2} \boldsymbol{\theta}_j \boldsymbol{\theta}_j'$ to be a positive definite matrix is that $\lambda_{\min}(n^{-1} \mathbf{X}' \mathbf{X}) - \lambda a^{-2} > 0$. From the above, if $\lambda_{\min}(n^{-1} \mathbf{X}' \mathbf{X}) > \lambda a^{-2}$, then $\nabla_{\text{vec}(\boldsymbol{\Theta}')}^2 Q_n(\boldsymbol{\Theta})$ is a positive definite matrix.

Next, consider the case where $\text{vec}(\boldsymbol{\Theta}')$ is a non-differentiable point. We need to show that for any directional derivative, the right derivative is greater than or equal to the left derivative, i.e.,

$$\forall \mathbf{v} \in \mathbb{R}^{kp}, \quad \nabla_{\mathbf{v}}^+ Q_n(\boldsymbol{\Theta}) - \nabla_{\mathbf{v}}^- Q_n(\boldsymbol{\Theta}) \geq 0.$$

Take any $\mathbf{v} \in \mathbb{R}^{kp}$. Let \mathbf{v}_ℓ ($\ell = 1, \dots, k$) be the ℓ -th block vector. For ℓ such that $\boldsymbol{\theta}_\ell = \mathbf{0}_p$, we show

$$\nabla_{\mathbf{v}_\ell}^+ \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a} \right) - \nabla_{\mathbf{v}_\ell}^- \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a} \right) \geq 0.$$

Let $\boldsymbol{\theta}_\ell = \mathbf{0}_p$ and use Taylor expansion. Then,

$$\begin{aligned}
\nabla_{\mathbf{v}_\ell}^+ \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a} \right) &= \lim_{h \rightarrow +0} \left(\frac{-e^{-\|\boldsymbol{\theta}_\ell + h \mathbf{v}_\ell\|/a} + e^{-\|\boldsymbol{\theta}_\ell\|/a}}{h} \right) \\
&= \lim_{h \rightarrow +0} \left(\frac{-e^{-h \|\mathbf{v}_\ell\|/a} + 1}{h} \right) \\
&= \lim_{h \rightarrow +0} \frac{h \|\mathbf{v}_\ell\| a^{-1} + O(h^2)}{h} \\
&= \frac{\|\mathbf{v}_\ell\|}{a}.
\end{aligned}$$

Similarly, for the left derivative,

$$\nabla_{\mathbf{v}_\ell}^- \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a}\right) = -\frac{\|\mathbf{v}_\ell\|}{a}$$

holds. Therefore,

$$\nabla_{\mathbf{v}_\ell}^+ \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a}\right) - \nabla_{\mathbf{v}_\ell}^- \left(1 - e^{-\|\boldsymbol{\theta}_\ell\|/a}\right) = \frac{2\|\mathbf{v}_\ell\|}{a} \geq 0$$

Thus, it is shown that the strict convexity condition for $Q_n(\boldsymbol{\Theta})$ is $\lambda_{\min}(n^{-1}\mathbf{X}'\mathbf{X}) > \lambda a^{-2}$.

Next, we show that if $n^{-1}\|\mathbf{x}_{(j)}\|^2 > \lambda a^{-2}$, $Q_n(\boldsymbol{\Theta})$ is a strictly convex function in the $\boldsymbol{\theta}_j$ direction. When $\boldsymbol{\theta}_j$ is a differentiable point,

$$\nabla_{\boldsymbol{\theta}_j} \left(\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2\right) = -\frac{1}{n}(\mathbf{x}_{(j)} \otimes \mathbf{I}_p)' \{\text{vec}(\mathbf{Y}') - (\mathbf{X} \otimes \mathbf{I}_p)\text{vec}(\boldsymbol{\Theta}')\},$$

so

$$\nabla_{\boldsymbol{\theta}_j}^2 \left(\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2\right) = \frac{1}{n}(\mathbf{x}'_{(j)}\mathbf{x}_{(j)} \otimes \mathbf{I}_p) = \frac{\|\mathbf{x}_{(j)}\|^2}{n}\mathbf{I}_p.$$

