Bernoulli **22**(3), 2016, 1937–1961
DOI: 10.3150/15-BEJ715

An analysis of penalized interaction models

JUNLONG ZHAO[1] and CHENLEI LENG[2]

[1] *School of Mathematics and Systems Science, Beihang University, LMIB of the Ministry of Education, China. E-mail: zhaojunlong928@126.com*
[2] *Department of Statistics, University of Warwick, UK. E-mail: C.Leng@warwick.ac.uk*

An important consideration for variable selection in interaction models is to design an appropriate penalty that respects hierarchy of the importance of the variables. A common theme is to include an interaction term only after the corresponding main effects are present. In this paper, we study several recently proposed approaches and present a unified analysis on the convergence rate for a class of estimators, when the design satisfies the restricted eigenvalue condition. In particular, we show that with probability tending to one, the resulting estimates have a rate of convergence $s\sqrt{\log p_1/n}$ in the ℓ_1 error, where p_1 is the ambient dimension, s is the true dimension and n is the sample size. We give a new proof that the restricted eigenvalue condition holds with high probability, when the variables in the main effects and the errors follow sub-Gaussian distributions. Under this setup, the interactions no longer follow Gaussian or sub-Gaussian distributions even if the main effects follow Gaussian, and thus existing works are not applicable. This result is of independent interest.

Keywords: convergence rate; hierarchical variable selection; high-dimensionality; interaction models; Lasso; restricted eigenvalue condition

1. Introduction

High-dimensional datasets are predominantly characterized by a large ambient dimension of the covariates p and a small number of the observations n. An important assumption in analyzing such datasets is that the true dimension of the relevant covariates s is often smaller than n. Let $(X_i, Y_i), i = 1, \dots, n$ be i.i.d. observations where $X_i = (X_{i1}, \dots, X_{ip})^T$ is a p-dimensional regressor and Y_i is a scalar response. Denote $\mathbb{X} = (X_1, \dots, X_n)^T \in R^{n\times p}$, $\mathbb{Y} = (Y_1, \dots, Y_n)^T \in R^n$. Consider the usual linear model

$$\mathbb{Y} = \mathbb{X}\alpha + \varepsilon, \tag{1.1}$$

where $\alpha \in R^p$ is a sparse vector, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ and ε_i are i.i.d. errors. There are a number of important approaches for recovering the sparse vector α from this model, such as Lasso (Tibshirani [23]), the SCAD (Fan and Li [7]), the Dantzig selector (Candes and Tao [5]) and many others. The main idea is to make use of a sparsity-encouraging penalty that gives coefficient estimation and variable selection at the same time. In particular, the popular Lasso method aims to estimate α as

$$\hat{\alpha} = \arg\min_{\alpha} \|\mathbb{Y} - \mathbb{X}\alpha\|_2^2 + \lambda_n \|\alpha\|_1,$$

where λ_n is a penalty parameter suitably chosen and $\|\cdot\|_q$ is the ℓ_q norm of a vector. A fundamental problem in analyzing such approaches is to derive sufficient conditions such that the resulting

1350-7265

estimates are consistent with appropriate rate of convergence. Zhang and Huang [31] analyzed the sparsity and bias of this formulation. Meinshausen and Yu [15] showed that with high probability, all important variables are selected by the Lasso. In an important work, Bickel, Ritov and Tsybakov [2] proposed the restricted eigenvalue (RE) condition. van de Geer and Bühlmann [25] considered a weaker compatibility condition to guarantee the consistency of Lasso in the ℓ_1 or ℓ_2 sense. Subsequently, many authors studied sufficient conditions under which the restricted eigenvalue condition holds. Raskutti, Wainwright and Yu [19] proved that the RE condition holds with large probability for Gaussian designs, and Rudelson and Zhou [21] showed that the RE condition holds for sub-Gaussian distributions with large probability.

In practice, it is often necessary, sometimes even mandatory, to consider interactions of the main variables and to select the variables in a hierarchical fashion. For example, in genetic studies, the main effects of genetic and environmental factors often fail to explain enough phenotypical variation. Gene-environment interactions are considered pivotal in this context. In analysis of variance models, interactions are often analyzed only when the corresponding main effects are present. In these settings, we consider not only the main effects of $X_{i1}, \dots, X_{ip}$ but also the effects of their interactions $X_{ik_1} X_{ik_2}$, for any $k_1, k_2 = 1, \dots, p$ with $k_1 < k_2$. Let $Z_i = (X_i^T, X_i^{*T})^T$ with $X_i = (X_{i1}, \dots, X_{ip})^T \in R^p$ and $X_i^* = (X_{i1}X_{i2}, \dots, X_{ik_1}X_{ik_2}, \dots, X_{i(p-1)}X_{ip})^T$, $i = 1, \dots, n$ and $\mathbb{Z} = (Z_1, \dots, Z_n)^T \in R^{n \times p_1}$ with $p_1 = p(p+1)/2$. Similar to the model that only contains the main effects, we write an interaction model as

$$\mathbb{Y} = \mathbb{Z}\beta + \varepsilon,$$

where $\beta \in R^{p_1}$ and $\mathbb{Z} \in R^{n \times p_1}$ with columns consisting of the main effects and interaction terms. In this paper, we only consider two-way interaction models. It is unclear if the main results can be extended to higher-order interaction models.

It is generally desirable to select an interaction effect only after the corresponding main effects are selected. As Bien, Taylor and Tibshirani [4] and the references therein argued, an interaction should be allowed into the model only if the corresponding main effects are present. A version of this statement can be found in Assumption (A1) in Section 2. To achieve this goal, several methods have been developed. Yuan, Joseph and Zou [29] proposed to use a two-step procedure after obtaining a first-step estimate that is consistent. The theoretical properties are studied when p is fixed. Radchenko and James [18] used a nonconvex penalty for enforcing this strong heredity constraint. A few approaches formulated as convex optimization were also studied. Zhao, Rocha and Yu [32] proposed to use the CAP approach and Radchenko and James [18] proposed the VANISH method for hierarchical selection of variables. Bach *et al.* [1] suggested a framework for structured penalty. Bien, Taylor and Tibshirani [4] studied a constrained optimization model for the same purpose. These three papers (Zhao, Rocha and Yu [32], Choi, Li and Zhu [6], Bach *et al.* [1]) consider convex optimization problems of the following type:

$$\hat{\beta} = \arg\min_{\theta} \|\mathbb{Y} - \mathbb{Z}\theta\|_2^2 + \lambda_n P_e(\theta),$$

where $P_e(\theta)$ is a convex penalty usually consisting of norms on groups of variables Yuan and Lin [30], Lin and Zhang [13]. A detailed discussion of this penalty is given in Section 2. A desirable property of this one-step formulation is that convex optimization techniques can be exploited to yield a global optimal solution. Among these convex formulations, only Radchenko and James [18] studied theoretical properties of their estimate in terms of model selection consistency when the design matrix satisfies strong conditions, similar to the irrepresentable conditions in Zhao and Yu [33]. It is thus of great interest to explore other theoretical aspects of a selected interaction model.

Although the convex optimization methods are attractive conceptually, the consistency of the estimates for large p in terms of the convergence rate is largely unknown, especially under weaker conditions than those in Radchenko and James [18]. The main goal of this paper is to consider this important issue for a class of penalties. Our results give a general method for proving consistency when interactions are present. In particular, two major contributions are made in this paper.

First, assuming sub-Gaussian distributions on the error and the variables in the main effects, we establish the consistency of penalized estimators in interaction models proposed by Zhao, Rocha and Yu [32], Radchenko and James [18], Bach *et al.* [1] and Bien, Taylor and Tibshirani [4]. In particular, we show that under appropriate conditions and a modified restricted eigenvalue (RE) condition Bickel, Ritov and Tsybakov [2], these penalized estimates achieve the rate of convergence in the order $s\sqrt{\log p_1/n}$, which matches the optimal rate of the Lasso when no interaction is considered. These results also motivate us to develop a unified analysis of a general class of methods for penalized variable selection in interaction models. The main results are nonasymptotic and presented in Theorem 2.1. We note that Negahban *et al.* [17] put forward a unified theory for high-dimensional M-estimation when the penalty satisfies a notion of decomposability. However, as will become clear later, the penalties we investigate in this paper are not necessarily decomposable, and thus the theory developed in Negahban *et al.* [17] is not directly applicable to the interaction models under consideration.

Second, we generalize the RE condition to the case where both the main and the interaction terms are presented in covariate Z_i. We remark that the presence of the interactions poses significant theoretical difficulties. For example, when the main effect covariates follow a multivariate Gaussian distribution, the joint distribution of the main effects and the interactions are no longer Gaussian, or even sub-Gaussian. Thus, the regularity conditions of some existing results are violated, for example, in Raskutti, Wainwright and Yu [19] and Rudelson and Zhou [21]. A new analysis in Theorem 3.1 is provided to address this issue. This result can be of independent interest.

There is a large literature more recently devoted to selecting variables in interaction models. One particular class is based on identifying a superset of the variables that includes the true model as a subset. Hao and Zhang [10–12] utilized an interesting algorithm to build a hierarchical model using the forward selection (Wang [27]). Another class is to exploit algorithms for interaction selection. Shah [22] exploited a backtracking iterative algorithm to identify a true model. Hall and Xue [9] adopted a simple recursive approach to allow identifications. Bickel, Ritov and Tsybakov [3] studied a procedure with a sequential Lasso fit.

The main content of this paper is arranged as follows. In Section 2, we establish the consistency of the estimators with penalty functions defined in Zhao, Rocha and Yu [32], Radchenko and James [18], and Bien, Taylor and Tibshirani [4]. We show that our result is applicable to a wider class of penalties appropriately defined. In Section 3, we generalize the RE condition to the interaction model setting. We develop sufficient conditions to guarantee the RE condition to hold. The proofs are delayed to Appendix A and some auxiliary results are presented in Appendices B and C.