Therefore, from (49),

$$\nabla_{\boldsymbol{\theta}_j}^2 Q_n(\boldsymbol{\Theta}) = \frac{\lambda}{a\|\boldsymbol{\theta}_j\|} e^{-\|\boldsymbol{\theta}_j\|/a} \left(\mathbf{I}_p - \frac{\boldsymbol{\theta}_j\boldsymbol{\theta}'_j}{\|\boldsymbol{\theta}_j\|^2}\right) + \frac{\|\mathbf{x}_{(j)}\|^2}{n}\mathbf{I}_p - e^{-\|\boldsymbol{\theta}_j\|/a} \frac{\lambda\boldsymbol{\theta}_j\boldsymbol{\theta}'_j}{a^2\|\boldsymbol{\theta}_j\|^2} \quad (50)$$

and the first term on the right-hand side of (50) is a positive semidefinite matrix. For the remaining terms,

$$\frac{\|\mathbf{x}_{(j)}\|^2}{n}\mathbf{I}_p - e^{-\|\boldsymbol{\theta}_j\|/a} \frac{\lambda\boldsymbol{\theta}_j\boldsymbol{\theta}'_j}{a^2\|\boldsymbol{\theta}_j\|^2} \geq \frac{\|\mathbf{x}_{(j)}\|^2}{n}\mathbf{I}_p - \frac{\lambda\boldsymbol{\theta}_j\boldsymbol{\theta}'_j}{a^2\|\boldsymbol{\theta}_j\|^2} \quad (51)$$

holds. Therefore, for the right-hand side of (51),

$$\left(\frac{\|\mathbf{x}_{(j)}\|^2}{n}\mathbf{I}_p - \frac{\lambda\boldsymbol{\theta}_j\boldsymbol{\theta}'_j}{a^2\|\boldsymbol{\theta}_j\|^2}\right)\boldsymbol{\theta}_j = \left(\frac{\|\mathbf{x}_{(j)}\|^2}{n} - \frac{\lambda}{a^2}\right)\boldsymbol{\theta}_j.$$

Furthermore, if $n^{-1}\|\mathbf{x}_{(j)}\|^2 - \lambda a^{-2} > 0$, then $\nabla_{\boldsymbol{\theta}_j}^2 Q_n(\boldsymbol{\Theta})$ is a positive definite matrix. The case where $\boldsymbol{\theta}_j$ is a non-differentiable point is the same as before. Thus, if $n^{-1}\|\mathbf{x}_{(j)}\|^2 > \lambda a^{-2}$, it is shown that $Q_n(\boldsymbol{\Theta})$ is strictly convex in the $\boldsymbol{\theta}_j$ direction. This concludes the proof of Proposition 1. \square

Appendix 7: Derivation of the update rule (10) in the GCD algorithm

We assume the strict convexity condition in the vector direction of the objective function $Q_n(\boldsymbol{\Theta})$, which is $n^{-1}\|\mathbf{x}_{(\ell)}\|^2 > \lambda a^{-2}$ ($\ell = 1, \dots, k$). We set the notation as follows: $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta}')$, $\mathbf{y} = \text{vec}(\mathbf{Y}')$, and $\mathbf{b} = (\mathbf{X}' \otimes \mathbf{I}_p)\text{vec}(\mathbf{Y}')$. Also, let $\mathbf{b}_j \in \mathbb{R}^p$ be the j -th block vector of \mathbf{b} , and set $\mathbf{c}_\ell = \mathbf{b}_\ell - \sum_{j \neq \ell} (\mathbf{x}'_{(\ell)}\mathbf{x}_{(j)})\boldsymbol{\theta}_j$. Then, the following holds:

$$\begin{aligned} Q_n(\boldsymbol{\Theta}) &= \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}\|^2 + \lambda \sum_{j=1}^k \left(1 - e^{-\|\boldsymbol{\theta}_j\|/a}\right) \\ &= \frac{1}{2n}\|\text{vec}(\mathbf{Y}') - (\mathbf{X} \otimes \mathbf{I}_p)\text{vec}(\boldsymbol{\Theta}')\|^2 + \lambda \sum_{j=1}^k \left(1 - e^{-\|\boldsymbol{\theta}_j\|/a}\right) \\ &= \frac{1}{2n}\{\text{vec}(\boldsymbol{\Theta}')(\mathbf{X}'\mathbf{X} \otimes \mathbf{I}_p)\text{vec}(\boldsymbol{\Theta}') - 2\text{vec}(\boldsymbol{\Theta}')(\mathbf{X}' \otimes \mathbf{I}_p)\text{vec}(\mathbf{Y}') + \text{vec}(\mathbf{Y}')'\text{vec}(\mathbf{Y}')\} + \lambda \sum_{j=1}^k \left(1 - e^{-\|\boldsymbol{\theta}_j\|/a}\right) \\ &= \frac{1}{2n}\{\boldsymbol{\theta}'(\mathbf{X}'\mathbf{X} \otimes \mathbf{I}_p)\boldsymbol{\theta} - 2\boldsymbol{\theta}'\mathbf{b} + \mathbf{y}'\mathbf{y}\} + \lambda \sum_{j=1}^k \left(1 - e^{-\|\boldsymbol{\theta}_j\|/a}\right) \\ &= \frac{1}{2n}(\|\mathbf{x}_{(\ell)}\|^2\|\boldsymbol{\theta}_\ell\|^2 - 2\mathbf{c}'_\ell\boldsymbol{\theta}_\ell) - \lambda e^{-\|\boldsymbol{\theta}_\ell\|/a} + (\text{terms not dependent on } \boldsymbol{\theta}_\ell). \end{aligned}$$

Here, by finding the subgradient with respect to $\boldsymbol{\theta}_\ell$ for $Q_n(\boldsymbol{\Theta})$, we obtain the following equation:

$$\frac{\|\mathbf{x}_{(\ell)}\|^2}{n}\boldsymbol{\theta}_\ell - \frac{1}{n}\mathbf{c}_\ell + \lambda\mathbf{s}_\ell = \mathbf{0}_p,$$

where

$$\mathbf{s}_\ell \in \begin{cases} \left\{ \frac{e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a\|\boldsymbol{\theta}_\ell\|} \boldsymbol{\theta}_\ell \right\} & \text{if } \boldsymbol{\theta}_\ell \neq \mathbf{0}_p \\ \left\{ \mathbf{v} \in \mathbb{R}^p \mid \|\mathbf{v}\| \leq \frac{1}{a} \right\} & \text{if } \boldsymbol{\theta}_\ell = \mathbf{0}_p \end{cases}.$$

Therefore, when $\boldsymbol{\theta}_\ell = \mathbf{0}_p$,

$$\begin{aligned} \frac{\|\mathbf{x}_{(\ell)}\|^2}{n} \boldsymbol{\theta}_\ell - \frac{1}{n} \mathbf{c}_\ell + \lambda \mathbf{s}_\ell = \mathbf{0}_p &\Leftrightarrow -\frac{1}{n} \mathbf{c}_\ell + \lambda \mathbf{s}_\ell = \mathbf{0}_p \\ &\Leftrightarrow \mathbf{s}_\ell = \frac{1}{n\lambda} \mathbf{c}_\ell \\ &\Leftrightarrow \frac{\|\mathbf{c}_\ell\|}{n\lambda} \leq \frac{1}{a} \\ &\Leftrightarrow \|\mathbf{c}_\ell\| \leq \frac{n\lambda}{a}. \end{aligned}$$

On the other hand, when $\boldsymbol{\theta}_\ell \neq \mathbf{0}_p$,

$$\begin{aligned} \frac{\|\mathbf{x}_{(\ell)}\|^2}{n} \boldsymbol{\theta}_\ell - \frac{1}{n} \mathbf{c}_\ell + \lambda \mathbf{s}_\ell = \mathbf{0}_p &\Leftrightarrow \frac{\|\mathbf{x}_{(\ell)}\|^2}{n} \boldsymbol{\theta}_\ell - \frac{1}{n} \mathbf{c}_\ell + \frac{\lambda e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a\|\boldsymbol{\theta}_\ell\|} \boldsymbol{\theta}_\ell = \mathbf{0}_p \\ &\Leftrightarrow \left(\frac{\|\mathbf{x}_{(\ell)}\|^2}{n} \mathbf{I}_p + \frac{\lambda e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a\|\boldsymbol{\theta}_\ell\|} \mathbf{I}_p \right) \boldsymbol{\theta}_\ell = \frac{1}{n} \mathbf{c}_\ell \\ &\Leftrightarrow \boldsymbol{\theta}_\ell = \left(\frac{\|\mathbf{x}_{(\ell)}\|^2}{n} + \frac{\lambda e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a\|\boldsymbol{\theta}_\ell\|} \right)^{-1} \frac{1}{n} \mathbf{c}_\ell. \end{aligned}$$