2. A theory for penalized estimation in interaction models

Recall that $p_1 = p(p+1)/2$ is the number of the variables in an interaction model with p main effect covariates. Consider the model

$$\mathbb{Y} = \mathbb{Z}\beta + \varepsilon,$$

where $E(\varepsilon) = 0$ and $\mathbb{Z} = (Z_1, \ldots, Z_n)^T$ is an n by p_1 matrix. The columns of $\mathbb{Z}$ consist of the main effects and the second-order interactions as $X_{ik_1}X_{ik_2}$, $k_1 < k_2$, and Z_i is independent of ε_i. Without loss of generality, we assume that $E(X_i) = 0$ and $E(X_{ik_1}X_{ik_2}) = 0$ for $k_1 < k_2$. Otherwise, we can replace X_i and $X_{ik_1}X_{ik_2}$ by $X_i - E(X_i)$ and $X_{ik_1}X_{ik_2} - E(X_{ik_1}X_{ik_2})$, respectively. Let $\mathrm{cov}(X_i) = \Sigma_x$ and $\lambda_{\max,x}, \lambda_{\min,x}$ be the corresponding largest and smallest eigenvalues, respectively. Similarly, we denote $\mathrm{cov}(Z_i) = \Sigma_z$ and define $\lambda_{\max,z}, \lambda_{\min,z}$. Denote the true parameter as $\beta = (\beta^{(1)T}, \beta^{(2)T})^T$ with $\beta^{(1)} = (\beta_1, \ldots, \beta_p)^T \in R^p$ and $\beta^{(2)} = (\beta_{12}, \ldots, \beta_{(p-1)p})^T \in R^{p_1 - p}$. Denote support sets of $\beta^{(1)}$, $\beta^{(2)}$ and β as $\mathcal{S}^{(1)}$, $\mathcal{S}^{(2)}$ and S, respectively. Then it is clear that $S = \mathcal{S}^{(1)} \cup \mathcal{S}^{(2)}$. Let $s = |S|$ be the cardinality of S. To simplify the description, we also denote $\beta = (\beta_1, \ldots, \beta_p, \ldots, \beta_{p_1})^T$ and use a single index j to denote either a main effect or an interaction. We focus on the strong hierarchy principle Yuan, Joseph and Zou [29], Choi, Li and Zhu [6] as stated in the following assumption.

(A1) If $(i, j) \in \mathcal{S}^{(2)}$ then both $i \in \mathcal{S}^{(1)}$ and $j \in \mathcal{S}^{(1)}$.

Assumption (A1) means that if an interaction is selected, that is, if $\beta_{ij} \neq 0$, then the corresponding main effects should also be selected, that is, $\beta_i \neq 0$ and $\beta_j \neq 0$, for any $1 \leq i, j \leq p$. This assumption is very natural to achieve hierarchical selection and is a well-accepted practice in statistics. See, for example, Section 1.2 of Bien, Taylor and Tibshirani [4]. However, this qualitative constraint also makes variable selection for interaction models extremely difficult. From a computational point of view, if one is content with selecting variables via the Lasso penalty as $P_e(\beta) = \sum_j |\beta_j|$, enforcing Assumption (A1) in this Lasso formulation would give a combinatorial problem that is challenging to solve. This is why a convex relaxation of (A1) such as that in Zhao, Rocha and Yu [32] is attractive and tractable. From a theoretical viewpoint, having a hierarchical structure also complicates the theoretical analysis. As will be seen later, the convex relaxations of Assumption (A1) in the literature eventually lead to overlapping group Lasso penalties, for which the familiar techniques such as those in Bickel, Ritov and Tsybakov [2] and Negahban *et al.* [17] fail to work. Finally, the main idea and analysis in this paper may also be generalized to the so-called weak hierarchy where $(i, j) \in \mathcal{S}^{(2)}$ leads to either $i \in \mathcal{S}^{(1)}$ or $j \in \mathcal{S}^{(1)}$. Since the majority of the works in the literature deal with strong hierarchy, we focus on this case in the following.

For hierarchical selection of variables that respects the strong heredity principle, many authors Zhao, Rocha and Yu [32], Radchenko and James [4] considered the penalized least squares problem

$$\hat{\beta} = \arg\min_{\theta} \left\{ \frac{1}{2n} \|\mathbb{Y} - \mathbb{Z}\theta\|_2^2 + \lambda_n P_e(\theta) \right\},$$

where $P_e(\theta)$ is a convex penalty. Let $\theta = (\theta_1, \dots, \theta_p, \theta_{12}, \dots, \theta_{(p-1)p})^T$. The general penalty used by the composite absolute penalty method (Zhao, Rocha and Yu [32], CAP) entertains

$$P_e(\theta) = \sum_{j=1}^{p} \left\| (\theta_j, \theta_{jk:\ j<k}, \theta_{kj:\ k<j}) \right\|_q + \sum_{j<k} |\theta_{jk}|, \tag{2.1}$$

where $q > 1$. The same penalty is studied in the framework of structured sparsity by Bach *et al.* [1]. A special case is thoroughly investigated by Radchenko and James [18] for $q = 2$. Bien, Taylor and Tibshirani [4] proposed a constrained Lasso formulation for hierarchical selection, which is equivalent to using the penalty

$$P_e(\theta) = \sum_{j=1}^{p} \max\left\{ |\theta_j|, \left\| (\theta_{jk:\ j<k}, \theta_{kj:\ k<j}) \right\|_1 \right\} + \sum_{j<k} |\theta_{jk}|, \tag{2.2}$$

which is similar in nature to but different in form from (2.1).

As will be shown later, $P_e(\theta)$ is upper and lower bounded by $c\|\theta\|_1$ for some constant $c > 0$. However, different amounts of penalty are applied on the terms in (2.1) or (2.2) to achieve the effect of hierarchical selection. Thus, the penalties in (2.1) and (2.2) are useful for selecting variables hierarchically, while the Lasso penalty $\|\theta\|_1$ ignores the relative importance of the main effects and interactions.

We note that Negahban *et al.* [17] developed a unified framework for M-estimator in high-dimensional setting if the regularizer is decomposable in the sense that $P_e(\theta) = P_e(\theta_A) + P_e(\theta_{A^c})$ for some index set A, where θ_A and θ_{A^c} are subvectors of θ. Decomposable regularizers encompass many useful penalties such as the ℓ_1 penalty of Lasso as special cases. As general as the decomposable assumption is, the penalties in (2.1) and (2.2) are not decomposable expect for some trivial cases. Indeed, it was pointed out by Negahban *et al.* [17] that their framework does not apply to the group Lasso with overlapping groups, a special case of which corresponds to the interaction model we study. Therefore, the conclusion of Negahban *et al.* [17] is not applicable here and new theory needs to be developed.

To study the consistency of the penalty functions defined in (2.1), and (2.2), we make the following assumption.

(A2) There exists fixed constants a_1, a_2 such that the eigenvalues of $\mathrm{cov}(Z_i)$ satisfies $0 < a_1 < \lambda_{\min,z} \le \lambda_{\max,z} < a_2 < \infty$.

Assumption (A2) is made on the population covariance of $\mathbb{Z}$, and is generally made in the literature Wang [27], Raskutti, Wainwright and Yu [19]. In particular, (A2) implies that the eigenvalues of $\mathrm{cov}(X_i)$ satisfies $0 < \lambda_{\min,x} \le \lambda_{\max,x} < \infty$ and that the diagonal elements of Σ_z, denoted by $\{h_j^2, 1 \le j \le p_1\}$ are finite, that is, $\max_j h_j \le h_0 < \infty$ for some h_0. From the proof of the theoretical results, Assumption (A2) can be relaxed to allow $\lambda_{\min,z}$ to tend to zero and $\lambda_{\max,z}$

to ∞, respectively, at the expense of slower convergence rates for the estimates. The following Proposition 2.1 shows that (A2) holds if X_i is normal.

Proposition 2.1. *If X_i follows a multivariate normal distribution with $0 < \tilde{a}_1 < \lambda_{\min,x} \leq \lambda_{\max,x} < \tilde{a}_2 < \infty$, for some fixed $\tilde{a}_1, \tilde{a}_2$, then* (A2) *holds.*

For simplicity, let $v = \hat{\beta} - \beta$ be the difference between the true β vector and the estimate obtained with penalty in (2.2) or (2.1). We are interested in the convergence rate of v when both p and n go to infinity. Denote $\|a\|_\Sigma = (a^T \Sigma_z a)^{1/2}$ for any $a \in R^{p_1}$. Let $S^{p_1-1} = \{a \in R^{p_1}: \|a\|_2 = 1\}$ be the p_1-dimensional sphere. For any vector $u \in R^{p_1}$ and any set $A \subseteq \{1, 2, \ldots, p_1\}$, we denote $u_A \in R^{p_1}$ as the vector with the jth element as u_j if $j \in A$, and 0 otherwise.

We develop the theory under the assumption that X_i and ε_i, $i = 1, \ldots, n$ are i.i.d. observations from sub-Gaussian distributions. Let

$$K_x = \|X_i\|_{\psi_2} = \sup_{\|u\|=1} \left\|u^T X_i\right\|_{\psi_2} = \sup_{\|u\|=1} \sup_{q \geq 1} q^{-1/2} \left(\mathrm{E}\left|u^T X_i\right|^q\right)^{1/q}$$

be the ψ_2 norm of vector X_i, and similarly denote $K_e = \|\varepsilon_i\|_{\psi_2}$ and $K_{xj} = \|X_{ij}\|_{\psi_2}$, $1 \leq j \leq p$. This definition of a ψ_2 norm comes from Vershynin [26]. By the basic property of sub-Gaussian distribution Vershynin [26], we have $K_{xj} \leq K_x$ for any $1 \leq j \leq p$. A summary of useful results for sub-Gaussian distributions are listed in the Appendix.

For any $0 < \delta \leq 1$, define the event

$$\mathcal{A}_0 = \left\{ \left\|\mathbb{Z}^T \varepsilon / n\right\|_\infty < C_{e,\delta} \sqrt{\log p_1 / n} \right\},$$

where $C_{e,\delta} = [(1 + \eta_0)/c]^{1/2} K_e h_0 (1 + \delta)$ for some positive constant c and η_0, is a constant depending only on K_x, K_e and δ. Now we show that event $\mathcal{A}_0$ holds with a probability close to one.

Proposition 2.2. *Suppose that X_i and ε_i, $i = 1, \ldots, n$, respectively, are i.i.d. sub-Gaussian. When $\log p_1 = \mathrm{o}(n^{1/3})$, for any $\delta \in (0, 1)$, as $n > C_1/\delta$ for some constant $C_1 > 0$, it holds that $P(\mathcal{A}_0) \geq q_{1n} \cdot q_{2n}$, where q_{1n} and q_{2n} defined in* (A.2) *and* (A.3), *tend to one as $n \to \infty$.*

The requirement $\log p_1 = \mathrm{o}(n^{1/3})$ is due to the fact that, generally, an interaction term is not sub-Gaussian any more and is heavy tailed. Therefore, a larger n is required, compared with existing results for high-dimensional linear models where only the main effects are considered and the order is typically $n > s \log p$. The requirement $\log p_1 = \mathrm{o}(n^{1/3})$ here may not be optimal. However, in this paper, we only assume that X_i is a general sub-Gaussian variable without further assumption on the distribution of X_i. The requirement on p and n can be seen as the price paid on the rate in exchange for the generality of our result.