Taking the norm of both sides, we have

$$\begin{aligned} \|\boldsymbol{\theta}_\ell\| &= \left(\frac{\|\mathbf{x}_{(\ell)}\|^2}{n} + \frac{\lambda e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a\|\boldsymbol{\theta}_\ell\|} \right)^{-1} \frac{1}{n} \|\mathbf{c}_\ell\| \Leftrightarrow \frac{\|\mathbf{x}_{(\ell)}\|^2}{n} \|\boldsymbol{\theta}_\ell\| + \frac{\lambda e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a} = \frac{1}{n} \|\mathbf{c}_\ell\| \\ &\Leftrightarrow \|\mathbf{x}_{(\ell)}\|^2 \|\boldsymbol{\theta}_\ell\| - \|\mathbf{c}_\ell\| + \frac{n\lambda e^{-\|\boldsymbol{\theta}_\ell\|/a}}{a} = 0. \end{aligned}$$

Now, let $g(x) = \|\mathbf{x}_{(\ell)}\|^2 x + n\lambda a^{-1} e^{-x/a}$. Then,

$$\begin{aligned} \nabla g(x) = 0 &\Leftrightarrow \|\mathbf{x}_{(\ell)}\|^2 - \frac{n\lambda}{a^2} e^{-x/a} = 0 \\ &\Leftrightarrow e^{x/a} = \frac{\lambda}{a^2} \cdot \frac{n}{\|\mathbf{x}_{(\ell)}\|^2} \\ &\Leftrightarrow x = a \log \frac{\lambda}{a^2} \cdot \frac{n}{\|\mathbf{x}_{(\ell)}\|^2}. \end{aligned}$$

Thus, since $\nabla g(x)$ is a strictly increasing function and from the strict convexity condition of the objective function in the vector direction, $a \log n\lambda/(a^2 \|\mathbf{x}_{(\ell)}\|^2) < 0$, we have $\nabla g(x) > 0$ on $(0, \infty)$, which means that $g(x)$ is a strictly increasing function on $(0, \infty)$. Since $g(0) = n\lambda/a (< \|\mathbf{c}_\ell\|)$ and $\lim_{x \rightarrow \infty} g(x) = \infty$, there exists exactly one positive solution to $\|\mathbf{x}_{(\ell)}\|^2 x - \|\mathbf{c}_\ell\| + n\lambda a^{-1} e^{-x/a} = 0$. We denote this solution by x_0 . Then, $\boldsymbol{\theta}_\ell = x_0 \|\mathbf{c}_\ell\|^{-1} \mathbf{c}_\ell$. From the above, we obtain the update rule (10) for the GCD algorithm.

Appendix 8: Proof of Lemma A.2

We prove this by mathematical induction. For the base case, let $n = 1$. By the Cauchy-Schwarz inequality,

$$\mathbb{E}[\|\boldsymbol{\varepsilon}' \mathbf{t}^{2+\delta}\|] \leq \mathbb{E}[\|\boldsymbol{\varepsilon}\|^{2+\delta} \|\mathbf{t}\|^{2+\delta}] = \mathbb{E}[\|\boldsymbol{\varepsilon}\|^{2+\delta}] = O(p^{1+\delta/2}).$$

This completes the proof for the case where $n = 1$. Next, for the inductive step, we assume the claim holds for $n = k - 1$ (≥ 1) and prove it for $n = k$. Let $\mathbf{t} \in \mathbb{R}^{kp}$ be an arbitrary vector with $\|\mathbf{t}\| = 1$, and let $\mathbf{t}_i \in \mathbb{R}^p$ be the i -th block vector of \mathbf{t} . Without loss of generality, assume $\|\mathbf{t}_k\| < 1$. Define $\tilde{\mathbf{t}}_i = (1 - \|\mathbf{t}_k\|^2)^{-1/2} \mathbf{t}_i$ ($i = 1, \dots, k - 1$), and set $S_k = \sum_{i=1}^k \boldsymbol{\varepsilon}'_i \mathbf{t}_i$ and $S_{k-1} = \sum_{i=1}^{k-1} \boldsymbol{\varepsilon}'_i \tilde{\mathbf{t}}_i$. Note that $S_k = (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} + \boldsymbol{\varepsilon}'_k \mathbf{t}_k$. Applying the Taylor theorem, we have