We now provide a sufficient condition that guarantees the consistency of the penalized estimate. Specially, we consider the restricted eigenvalue (RE) condition introduced by Bickel, Ritov and Tsybakov [2].

RE condition. Assume

$$\min_{\substack{J\subseteq\{1,\dots,p_1\}\\ |J|\le s}} \min_{\|\alpha_{J^c}\|_1\le k_0\|\alpha_J\|_1} \frac{\|\mathbb{Z}\alpha\|_2}{\sqrt{n}\|\alpha_J\|_2} = M(k_0,s) > 0.$$

This version of RE condition was introduced by Bickel, Ritov and Tsybakov [2] to prove the consistency of the Lasso and the Dantzig selector in absence of the interactions. There are many different conditions of this type in the literature, imposing different constraints on the design matrix $\mathbb{X} = (X_1,\dots,X_n)^T$ for establish convergence results. Noticeably, Candes and Tao [5] introduced the uniform uncertainty principle condition on sparse eigenvalues. Zhang and Huang [31] introduced the sparse Reisz condition. The RE condition defined by Bickel, Ritov and Tsybakov [2] is weaker than the that of Candes and Tao [5]. It should also be pointed out that for both the ℓ_1 and ℓ_2 losses, strictly weaker conditions than the RE were given in van de Geer [24] and Ye and Zhang [28]. With proper parameters, the conditions in Ye and Zhang [28] include that of Bickel, Ritov and Tsybakov [2] and that of van de Geer [24] as the special case. In addition, Raskutti, Wainwright and Yu [20] also defined a similar restrictive eigenvalue condition to obtain the minimax rate high-dimensional linear regression over ℓ_q ball.

Our RE assumption can be seen as an extension of RE condition of Bickel, Ritov and Tsybakov [2] to the interaction model. For the main effects model, Raskutti, Wainwright and Yu [19] proved that the RE condition holds under the Gaussian assumption, and Rudelson and Zhou [21] showed that the RE condition still holds under the sub-Gaussian assumption. Here, we need the RE condition for variable $\mathbb{Z}$, which contains both the main effects and interaction terms and is clearly not sub-Gaussian. We give sufficient conditions for the RE condition to hold for the general case Section 3. Based on the results of Proposition 2.2 and the RE condition, we have the following on the consistency of the penalized estimators.

Theorem 2.1. *Consider the penalty defined in* (2.1) *and* (2.2) *for hierarchical selection. Suppose that* (A1), (A2) *and the conditions of Proposition* 2.2 *holds and that the RE condition holds with* $k_0 = 7$. *If* $\lambda_n \ge C_{e,\delta}\sqrt{\log p_1/n}$, *then with probability* $q_{1n}q_{2n}$, *which tends to* 1 *as* $n\to\infty$, *we have*

$$\|v\|_1 \le \lambda_n D(s) \quad \textit{and} \quad P_e(v) \le 3\lambda_n D(s),$$

where $D(s) = 16s/M^2(7,s)$.

Theorem 2.1 shows that the penalty in Zhao, Rocha and Yu [32], Radchenko and James [18] and Bien, Taylor and Tibshirani [4] designed for hierarchical variable selection all lead to consistent estimate in the ℓ_1 norm and the norm defined by $P_e(\cdot)$ under suitable conditions. In particular, the ℓ_1 norm $v = \hat{\beta} - \beta$ is of the order $s\sqrt{\log p_1/n}$, which matches the rate of convergence for the Lasso when no interaction is considered. Assumption (A1) is very important for the convergence rate of the estimate; otherwise, the estimate may get lower convergence order. To better understand the behavior of penalty functions introduced by Zhao, Rocha and Yu [32], Radchenko and James [18] and Bien, Taylor and Tibshirani [4], we consider a more general class of penalty functions. In fact it can be verified that under assumption (A1), the penalty functions defined in Zhao, Rocha and Yu [32], Radchenko and James [18] and Bien, Taylor and Tibshirani [4] are special cases of a general class of penalty functions with $L_1 = 1$ and $L_2 = 3$ satisfying the following assumption.

(A3) Suppose that $P_e(0) = 0$. For any $\theta_1, \theta_2 \in R^{p_1}$, we have $P_e(\theta_1 + \theta_2) \leq P_e(\theta_1) + P_e(\theta_2)$. For any $\theta \in R^{p_1}$, $P_e(\theta) \geq P_e(\theta_S) + L_1\|\theta_{S^c}\|_1$ with $L_1 > 1/2$ and $P_e(\theta_S) \leq L_2\|\theta_S\|_1$.

Assumption $P_e(0) = 0$ holds for all penalties designed for variable selection in the literature. By triangular inequality, the requirement $P_e(\theta_1 + \theta_2) \leq P_e(\theta_1) + P_e(\theta_2)$ holds for properly defined norms or seminorms. We can see (A3) as a weakened form of the decomposable assumption in Negahban *et al.* [17]. We now give a theorem for any penalty function that satisfies (A3).

Proposition 2.3. *Assume that* (A2), *the conditions of Proposition* 2.2 *and the RE condition hold with* $k_0 = \frac{2L_2+1}{2L_1-1}$. *Suppose the penalty* $P_e(\theta)$ *satisfies* (A3). *For the estimator* $\hat{\beta}$ *associated with penalty* $P_e(\theta)$, *we have* $\|v\|_1 \leq \lambda_n D(s)$ *with probability at least* $q_{1n}q_{2n}$, *where* $D(s) = \frac{(L_2+L_1)^2 s}{(2L_1-1)M^2(k_0,s)}$.

Following the proof of this proposition, it is not difficult to show that the rate of convergence is the same when the ℓ_1 penalty is used without enforcing the hierarchical structure in (A1). From this perspective, Assumption (A3) connects interaction selection with the ℓ_1 penalty and various hierarchical penalty functions $P_e(\theta)$. The hierarchical penalty functions typically differentiate the main and interaction terms by penalizing the main effects θ_j less and the interactions θ_{ij} more. At the same time, $P_e(\theta)$ should behave similarly to $\ell_q, q \leq 1$, that is used for variable selection. In fact, letting $\theta = \theta_{S^c}$ and $\theta_S = 0$ in (A3), we have $P_e(\theta_{S^c}) \geq L_1\|\theta_S^c\|_1$. That is, although the penalty in $P_e(\theta)$ on $\theta_j, j \notin \mathcal{S}^{(1)}$ are different from that on $\theta_{ij}, (i,j) \notin \mathcal{S}^{(2)}$, both behave similarly to the ℓ_1 penalty. On the other hand, $P_e(\theta_S) \leq L_2\|\theta_S\|_1$, indicating that the penalty in $P_e(\theta)$ on θ_S is less than or equal to that in the ℓ_1 penalty. Recall that $S = \mathcal{S}^{(1)} \cup \mathcal{S}^{(2)}$ and that (A3) does not assume any specific relationship between $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$. As $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ satisfy (A1), the penalty functions defined by Zhao, Rocha and Yu [32], Radchenko and James [18] and Bien, Taylor and Tibshirani [4] are members of the penalty functions satisfying (A3) with $L_1 = 1$ and $L_2 = 3$. Consequently, it is expected that the convergence rates using these methods are similar to that with the ℓ_1 penalty. However, a key advantage of the hierarchical penalties is that the chosen model is more likely to have a hierarchical structure, a parsimony principle often desirable in practice. We note, however, that the rate may not be improvable. To see this, consider a special scenario where the true model is sparse and only contains $s \ll p$ main effects. Even one is able to fix all the coefficients of the interactions as zero, the optimal rate of fitting a main effects model via the Lasso penalty is $s\sqrt{\log p/n}$ (Bickel, Ritov and Tsybakov [2]), the same as the rate $s\sqrt{\log p_1/n}$ obtained in this paper as $p_1 = \mathrm{O}(p^2)$ in an interaction model. Whether the rate is optimal is a direction for future research.

Motivated by Proposition 2.3, we can define new penalties such that the rate of convergence result in Theorem 2.1 is achieved. We now discuss a few concrete examples when the true model obeys the strong hierarchical property (A1).

Example 1. For $q > 1$, define $P_e(\theta)$ as

$$(p-1)^{-1} \sum_{1 \le j < k \le p} \left\{ \left\| (\theta_j, \theta_{jk}) \right\|_q + \left\| (\theta_k, \theta_{jk}) \right\|_q \right\} + \sum_{1 \le j < k \le p} |\theta_{jk}|.$$

This penalty follows the strong hierarchical principle and satisfies (A3) with $L_1 = 1$, $L_2 = 1 + (p-1)^{-1} \le 2$. Thus, Proposition 2.3 is applicable.

This penalty can be seen as an alternative to (2.1) and (2.2) for hierarchical variable selection, where the factor $p-1$ reflects the fact that every main effect appears $p-1$ times in the penalty. In particular, we construct groups (θ_j, θ_{jk}) and (θ_k, θ_{jk}) and impose $\ell_q, q > 1$ penalty to encourage sparsity in these groups. An additional penalty is imposed on the interaction as in $|\theta_{jk}|$ to encourage greater sparsity of θ_{jk}.

Example 2. Due to the connection between variables, for example, in times series analysis, we want to select variables and their interactions in a contiguous pattern. The idea of a block convex structure (Bach *et al.* [1]) is useful for such a setting. To select main effects and their interactions in the contiguous pattern of order d_0, we can consider the groups as follows. Let $T_k = (\theta_{km:\, m>k}, \theta_{mk:\, m<k})$ and construct groups as $G_j = (\theta_{j-d_0+1}, \dots, \theta_j, H_j)$ with $H_j = (T_{j-d_0+1}, \dots, T_j)$ for $j \ge d_0$. Then for $q > 1$, we define the penalty

$$P_e(\theta) = \sum_{j \ge d_0}^{p} \left(\|G_j\|_q + 2^{-1} \|H_j\|_1 \right).$$

Clearly, if $d_0 = 1$, we have the penalty defined in (2.1). For a fixed d_0, it is easy to verify that $L_1 = 1$ and $L_2 = 3d_0$.

The consideration of using this penalty comes from time series data analysis. Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_p)^T \in R^p$ be the population version of X_i. A variable $\mathbf{x}_j$ is considered important variable if both $\mathbf{x}_j$ and the variables close to $\mathbf{x}_j$ have large main effects and interaction effects. Therefore, we construct groups, which involve the main effects and interaction effects of d_0 variables that are contiguous, that is, $\mathbf{x}_{j-d_0+1}, \dots, \mathbf{x}_j$. Note that T_k consists of the interaction effect of $\mathbf{x}_k$ with all the other variables. The group G_j is the collection d_0 contiguous main effect $\theta_{j-d_0+1}, \dots, \theta_j$ and their interactions $T_{j-d_0+1}, \dots, T_j$ with all the other variable in the study. If the tuning parameter is chosen appropriately, this penalty is more likely to give a model with either contiguous sequences of variables with main effects, or contiguous sequences of variables with their main effects and interactions among those nonzero variables.

Example 3. Suppose that we want to consider the interactions of the main effects in a nested manner, that is, we want to consider the interaction of the jth variable $\mathbf{x}_j$ and the k-variable $\mathbf{x}_k$ with $k < j$ in a contiguous setting. We can define $\tilde{G}_j = (\theta_j, \tilde{H}_j)$, where $\tilde{H}_j = (\theta_{kj:\, k<j})$ for $1 \le j \le p$. For $q > 1$, we can define

$$P_e(\theta) = \sum_{j=1}^{p} \left(\|\tilde{G}_j\|_q + \|\tilde{H}_j\|_1 \right).$$

For this penalty satisfying the strong hierarchical principle, it is easy to verify that (A3) holds with $L_1 = 1$ and $L_2 = 2$. Thus, the conclusion of Proposition 2.3 and that of Theorem 2.1 hold.

If the tuning parameter is chosen appropriately, this penalty would give a model consisting of θ_j and all (or some) of its interactions with variables $\mathbf{x}_k$ as θ_{kj} for $k < j$. More precisely, this penalty would give models with main effects and its interactions with variables that have smaller indices. This is interesting when variables are organized along time. When a variable has no effect, all its interactions with preceding variables have zero effect but its interactions with later variables may be nonzero.

3. The RE condition

In our setting, variable Z_i, containing both the main and interaction terms of X_i, is not sub-Gaussian if X_i follows multivariate normal. Thus, we need to develop sufficient conditions for the RE condition to hold. The previous results in the literature assuming sub-Gaussianity are not applicable here.

To simplify our proof, we first introduce a lemma. Let X, X^*, Z be the population version of X_i, X_i^*, Z_i. We denote $Z - E(Z) = (\tilde{X}^T, \tilde{Z}^T)^T$, where $\tilde{X} = X - E(X)$ and $\tilde{Z} = X^* - E(X^*)$ and for any $u \in S^{p_1-1}$, $u = (u^{(1)T}, u^{(2)T})^T$ with $u^{(1)} \in R^p$ and $u^{(2)} = (w_{12}, \dots, w_{1p}, w_{23}, \dots, w_{2p}, \dots, w_{(p-1)p})^T \in R^{p_1-p}$. Define $w_{ji} = w_{ij}$ for $i < j$ and $w_{ii} = 0$, $i = 1, \dots, p$ and construct the symmetric matrix $W = (w_{ij})$. It is easy to see that $Z^T u = X^T u^{(1)} + 2^{-1} X^T W X$. For any $u \in S^{p_1-1}$, we distinguish three cases: (i) $u^{(1)} = 0, u^{(2)} \neq 0$, (ii) $u^{(1)} \neq 0, u^{(2)} = 0$ and (iii) $u^{(1)} \neq 0, u^{(2)} \neq 0$. For u in case (iii), we define the correlation coefficient as $\rho_u = \operatorname{corr}(X^T u^{(1)}, X^T W X)$, and for that in (i) and (ii), $\rho_u = 0$. Let $X_W^* = [\operatorname{var}(X^T W X)]^{-1/2} X^T W X$ for case (i) and (iii), and $X_W^* = 0$ for case (ii). We make the following assumptions.

(A4) We assume $\sup_{u \in S^{p_1-1}} |\rho_u| \le \bar{\rho} < 1$ for some $\bar{\rho}$.

(A5) The random vector X_W^* is subexponential with ψ_1 norm K_u and there exists K_0, such that $\sup_{u \in S^{p_1-1}} K_u \le K_0 < \infty$.

It can be verified that (A4) and (A5) hold when X is multivariate normal. In this case, for any $u \in S^{p_1-1}$, $X^T u^{(1)}$ is normal and $X^T W X$ follows a weighted χ^2 distribution; thus the correlation ρ_u is strictly less than 1. Due to the closeness of the set S^{p_1-1}, we have $\sup_{u \in S^{p_1-1}} |\rho_u| \le \bar{\rho} < 1$. Particularly, as $X \sim N(0, \Sigma_x)$, we have $\rho_u = 0$ for any $u \in S^{p_1-1}$ due to the symmetry. In fact, we have shown in the proof of Proposition 2.1 that (A4) holds when X is continuous with $\lambda_{\min,x} > 0$. Since X is normal, by Magnus and Neudecker [14], we have $\operatorname{var}(X^T W X) = \operatorname{tr}(W\Sigma_x)^2 = \|W\Sigma_x\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm. Let $W_{\Sigma_x} = \Sigma_x^{1/2} W \Sigma_x^{1/2}$ with associated eigenvalue decomposition $W_{\Sigma_x} = U \Upsilon U^T$ where $\Upsilon = \operatorname{diag}(\tau_1, \dots, \tau_p)$ are the eigenvalues and the columns of U are eigenvectors. Then $Y = U^T \Sigma_x^{-1/2} X \sim N(0, I_p)$. Then we have

$$X_W^* = \|W\Sigma_x\|_F^{-1} X^T W X = \|W\Sigma_x\|_F^{-1} Y^T \Upsilon Y = \|W\Sigma_x\|_F^{-1} \sum_{i=1}^{p} \tau_i Y_i^2.$$

Note that, for $i = 1, \dots, n$, Y_i^2's are independent χ_1^2 that is subexponential with the ψ_1 norm $\|Y_i^2\|_{\psi_1} \le 2\|Y_i\|_{\psi_2} \le 2K_x$, and that $\sum \tau_i^2 / \|W\Sigma_x\|_F^2 = 1$ for any $u \in S^{p_1-1}$. By Vershynin

[26] (see Proposition C.1 in Appendix C), for any $u \in S^{p_1-1}$, X_W^* is subexponential with $K_u = \|Y_i^2\|_{\psi_1}$. Therefore, it is sufficient to take $K_0 = \|Y_i^2\|_{\psi_1}$. We have the following tail bounds.

Lemma 3.1. *Let* $Z_{u,\Sigma} = [Z - E(Z)]^T u \|u\|_\Sigma^{-1}$. *Suppose that* (A4)–(A5) *hold. Then for any* $t > 0$,

$$\sup_{u \in S^{p_1-1}} P\big(|Z_{u,\Sigma}| > t\big) \le C \exp\big[-\min\big(t^2/\tilde{K}_0, t/K_0\big)\big],$$

where K_0, $\tilde{K}_0$ *are constant depend only on* K_x. *Furthermore, it holds that*

$$\sup_{u \in S^{p_1-1}} \operatorname{var}\big[\big(Z^T u\big)^2\big] < \infty.$$

Theorem 3.1. *Suppose that X is sub-Gaussian and the assumptions of Lemma 3.1 hold. For any* $0 < \epsilon < \lambda_{\min,z}^{1/2}(8\lambda_{\max,z}^{1/2} + \lambda_{\min,z}^{1/2})^{-1}$, *if*

$$n > \epsilon^{-1} \max\big\{C_1, \big(c_1 m_1 \tilde{C}_K^{-1} \log(c_1 m_1 p_1)\big)^3, \big(4^{-1} \tilde{C}_K^{-1} c_1 m_1 \log(4 c_1 m_1 p_1)\big)^3\big\},$$

where $m_1 = 16s(1 + k_0)^2$ *and* $c_1, C_1, \tilde{C}_K$ *are positive constants with* $\tilde{C}_K < C_K$, *then the RE condition holds with a high probability that is at least* $\tilde{p}_{0n}$ *where* $\tilde{p}_{0n}$ *defined in* (A.15) *tends to* 1, *as n goes to infinity.*

When there is no interaction and X is sub-Gaussian, we can show that the order on the right-hand side of the inequality in Lemma B.4 can be improved to $\exp(-C_K n\delta)$. As a result, the lower bound of the sample size n in Theorem 3.1 can be improved to a smaller order $\epsilon^{-1} \max\{C_1, c_1 m_1 \tilde{C}_K^{-1} \log(c_1 m_1 p), 4^{-1} \tilde{C}_K^{-1} c_1 m_1 \log(4 c_1 m_1 p)\}$. This observation again reflects that selecting variables in interaction models is more difficult.

Appendix A: Proofs of the main results

We provide the proofs for the theorems. A few additional lemmas that are needed are given in Appendix B.

A.1. Proof of Proposition 2.1

We use the notation defined in Section 3. Recall that $Z = (X^T, X^{*T})^T$. For any $u \in S^{p_1-1}$, $u = (u^{(1)T}, u^{(2)T})^T$ with $u^{(1)} \in R^p$ and $u^{(2)} = (w_{12}, \dots, w_{1p}, w_{23}, \dots, w_{2p}, \dots, w_{(p-1)p})^T \in R^{p_1-p}$. Define the symmetric matrix $W = (w_{ij})$ based on $u^{(2)}$ as in Section 3. It is easy to see that $Z^T u = X^T u^{(1)} + 2^{-1} X^T W X$. Then $\lambda_{\max,z} = \sup_{u \in S^{p_1-1}} \operatorname{cov}(Z^T u)$ and $\lambda_{\min,z} = \inf_{u \in S^{p_1-1}} \operatorname{cov}(Z^T u)$.