$$\begin{aligned} |S_k|^{2+\delta} &= \left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} + \boldsymbol{\varepsilon}'_k \mathbf{t}_k \right|^{2+\delta} \\ &= \left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} \right|^{2+\delta} + (2 + \delta) \text{sign}(S_{k-1}) \left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} \right|^{1+\delta} \boldsymbol{\varepsilon}'_k \mathbf{t}_k \end{aligned}$$

$$+ \frac{1}{2}(2 + \delta)(1 + \delta) \left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} + \xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k \right|^\delta (\boldsymbol{\varepsilon}'_k \mathbf{t}_k)^2 \quad (\xi \in [0, 1]).$$

Furthermore, by the triangle inequality and $\xi \in [0, 1]$,

$$\begin{aligned} \left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} + \xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k \right|^\delta &\leq \left\{ (1 - \|\mathbf{t}_k\|^2)^{1/2} |S_{k-1}| + |\xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k| \right\}^\delta \\ &\leq 2^\delta \left\{ (1 - \|\mathbf{t}_k\|^2)^{-\delta/2} |S_{k-1}|^\delta + |\xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k|^\delta \right\}. \end{aligned}$$

Applying Hölder's inequality, we have

$$\begin{aligned} \mathbb{E} \left[|S_{k-1}|^\delta \|\boldsymbol{\varepsilon}_k\|^2 \right] &\leq \mathbb{E} \left[|S_{k-1}|^{2+\delta} \right]^{\delta/(2+\delta)} \mathbb{E} \left[\|\boldsymbol{\varepsilon}_k\|^{2+\delta} \right]^{2/(2+\delta)} \\ &= \left\{ O(p^{1+\delta/2}) \right\}^{\delta/(2+\delta)} \left\{ O(p^{1+\delta/2}) \right\}^{2/(2+\delta)} \\ &= O(p^{1+\delta/2}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E} \left[|S_k|^{2+\delta} \right] &= \mathbb{E} \left[\left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} \right|^{2+\delta} \right] + \frac{1}{2}(2 + \delta)(1 + \delta) \mathbb{E} \left[\left| (1 - \|\mathbf{t}_k\|^2)^{1/2} S_{k-1} + \xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k \right|^\delta (\boldsymbol{\varepsilon}'_k \mathbf{t}_k)^2 \right] \\ &\leq (1 - \|\mathbf{t}_k\|^2)^{1+\delta/2} \mathbb{E} \left[|S_{k-1}|^{2+\delta} \right] + \frac{1}{2}(2 + \delta)(1 + \delta) \mathbb{E} \left[2^\delta \left\{ (1 - \|\mathbf{t}_k\|^2)^{-\delta/2} |S_{k-1}|^\delta + |\xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k|^\delta \right\} (\boldsymbol{\varepsilon}'_k \mathbf{t}_k)^2 \right] \\ &\leq \mathbb{E} \left[|S_{k-1}|^{2+\delta} \right] + 2^{\delta-1} (2 + \delta)(1 + \delta) \mathbb{E} \left[\left(|S_{k-1}|^\delta + |\xi \boldsymbol{\varepsilon}'_k \mathbf{t}_k|^\delta \right) \|\boldsymbol{\varepsilon}_k\|^2 \|\mathbf{t}_k\|^2 \right] \\ &\leq O(p^{1+\delta/2}) + 2^{\delta-1} (2 + \delta)(1 + \delta) \left(\mathbb{E} \left[|S_{k-1}|^\delta \|\boldsymbol{\varepsilon}_k\|^2 \right] + \mathbb{E} \left[\|\boldsymbol{\varepsilon}_k\|^{2+\delta} \right] \right) \\ &\leq O(p^{1+\delta/2}) + 2^{\delta-1} (2 + \delta)(1 + \delta) \left\{ O(p^{1+\delta/2}) + O(p^{1+\delta/2}) \right\} \\ &= O(p^{1+\delta/2}). \end{aligned}$$

This completes the proof of Lemma A.2. □

Acknowledgments

The authors wish to thank Prof. Shinpei Imori of Hiroshima University for drawing our attention to Lemma A.1, which is essential to the proof of (i) in Theorem 4, and FORTE Science Communications (<https://www.forte-science.co.jp/>) for English language editing. This work was partially supported by JSPS KAKENHI Grant Number 25K17296.