First, we consider $\lambda_{\max,z}$. It holds that

$$\lambda_{\max,z} \le \sup_{u \in S^{p_1-1}} 2\big[\operatorname{var}\big(X^T u^{(1)}\big) + \operatorname{var}\big(2^{-1} X^T W X\big)\big].$$

Since X is normal, from Magnus and Neudecker [14], we have $\mathrm{var}(X^T WX) = \mathrm{tr}(W\Sigma_x)^2 = \|W\Sigma_x\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm. For symmetric positive semidefinite matrices A and B, by Fang, Loparo and Feng [8], we have that $\lambda_{\min,A}\,\mathrm{tr}(B) \le \mathrm{tr}(AB) \le \lambda_{\max,A}\,\mathrm{tr}(B)$. Since W and Σ_x are symmetric matrices, we have

$$\mathrm{tr}(W^2)\lambda_{\min,x}^2 \le \mathrm{var}(X^T WX) = \mathrm{tr}(W^2\Sigma_x^2) \le \mathrm{tr}(W^2)\lambda_{\max,x}^2. \tag{A.1}$$

Moreover, we have $\mathrm{var}(Z^T u^{(1)}) \le \lambda_{\max,x}\|u^{(1)}\|_2^2$. Also, note that $\mathrm{tr}(W^2) + \|u^{(1)}\|_2^2 = \|u\|_2^2 = 1$. Then we have

$$\begin{aligned}\lambda_{\max,z} &\le 2 \sup_{u\in S^{p_1-1}} \left[\lambda_{\max,x}\|u^{(1)}\|_2^2 + 4^{-1}\mathrm{tr}(W^2)\lambda_{\max,x}^2\right] \\ &\le 2\max\{\lambda_{\max,x}, \lambda_{\max,x}^2\} \le 2\max\{\tilde{a}_2, \tilde{a}_2^2\}.\end{aligned}$$

Second, we consider $\lambda_{\min,z}$. Note that $\lambda_{\min,x} > \tilde{a}_1 > 0$. Then X is nondegenerate in any direction. For any $u^{(1)} \ne 0, u^{(2)} \ne 0$, we have $\mathrm{var}(X^T u^{(1)}) > 0$ and $\mathrm{var}(X^T WX) > 0$. Recall the notation $\rho_u = \mathrm{corr}(X^T u^{(1)}, X^T WX)$ in Section 3 which is defined as 0 when $u^{(1)} = 0$ or $u^{(2)} = 0$. For any $u \in S^{p_1-1}$ with $u^{(1)} \ne 0$ and $u^{(2)} \ne 0$, we have given a simple argument in Section 3 that, when X is normal, $\sup_{u\in S^{p_1-1}} \rho_u \le \bar{\rho} < 1$, for some $\bar{\rho}$. Here, we show a more general result that if X is continuous with $\lambda_{\min,x} > 0$, this conclusion still holds.

Suppose that on the contrary we have $\sup_{u\in S^{p_1-1}} \rho_u = 1$. Then due to the closeness of S^{p_1-1}, we have for some $u_0 = (u_0^{(1)T}, u_0^{(2)T})^T$ with $u_0^{(i)} \ne 0, i = 1, 2$, such that $\rho_{u_0} = 1$. Let $W_0 = (w_{ij})$ be the matrix constructed from $u_0^{(2)}$. Then there are some constants a_0 and a_1 in R such that $X^T u_0^{(1)} = a_1(X^T W_0 X) + a_0$ holds with probability 1. Since $u_0^{(1)} \ne 0$ and $\lambda_{\min,x} > 0$, we have $a_1 \ne 0$. Since W_0 is a symmetric matrix, its eigenvalue decomposition is denoted as $W_0 = U\Lambda U^T$, where $\lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ is a diagonal matrix of eigenvalues and U consists of the eigenvectors. Let $u_0^{(1)} = Ub$ for some $b \in R^p$. Then $V = U^T X := (\mathbf{v}_1, \dots, \mathbf{v}_p)^T$ is a nondegenerate variable in R^p, since $\mathrm{cov}(V) = \mathrm{cov}(X)$ and $\lambda_{\min,x} > 0$. Therefore, we have with probability one,

$$a_1^{-1}(V^T b - a_0) = V^T\Lambda V = \sum_{j=1}^p \lambda_j \mathbf{v}_j^2.$$

Due to $u^{(2)} \ne 0$, we have $W_0 \ne 0$, and consequently at least one of $\lambda_i, 1 \le i \le p$, is nonzero. Therefore, the right-hand side is a hyperquadric and the left-hand side is a hyperplane. Since the set of intersection of hyperquadric and hyperplane has Lebesgue measure 0 and V is a continuous and nondegenerate variable in R^p, the equation holds with probability 0. This leads to contradiction.

Furthermore, by the inequality $\sqrt{ab} \le 2^{-1}(a + b)$ for any $a > 0, b > 0$ and definition of correlation, we have

$$\begin{aligned}\mathrm{cov}(Z^T u) &\ge \mathrm{var}(X^T u^{(1)}) + \mathrm{var}(2^{-1}X^T WX) - 2\left|\mathrm{cov}(Z^T u^{(1)}, 2^{-1}X^T WX)\right| \\ &\ge (1-\bar{\rho})\left[\mathrm{var}(Z^T u^{(1)}) + \mathrm{var}(2^{-1}X^T WX)\right] \\ &\ge (1-\bar{\rho})\left[\lambda_{\min,x}\|u^{(1)}\|_2^2 + 4^{-1}\mathrm{tr}(W^2)\lambda_{\min,x}^2\right],\end{aligned}$$

where we use (A.1) in the third inequality. Consequently, it holds that

$$\lambda_{\min,z} = \inf_{u\in S^{p_1-1}} \operatorname{cov}\big(Z^T u\big) \geq 4^{-1}(1-\bar{\rho})\min\big(\lambda_{\min,x}, \lambda_{\min,x}^2\big) \geq 4^{-1}(1-\bar{\rho})\min\big(\tilde{a}_1, \tilde{a}_1^2\big),$$

where we use the fact $\|u^{(1)}\|_2^2 + \operatorname{tr}(W^2) = \|u\|_2^2 = 1$. This completes the proof.

A.2. Proof of Proposition 2.2

Recall that $E(Z_{ij}) = 0$ and $h_j^2 = \operatorname{var}(Z_{ij}) \leq h_0^2 < \infty$, for $1 \leq j \leq p_1$. Define $\mathcal{T} = \{\mathbb{Z}\colon |(\sqrt{n} \times h_j)^{-1}\|\mathbb{Z}_j\|_2 - 1| < \delta$, for all $j = 1, \dots, p_1\}$. We now consider the probability of event $\mathcal{T}$. In fact, let e_j, $j = 1, \dots, p_1$ be the vector with the jth element 1 and 0 elsewhere. Note that for any $1 \leq j_1 < j_2 \leq p$, X_{ij_1} and X_{ij_2} are sub-Gaussian with ψ_2 norm less than K_x. By Vershynin [26], $X_{ij_1}X_{ij_2}$ is subexponential with ψ_1 norm no more than $2K_x^2$ (see also Appendix C). By the results of Lemma B.4, as $n > C_1/\delta$, we have

$$\begin{aligned} P(\mathcal{T}) &= 1 - P\big(\mathcal{T}^c\big) \\ &\geq 1 - \sum_{j=1}^{p_1} P\big(\big|n^{-1}h_j^{-2}\|\mathbb{Z}e_j\|_2^2 - 1\big| \geq \delta\big) \\ &\geq 1 - 2p_1 \exp\big(-C_K(n\delta)^{1/3}\big) \\ &= 1 - 2\exp\big[-C_K(n\delta)^{1/3} + \log p_1\big] := q_{1n}, \end{aligned} \tag{A.2}$$

where C_K depends only on K_x and h_0. It is clear that $q_{1n} \to 1$, when $n \to \infty$ due to the assumption $\log p_1 = o(n^{1/3})$. On the other hand, letting $V_j = \mathbb{Z}_j^T \varepsilon$, $j = 1, \dots, p_1$, we have

$$P\big(\big\|\mathbb{Z}^T\varepsilon/n\big\|_\infty > t|\mathcal{T}\big) = P\Big(\max_j |V_j| > nt|\mathcal{T}\Big) \leq p_1 \max_j P\big(|V_j| > nt|\mathcal{T}\big).$$

By the independence of Z_i and ε_i, we have that $\mathbb{Z}_j^T\varepsilon|\mathbb{Z}_j$ is also sub-Gaussian with ψ_2 norm $\tilde{K}_{j,e} = C\|\mathbb{Z}_j\|K_e$ for some constant $C > 0$. Then we have

$$\max_j P\big(|V_j| > nt|\mathbb{Z}_j\big) \leq \max_j e\exp\bigg(-\frac{cn^2t^2}{K_e^2\|\mathbb{Z}_j\|_2^2}\bigg),$$

where $c > 0$ is constant. Moreover, we have

$$\max_j P\big(|V_j| > nt|\mathcal{T}\big) \leq e\exp\bigg(-\frac{ct^2n}{K_e^2(1+\delta)^2h_0^2}\bigg).$$

Consequently, we have

$$P\big(\big\|\mathbb{Z}^T\varepsilon/n\big\|_\infty > t|\mathcal{T}\big) \leq p_1 e\exp\bigg(-\frac{ct^2n}{K_e^2(1+\delta)^2h_0^2}\bigg).$$

Taking $t = C_{e,\delta}\sqrt{\log p_1/n}$, with $C_{e,\delta} = [(1+\eta_0)/c]^{1/2} K_e h_0(1+\delta)$ for any $\eta_0 > 0$, we have

$$P\left(\left\|\mathbb{Z}^T \varepsilon/n\right\|_\infty < C_{e,\delta}\sqrt{\log p_1/n}|\mathcal{T}\right) \geq 1 - p_1^{-\eta_0} := q_{2n}, \tag{A.3}$$

where $q_{2n} \to 1$ with $n \to \infty$. Therefore, it holds that

$$P\left(\left\|\mathbb{Z}^T \varepsilon/n\right\|_\infty \leq C_{e,\delta}\sqrt{\log p_1/n}\right) > q_{1n}q_{2n}.$$

The proof is completed.

A.3. Proof of Theorem 2.1

Step 1. For the penalty function defined in (2.1) and (2.2), we show that under (A1), assumption (A3) holds with $L_1 = 1$ and $L_2 = 3$. We take the penalty in (2.1) as example for illustration. The proof for penalty in (2.2) can be checked similarly.

For any $j \notin \mathcal{S}^{(1)}$, by Assumption (A1), pair $(j,k) \notin \mathcal{S}^{(2)}$ for all $j \neq k$. Consequently,

$$\left[|\theta_j|^q + \sum_{k:\ j<k} |\theta_{jk}|^q + \sum_{k:\ k<j} |\theta_{kj}|^q\right]^{1/q} + \sum_{k:\ j<k} |\theta_{jk}| \geq |\theta_j| + \sum_{k:\ j<k} |\theta_{jk}|. \tag{A.4}$$

And for any $j \in \mathcal{S}^{(1)}$,

$$\begin{aligned}
&\left[|\theta_j|^q + \sum_{k:\ j<k} |\theta_{jk}|^q + \sum_{k:\ k<j} |\theta_{kj}|^q\right]^{1/q} + \sum_{k:\ j<k} |\theta_{ij}| \\
&\quad \geq \left[|\theta_j|^q + \sum_{\substack{k:\ (j,k)\in\mathcal{S}^{(2)} \\ j<k}} |\theta_{jk}|^q + \sum_{\substack{k:\ (j,k)\in\mathcal{S}^{(2)} \\ k<j}} |\theta_{kj}|^q\right]^{1/q} \\
&\quad\quad + \sum_{\substack{k:\ (j,k)\in\mathcal{S}^{(2)} \\ j<k}} |\theta_{jk}| + \sum_{\substack{k:\ (j,k)\notin\mathcal{S}^{(2)} \\ j<k}} |\theta_{jk}|.
\end{aligned} \tag{A.5}$$

It is easy to see that sum of the left-hand side in (A.4) and (A.5) equals $P_e(\theta)$ and that of the right-hand sides equals $P_e(\theta_S) + \|\theta_{S^c}\|_1$. Therefore, $P_e(\theta) \geq P_e(\theta_S) + \|\theta_{S^c}\|_1$, that is $L_1 = 1$. Noting the fact that for $q > 1$, $\|\alpha\|_q \leq \|\alpha\|_1$ for any vector α, we can see

$$\begin{aligned}
P_e(\theta) &= \sum_j \left[|\theta_j|^q + \sum_{j<k} |\theta_{jk}|^q + \sum_{k<j} |\theta_{kj}|^q\right]^{1/q} + \sum_{j<k} |\theta_{jk}| \\
&\leq \sum_j \left[|\theta_j| + \sum_{j<k} |\theta_{jk}| + \sum_{k<j} |\theta_{kj}|\right] + \sum_{j<k} |\theta_{jk}| \\
&= \sum_j |\theta_j| + 3\sum_{j<k} |\theta_{jk}| \leq 3\|\theta\|_1.
\end{aligned}$$

Therefore, $L_2 = 3$.

Step 2. Note that the penalty function defined in (2.1) and (2.2) are special cases of the function class satisfying (A3). Consequently, under (A1), (A2) and the RE condition with $k_0 = (2L_2 + 1)/(2L_1 - 1) = 7$, by Proposition 2.3, we have $\|v\|_1 \le \lambda_n D(s)$ with probability $p_n q_n$, tending to 1 as $n \to \infty$. In addition, due to $P_e(\theta) \le 3\|\theta\|_1$ in Step 1, we have $P_e(v) \le 3\lambda_n D(s)$. This completes the proof.

A.4. Proof of Proposition 2.3

Step 1. By the definition of $\hat{\beta}$ and convexity of $P_e(\theta)$, we have

$$\frac{1}{2n}\|\mathbb{Y} - \mathbb{Z}\hat{\beta}\|_2^2 + \lambda_n P_e(\hat{\beta}) \le \frac{1}{2n}\|\mathbb{Y} - \mathbb{Z}\beta\|_2^2 + \lambda_n P_e(\beta).$$

Conditioning on the set $\mathcal{A}_0$, taking $\lambda_n > 2C_{e,\delta}\sqrt{\log p_1/n}$, and noting that $\hat{\beta} = \hat{\beta}_S + v_{S^c}$, we have

$$\begin{aligned}
\frac{1}{n}\|\mathbb{Z}v\|_2^2 &\le 2\lambda_n P_e(\beta) - 2\lambda_n P_e(\hat{\beta}) + 2\frac{v^T \mathbb{Z}^T \varepsilon}{n} \\
&\le 2\lambda_n P_e(\beta) - 2\lambda_n P_e(\hat{\beta}) + 2\left\|\frac{\mathbb{Z}^T\varepsilon}{n}\right\|_\infty \|v\|_1 \\
&\le 2\lambda_n\big[P_e(\beta) - P_e(\hat{\beta})\big] + \lambda_n\|v\|_1.
\end{aligned}$$

And consequently,

$$\begin{aligned}
\frac{1}{n}\|\mathbb{Z}v\|_2^2 &\le 2\lambda_n\big[P_e(\beta) - P_e(\hat{\beta}_S) - L_1\|v_{S^c}\|_1\big] + \lambda_n\|v\|_1 \\
&\le 2\lambda_n\big[P_e(-v_S) - L_1\|v_{S^c}\|_1\big] + \lambda_n\|v\|_1 \\
&\le 2\lambda_n[L_2\|v_S\|_1 - L_1\|v_{S^c}\|_1] + \lambda_n\big(\|v_S\|_1 + \|v_{S^c}\|_1\big) \\
&= \lambda_n\big[(2L_2 + 1)\|v_S\|_1 - (2L_1 - 1)\|v_{S^c}\|_1\big].
\end{aligned} \tag{A.6}$$

Thus, we have

$$\|v_{S^c}\|_1 \le \frac{2L_2 + 1}{2L_1 - 1}\|v_S\|_1 := k_0\|v_S\|_1.$$

Recall that the RE condition holds with $M(k_0, s) > 0$. Conditioning on $\mathcal{A}_0$ and this condition, by (A.6), we have

$$\begin{aligned}
\|\mathbb{Z}v\|_2^2/n + \lambda_n(2L_1 - 1)\|v\|_1 &\le 2\lambda_n(L_2 + L_1)\|v_S\|_1 \\
&\le 2\lambda_n(L_2 + L_1)\sqrt{s}\|v_S\|_2 \\
&\le 2\lambda_n(L_2 + L_1)\sqrt{s}\frac{\|\mathbb{Z}v\|_2/\sqrt{n}}{M(k_0, s)} \\
&\le \frac{\lambda_n^2(L_2 + L_1)^2 s}{M^2(k_0, s)} + \left(\frac{\|\mathbb{Z}v\|_2}{\sqrt{n}}\right)^2.
\end{aligned}$$

Therefore, we have

$$\lambda_n(2L_1-1)\|v\|_1 \le \frac{\lambda_n^2(L_2+L_1)^2 s}{M^2(k_0,s)},$$

and

$$\|v\|_1 \le \lambda_n \frac{(L_2+L_1)^2 s}{(2L_1-1)M^2(k_0,s)} := \lambda_n D(s).$$

A.5. Proof of Lemma 3.1

The result $\sup_{u\in S^{p_1-1}} \mathrm{var}[(Z_i^T u)^2] < \infty$ follows easily from the inequality in Lemma 3.1 and assumption $\lambda_{\max,z} < \infty$ in (A2). We only need to prove the main inequality in the lemma.

Step 1. We have $[Z - E(Z)]^T u = \tilde{X}^T u^{(1)} + \tilde{Z}^T u^{(2)}$. Then by the definition of W, it is easy to see $\tilde{Z}^T u^{(2)} = 2^{-1}[X^T W X - E(X^T W X)]$. Then

$$\begin{aligned} D_u &:= P\big(|Z_{u,\Sigma}| > t\big) \\ &= P\big(\|u\|_\Sigma^{-1}\big|\tilde{X}^T u^{(1)} + \tilde{Z}^T u^{(2)}\big| > t\big) \\ &\le P\big(\big|\tilde{X}^T u^{(1)}\big| > t\cdot 2^{-1}\|u\|_\Sigma\big) + P\big(\big|\tilde{Z}^T u^{(2)}\big| > t\cdot 2^{-1}\|u\|_\Sigma\big) \\ &:= D_{1,u} + D_{2,u}. \end{aligned}$$

If $u^{(1)} = 0$, it is clear that $D_{1,u} = 0$. If $u^{(1)} \neq 0$, due to $\|u^{(1)}\|_2 \le \|u\|_2$, it follows that

$$\begin{aligned} D_{1,u} &= P\big(\big|\tilde{X}^T u^{(1)}\big|\big\|u^{(1)}\big\|_2^{-1} > t\cdot 2^{-1}\|u\|_\Sigma \big\|u^{(1)}\big\|_2^{-1}\big) \\ &\le P\big(\big|\tilde{X}^T u^{(1)}\big|\big\|u^{(1)}\big\|_2^{-1} > t\cdot 2^{-1}\|u\|_\Sigma \|u\|_2^{-1}\big) \\ &\le P\big(\big|\tilde{X}^T u^{(1)}\big|\big\|u^{(1)}\big\|_2^{-1} > t\cdot 2^{-1}\lambda_{\min,z}^{1/2}\big). \end{aligned}$$

Note that $\tilde{X}$ is sub-Gaussian with ψ_2 norm $K_x < \infty$. Then for any $u^{(1)} \neq 0$, $|X^T u^{(1)}|\|u^{(1)}\|_2^{-1}$ is sub-Gaussian with ψ_2 norm less than K_x. Therefore, since $\lambda_{\min,z}$ is finite, we have

$$\sup_{u\in S^{p_1-1}} D_{1,u} \le \exp\big(-t^2/\tilde{K}_x^2\big),$$

where $\tilde{K}_x$ is a constant depending only on K_x.

Step 2. We show the exponential rate of $\sup_{u\in S^{p_1-1}} D_{2,u}$ and the final conclusion.

Step 2.1. We first show that $\sup_{u\in S^{p_1-1}} D_W \le L_0 < \infty$, where $D_W = [\mathrm{var}(X^T W X)]^{1/2}$. If $u^{(2)} = 0$, $\mathrm{var}(X^T W X) = 0$. We only need to consider u satisfying $u^{(2)} \neq 0$. Recall that $\|u\|_\Sigma = (u^T \Sigma_z u)^{1/2}$. It follows that $\|u\|_\Sigma^2 = u^T \Sigma_z u = \mathrm{var}(X^T u^{(1)} + 2^{-1} X^T W X) \le \lambda_{\max,z} < \infty$. That is,

$$\lambda_{\max,z} \ge \mathrm{var}\big(X^T u^{(1)} + 2^{-1} X^T W X\big) = A_u^2 + B_u^2 + 2\rho_u A_u B_u,$$

where $A_u^2 = \mathrm{var}(X^T u^{(1)})$ and $B_u^2 = \mathrm{var}(2^{-1} X^T W X)$. It is sufficient to show that $\sup_{u \in S^{p_1-1}} B_u$ is bounded. Suppose that, on the contrary, there exists series $\{u_n, n \geq 1\}$, such that $B_{u_n} \to \infty$; for simplicity we denote B_{u_n} as B_n, similarly we define A_n, ρ_n. It is easy to see

$$\lambda_{\max,z} \geq A_n^2 + B_n^2 + 2\rho_n A_n B_n \geq 2(1+\rho_n) A_n B_n \geq 2(1-\bar{\rho}) A_n B_n. \tag{A.7}$$

Consequently, $A_n \to 0$ and $\sup_n |\rho_n A_n B_n| \leq \lambda_{\max,z} \bar{\rho}/[2(1-\bar{\rho})]$ due to (A4). Taking $n \to \infty$, we have $A_n^2 + B_n^2 + 2\rho_n A_n B_n \to \infty$. This contradicts with the first inequality in (A.7). Consequently, $\sup_{u \in S^{p_1-1}} B_u$ is bounded.