References

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, F. Csaki (Eds.), *Second International Symposium on Information Theory*, 267–281.
- [2] Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica*, **28**, 43–62.
- [3] Bertsimas, D., King, A. and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.*, **44**, 813–852.
- [4] Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, **71**, 731–740.
- [5] Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.*, **25**, 173–187.
- [6] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350–2383.
- [7] Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **72** (1), 3–25.
- [8] Chen, Y., Luo, Z. and Kong, K. (2021). $\ell_{2,0}$ -norm based selection and estimation for multivariate generalized linear models. *J. Multivariate Anal.*, **185**, 104782.
- [9] Cui, J. and Yi, G. Y. (2024). Variable selection in multivariate regression models with measurement error in covariates. *J. Multivariate Anal.*, **202**, 105299.
- [10] Dicker, L., Huang, B. and Lin, X. (2013). Variable selection and estimation with the seamless- ℓ_0 penalty. *Statist. Sinica*, **23**, 929–962.

- [11] Dong, Y., Song, L. and Amin, M. (2010). SCAD-Ridge penalized likelihood estimators for ultra-high dimensional models. *Comput. Statist. Data Anal.*, **54**, 2230–2243.
- [12] VanDerwerken, D. N. (2011). Variable selection and parameter estimation using a continuous and differentiable approximation to the ℓ_0 penalty function. *All Theses and Dissertations*, Paper 2486.
- [13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.
- [14] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, **32**, 928–961.
- [15] Friedman, J. H., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Probab.*, **2**, 302–332.
- [16] Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.
- [17] Fujikoshi, Y. and Kan, T. (2018). A review of discriminant analysis by regression approach. TR No. 18–06, *Statistical Research Group*, Hiroshima University.
- [18] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statist. Sci.*, **27**, 481–499.
- [19] Huang, J., Breheny, P., Ma, S. and Zhang, C. H. (2016). The Mnet method for variable selection. *Statist. Sinica*, **26**, 903–923.
- [20] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.
- [21] Mallows, C. (1973). Some comments on C_p . *Technometrics*, **15**, 661–675.
- [22] Oda, R. (2020). Consistent variable selection criteria in multivariate linear regression even when dimension exceeds sample size. *Hiroshima Math. J.*, **50**, 339–374.
- [23] Oda, R., Ohishi, M., Suzuki, Y. and Yanagihara, H. (2023). An $\ell_{2,0}$ -norm constrained matrix optimization via extended discrete first-order algorithms. *Hiroshima. Math. J.*, **53**, 251–267.
- [24] Ohishi, M., Yanagihara, H. and Fujikoshi, Y. (2020). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. *J. Statist. Plann. Inference*, **204**, 187–205.
- [25] Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.
- [26] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
- [27] Shao, J. (2003). *Mathematical Statistics (Second Edition)*. Springer-Verlag, New York.
- [28] Sofer, T., Dicker, L. and Lin, X. (2014). Variable selection for high dimensional multivariate outcomes. *Statist. Sinica*, **24**, 1633–1654.
- [29] Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. John Wiley & Sons, New York.
- [30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.
- [31] Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer-Verlag, New York.
- [32] Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486–1494.
- [33] Wang, Y., Fan, Q. and Zhu, L. (2018). Variable selection and estimation using a continuous approximation to the ℓ_0 penalty. *Ann. Inst. Statist. Math.*, **70**, 191–214.
- [34] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **68**, 49–67.
- [35] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
- [36] Zhu, J., Wen, C., Zhu, J., Zhang, H. and Wang, X. (2020). A polynomial algorithm for best-subset selection problem. *Proc. Natl. Acad. Sci. USA*, **52**, 33117–33123.
- [37] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.
- [38] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **67**, 301–320.
- [39] Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, **37**, 1733–1751.