Step 2.2. First, for any $u \in S^{p_1-1}$, we have $\|u\|_\Sigma \geq \lambda_{\min,z}^{1/2} > 0$. Noting that $\tilde{Z}^T u^{(2)} = 2^{-1}(X^T W X - E(X^T W X))$, we have

$$\begin{aligned} D_{2,u} &= P\left(D_W^{-1} \left|\tilde{Z}^T u^{(2)}\right| > t \cdot 2^{-1} \|u\|_\Sigma D_W^{-1}\right) \\ &= P\left(\left|X_W^* - E\left(X_W^*\right)\right| > t \|u\|_\Sigma D_W^{-1}\right) \\ &\leq P\left(\left|X_W^* - E\left(X_W^*\right)\right| > t \cdot \lambda_{\min,z}^{1/2} L_0^{-1}\right). \end{aligned} \tag{A.8}$$

Since X_W^* is subexponential with $K_u \leq K_0$, by Vershynin [26], we have

$$\sup_{u \in S^{p_1-1}} D_{2,u} \leq 2 \exp\left[-c \min\left(t^2/K_0^2, t/K_0\right)\right],$$

where c is an absolute constant depending on L_0 and $\lambda_{\min,z}$. We now prove the final conclusion. From Steps 1 and 2, we easily have

$$\begin{aligned} \sup_{u \in S^{p_1-1}} D_u &\leq \exp\left(-t^2/\tilde{K}_x^2\right) + 2\exp\left[-c \min\left(t^2/K_0^2, t/K_0\right)\right] \\ &\leq C \exp\left[-\min\left(t^2/\tilde{K}_0, t/K_0\right)\right], \end{aligned}$$

where $\tilde{K}_0 = \max(\tilde{K}_x, K_0)$.

A.6. Proof of Theorem 3.1

Step 1. We first prove that the RE condition holds under the event $\mathcal{A}_1 \cap \mathcal{A}_2$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined below. Define the event

$$\mathbb{B}_1 = \left\{x\colon\ x \in R^{p_1}, \exists T,\ \text{with } |T| \leq s,\ \text{such that } \|x_{T^c}\|_1 \leq k_0 \|x_T\|_1\right\} \cap S^{p_1-1},$$

where $p_1 = p(p+1)/2$ and S^{p_1-1} is the p_1-dimensional unit sphere.

Due to the fact that $\frac{\|\mathbb{Z}x\|_2}{\sqrt{n}\|x_S\|_2} \geq \frac{\|\mathbb{Z}x\|_2}{\sqrt{n}\|x\|_2}$, for any $S \subseteq \{1, \ldots, p\}$, we only need to show that

$$\min_{|J| \leq s} \min_{\|x_{J^c}\|_1 \leq k_0 \|x_J\|_1} \frac{\|\mathbb{Z}x\|_2}{\sqrt{n}\|x\|_2} > 0.$$

This is equivalent to show

$$\min_{x\in\mathbb{B}_1} \frac{\|\mathbb{Z}x\|_2}{\sqrt{n}} > 0.$$

Based on the results of Lemma B.3 and Λ_1, Λ_2 defined there, we have $\mathbb{B}_1 \subseteq 4\operatorname{conv}\Lambda_1$, where $4\operatorname{conv}\Lambda_1 = \{4x\colon x \in \operatorname{conv}\Lambda_1\}$ and $\operatorname{conv}\Lambda_1$ for the finite set Λ_1 denotes the convex hull of set Λ_1. Consider the events

$$\mathcal{A}_1 = \left\{ \left| \frac{\|\mathbb{Z}x_0\|_2}{\sqrt{n}\|x_0\|_\Sigma} - 1 \right| \le \epsilon \text{ for all } x_0 \in \Lambda_1 \right\},$$

and

$$\mathcal{A}_2 = \left\{ \left| \frac{\|\mathbb{Z}x\|_2}{\sqrt{n}\|x\|_\Sigma} - 1 \right| \le \epsilon \text{ for all } x \in \Lambda_2 \right\},$$

where $\|x_0\|_\Sigma = (x_0^T \Sigma x_0)^{1/2}$ and $\Lambda_1 \subseteq S^{p_1-1}$, $\Lambda_2 \subseteq B_2^{p_1}$ are defined in Lemma B.3. Conditioning on $\mathcal{A}_2$, we have

$$\frac{\|\mathbb{Z}x\|_2}{\sqrt{n}\|x\|_2} \le \frac{\|x\|_\Sigma}{\|x\|_2}(1+\epsilon) \le \lambda_{\max,z}^{1/2}(1+\epsilon) \qquad \text{for any } x \in \Lambda_2.$$

Consequently,

$$\|\mathbb{Z}x\|_2/\sqrt{n} \le \lambda_{\max,z}^{1/2}(1+\epsilon)\|x\|_2 \qquad \text{for any } x \in \Lambda_2. \tag{A.9}$$

By the results of Lemma B.3, for any $x \in \mathbb{B}_1$, there exists $x_0 \in \Lambda_1$ such that $\|x - x_0\|_2 < \epsilon$. Therefore, we have

$$\|\mathbb{Z}x_0\|_2/\sqrt{n} - \|\mathbb{Z}(x-x_0)\|_2/\sqrt{n} \le \|\mathbb{Z}x\|_2/\sqrt{n} \le \|\mathbb{Z}x_0\|_2/\sqrt{n} + \|\mathbb{Z}(x-x_0)\|_2/\sqrt{n}. \tag{A.10}$$

Note that $x - x_0 \in (\mathbb{B}_1 - \mathbb{B}_1) \cap \epsilon B_2^{p_1}$ and recall from Lemma B.3 that $(\mathbb{B}_1 - \mathbb{B}_1) \cap \epsilon B_2^{p_1} \subseteq 4\operatorname{conv}(\Lambda_2)$. Combining with the definition of Λ_2 and (A.9), we have

$$\|\mathbb{Z}(x-x_0)\|_2/\sqrt{n} \le 4 \sup_{z\in\Lambda_2} \|\mathbb{Z}z\|_2/\sqrt{n} \le 4\lambda_{\max,z}^{1/2}(1+\epsilon)\|z\|_2 \le 4\lambda_{\max,z}^{1/2}(1+\epsilon)\epsilon. \tag{A.11}$$

In addition, conditioning on $\mathcal{A}_1$, we have

$$\|x_0\|_\Sigma(1-\epsilon) < \|\mathbb{Z}x_0\|_2/\sqrt{n} \le \|x_0\|_\Sigma(1+\epsilon). \tag{A.12}$$

There, conditioning on $\mathcal{A}_1 \cap \mathcal{A}_2$ and combining (A.10)–(A.12), as $0 < \epsilon < \lambda_{\min,z}^{1/2}(8\lambda_{\max,z}^{1/2} + \lambda_{\min,z}^{1/2})^{-1}$, we have, for any $x \in \mathbb{B}_1$,

$$\|\mathbb{Z}x\|_2/\sqrt{n} \ge \|x_0\|_\Sigma(1-\epsilon) - 4\lambda_{\max,z}^{1/2}(1+\epsilon)\epsilon \ge \lambda_{\min,z}^{1/2}(1-\epsilon) - 4\lambda_{\max,z}^{1/2}(1+\epsilon)\epsilon > 0.$$

That is, with probability $\tilde{p}_{0n} := P(\mathcal{A}_1 \cap \mathcal{A}_2)$, the RE condition holds with $M(k_0, s) > 0$.

Step 2. Finally, we show the bound of $\tilde{p}_{0n}$. In fact, for any $x_0 \in S^{p_1-1}$, it holds that

$$\left|\frac{\|\mathbb{Z}x_0\|_2}{\sqrt{n}\|x_0\|_\Sigma} - 1\right| \le \left|\frac{\|\mathbb{Z}x_0\|_2^2}{n\|x_0\|_\Sigma^2} - 1\right|.$$

In addition, it is easy to see that

$$P(\mathcal{A}_1 \cap \mathcal{A}_2) \ge 1 - \sum_{x_0\in\Lambda_1} P\left(\left|\frac{\|\mathbb{Z}x_0\|_2^2}{n\|x_0\|_\Sigma^2} - 1\right| > \epsilon\right) - \sum_{x\in\Lambda_2} P\left(\left|\frac{\|\mathbb{Z}x\|_2^2}{n\|x\|_\Sigma^2} - 1\right| \ge \epsilon\right). \tag{A.13}$$

Recall the assumption $E(Z) = 0$ and $\Lambda_1 \subset S^{p_1-1}$. For any $x_0 \in \Lambda_1$, by Lemma 3.1 and consequently by Lemma B.4, for any $n > C_1\epsilon^{-1}$ for some constant $C_1 > 0$ defined in Lemma B.4, we have

$$P\left(\left|\frac{\|\mathbb{Z}x_0\|_2^2}{n\|x_0\|_\Sigma^2} - 1\right| > \epsilon\right) \le C_0 \exp\bigl(-C_K(n\epsilon)^{1/3}\bigr).$$

Furthermore, recall that $\Lambda_2 \subset B_2^{p_1}$. For any $x \in \Lambda_2$, by letting $\tilde{x} = x\|x\|_2^{-1} \in S^{p_1-1}$, we note that

$$\frac{\|\mathbb{Z}x\|_2^2}{n\|x\|_\Sigma^2} = \frac{\|\mathbb{Z}x\|x\|_2^{-1}\|_2^2}{n\|x\|x\|_2^{-1}\|_\Sigma^2} = \frac{\|\mathbb{Z}\tilde{x}\|_2^2}{n\|\tilde{x}\|_\Sigma^2}.$$

By the same argument, for any $x \in \Lambda_2$, as $n > C_1\epsilon^{-1}$ where constant C_1 is defined in Lemma B.4, we have

$$P\left(\left|\frac{\|\mathbb{Z}x\|_2^2}{n\|x\|_\Sigma^2} - 1\right| > \epsilon\right) \le C_0 \exp\bigl(-C_K(n\epsilon)^{1/3}\bigr). \tag{A.14}$$

Combining (A.13)–(A.14), (B.2) and (B.3), we have

$$\begin{aligned} P(\mathcal{A}_1 \cap \mathcal{A}_2) &\ge 1 - \bigl(|\Lambda_1| + |\Lambda_2|\bigr)C_0 \exp\bigl(-C_K(n\epsilon)^{1/3}\bigr) \\ &\ge 1 - 2C_0 \max\bigl\{|\Lambda_1|, |\Lambda_2|\bigr\} \exp\bigl(-C_K(n\epsilon)^{1/3}\bigr) \\ &= 1 - 2C_0 \max\left\{\left(\frac{c_1p_1}{m_1}\right)^{c_1m_1}, \left(\frac{4c_1p_1}{m_1}\right)^{c_1m_1/4}\right\} \exp\bigl(-C_K(n\epsilon)^{1/3}\bigr). \end{aligned}$$

Therefore, as $n > \epsilon^{-1}\max\{C_1, (c_1m_1\tilde{C}_K^{-1}\log(c_1m_1p_1))^3, (4^{-1}\tilde{C}_K^{-1}c_1m_1\log(4c_1m_1p_1))^3\}$ where $\tilde{C}_K = C_K - \delta_0$ for some $\delta_0 > 0$ being small, we have

$$P(\mathcal{A}_1 \cap \mathcal{A}_2) > 1 - 2C_0 \exp\bigl(-n^{1/3}\epsilon^{1/3}\delta_0\bigr) := \tilde{p}_{0n} \to 1. \tag{A.15}$$

Appendix B: Some auxiliary results

B.1. On the coverage of $\mathbb{B}_1$

In this section, we consider the coverage of the set $\mathbb{B}_1$ used in the proof of Proposition 3.1. For simplicity, for any finite set A, let $\operatorname{conv} A$ stand for the convex hull of A and for any constant C, $C\operatorname{conv}(A) = \{C \cdot x\colon\ x \in \operatorname{conv} A\}$. For $q = 1, 2$ and any positive integer m, define $B_q^m = \{x\colon\ \|x\|_q \le 1, x \in R^m\}$.

Lemma B.1 (Mendelson, Pajor and Tomczak-Jaegermann [16]). *Let $m \ge 1$ and $\varepsilon > 0$. There exists an ε cover $\Lambda \subset B_2^m$ of B_2^m with respect to the Euclidean metric such that $B_2^m \subset (1-\varepsilon)^{-1}\operatorname{conv}\Lambda$ and $|\Lambda| \le (1+2/\varepsilon)^m$. Similarly there exists $\Lambda' \subset S^{p-1}$ which is an ε cover of the sphere S^{m-1} and $|\Lambda'| \le (1+2/\varepsilon)^m$.*

Define $\tilde{U}_m = \{x \in B_2^p\colon\ |\operatorname{supp} x| \le m\}$ and $U_m = \{x \in S^{p-1}\colon\ |\operatorname{supp} x| \le m\}$. The following result holds.

Lemma B.2 (Mendelson, Pajor and Tomczak-Jaegermann [16]). *There exists an absolute constant c for which the following holds. For every $0 < \varepsilon \le 1/2$ and every $1 \le m \le p$, there is a set $\Lambda \subset B_2^p$ which is an ε cover of $\tilde{U}_m$, such that $\tilde{U}_m \subset 2\operatorname{conv}\Lambda$ and $|\Lambda|$ is at most*

$$\exp\left(cm\log\left(\frac{cp}{m\varepsilon}\right)\right). \tag{B.1}$$

Moreover, there exists an ε cover $\Lambda' \subset S^{p-1}$ of U_m with cardinality at most (B.1) Furthermore, for any $0 \le r \le 1$ there exists $\bar{\Lambda} \subset rB_2^p$ such that $(U_m - U_m) \cap rB_2^p \subset 2\operatorname{conv}\bar{\Lambda}$ and $|\bar{\Lambda}|$ is at most (B.1).

For any $x \in R^{p_1}$, let $T_{0,x}$ with $|T_{0,x}| = s$ be the index of the largest s elements of $|x| = (|x_1|, \dots, |x_{p_1}|)^T$. Based on the results of Mendelson, Pajor and Tomczak-Jaegermann [16], we consider the ϵ-cover of the set

$$\begin{aligned}
\mathbb{B}_1 &= \left\{x \in R^{p_1}\colon\ \exists T \text{ with } |T| \le s,\ \text{ such that } \|x_{T^c}\|_1 \le k_0\|x_T\|_1\right\} \cap S^{p_1-1} \\
&= \left\{x \in R^{p_1}\colon\ \exists T \text{ with } |T| \le s,\ \text{ such that } \|x_{T^c}\|_1 \le k_0\|x_T\|_1\right\} \cap \left\{x\colon\ \|x\|_2 \le 1\right\} \cap S^{p_1-1} \\
&:= A_0 \cap \left\{x\colon\ \|x\|_2 \le 1\right\} \cap S^{p_1-1}.
\end{aligned}$$

By the result of Rudelson and Zhou [21], it is easy to see

$$\begin{aligned}
A_0 &= \left\{x\colon\ x \in R^{p_1}, \exists T \text{ with } |T| \le s,\ \text{ such that } \|x_{T^c}\|_1 \le k_0\|x_T\|_1\right\} \\
&= \left\{x\colon\ x \in R^{p_1}, \|x_{T_{0,x}^c}\|_1 \le k_0\|x_{T_{0,x}}\|_1\right\}.
\end{aligned}$$

Therefore, we consider the coverage of

$$\begin{aligned}\mathbb{B}_1 &= \left\{x \in R^{p_1}\colon \|x_{T_{0,x}^c}\|_1 \le k_0\|x_{T_{x,0}}\|_1, \|x\|_2 \le 1\right\} \cap S^{p_1-1}\\ &:= \tilde{A}_0 \cap S^{p_1-1}.\end{aligned}$$

For $1 \le m \le p_1$, let

$$U_m = \left\{u \in S^{p_1-1}\colon \operatorname{supp}(u) \le m\right\}.$$

Lemma B.3. *The following two conclusions hold.* (1) *There exists* U_{m_1} *of* $\mathbb{B}_1$, *such that* $\mathbb{B}_1 \subseteq 2\operatorname{conv}(U_{m_1})$ *with* $m_1 = 16s[(1+k_0)]^2$; *in addition there exists* $\Lambda_1 \subset S^{p_1-1}$ *such that* $U_{m_1} \subset 2\operatorname{conv}\Lambda_1$ *and the cardinality of* Λ_1 *is given in* (B.2). (2) *There exists* $\Lambda_2 \subseteq \epsilon B_2^{p_1}$, *such that* $(\mathbb{B}_1 - \mathbb{B}_1) \cap \epsilon B_2^{p_1} \subseteq 4\operatorname{conv}\Lambda_2$, *for any* $0 \le \epsilon \le 1$. *And the cardinality of* Λ_2 *is given in* (B.3).

Proof. *Step* 1. The proof of conclusion (1). For $x \in \tilde{A}_0$, we have

$$\|x\|_1 \le (1+k_0)\|x_{T_0}\|_1 \le s^{1/2}(1+k_0)\|x_{T_0}\|_2 \le s^{1/2}(1+k_0) := q(s).$$

That is, $x \in q(s)B_1^{p_1}$. Therefore, we have

$$\mathbb{B}_1 \subset q(s)B_1^{p_1} \cap S^{p_1-1} := \tilde{\mathbb{B}}_1.$$

Therefore by Lemma 3.8 of Mendelson, Pajor and Tomczak-Jaegermann [16], it follows that U_{m_1} with $m_1 = 4s[(1+k_0)/\epsilon]^2$ is a ϵ-cover of $\tilde{\mathbb{B}}_1$, for any $0 < \epsilon \le 1/2$. Taking $\epsilon = 1/2$ leads to $m_1 = 16s[(1+k_0)]^2$. By Lemma B.2, there is a set $\Lambda_1 \subset S^{p_1-1}$, such that $U_{m_1} \subset 2\operatorname{conv}\Lambda_1$ and

$$|\Lambda_1| \le \exp\bigl(c_1 m_1 \log[c_1 p_1/m_1]\bigr) \tag{B.2}$$

for some constant $c_1 > 0$. Therefore, we have $\mathbb{B}_1 \subset 4\operatorname{conv}\Lambda_1$.

Step 2. The proof of the second conclusion.

Since $\mathbb{B}_1 \subset \tilde{\mathbb{B}}_1$ and $\tilde{\mathbb{B}}_1$ is convex and star shaped, for any $0 < \epsilon \le 1$, it holds that

$$(\mathbb{B}_1 - \mathbb{B}_1) \cap \epsilon B_2^{p_1} \subseteq (\tilde{\mathbb{B}}_1 - \tilde{\mathbb{B}}_1) \cap \epsilon B_2^{p_1} \subseteq \bigl[2q(s)B_1^{p_1} \cap 2B_2^{p_1}\bigr] \cap \epsilon B_2^{p_1} \subseteq 2q(s)B_1^{p_1} \cap B_2^{p_1}.$$

By Lemma 3.7 of Mendelson, Pajor and Tomczak-Jaegermann [16], for $m_2 = 4q^2(s) = 4s(1+k_0)^2$, we have

$$(\mathbb{B}_1 - \mathbb{B}_1) \cap B_2^{p_1} \subseteq 2\operatorname{conv}(\tilde{U}_{m_2}).$$

By Lemma B.2, there is a set $\Lambda_2' \subset B_2^{p_1}$ such that $\tilde{U}_{m_2} \subset 2\operatorname{conv}\Lambda_2'$ and

$$\bigl|\Lambda_2'\bigr| \le \exp\bigl(c_1 m_2 \log(c_1 p_1/m_2)\bigr).$$

Therefore, letting $\Lambda_2 = \epsilon\Lambda_2'$ yields $(\mathbb{B}_1 - \mathbb{B}_1) \cap \epsilon B_2^{p_1} \subset 4\operatorname{conv}\Lambda_2$ and $|\Lambda_2| = |\Lambda_2'| \le \exp(c_1 m_2 \log[c_1 p_1/m_2])$. Recalling the definition of m_1 and m_2, we have $m_2 = m_1/4$ and

$$|\Lambda_2| = \bigl|\Lambda_2'\bigr| \le \exp\bigl(4^{-1} c_1 m_1 \log[4c_1 p_1/m_1]\bigr). \tag{B.3}$$

$\square$

B.2. Concentration inequality of the squares of subexponential type variables

Lemma B.4. *Let $Z \in R$ be a subexponential type variable that satisfies $P(|Z| > t) < C\exp(-t/K)$ as $t > C_0$ for some positive constants C, K and C_0. Assume that $E(Z) = 0$ and that both $\operatorname{var}(Z)$ and $\operatorname{var}(Z^2)$ are finite. Let Z_i, $i = 1, \ldots, n$ be the i.i.d. realizations of Z. Then, for any $\delta > 0$, as $n\delta > C_1$ for some $C_1 > 0$, we have*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - \operatorname{var}(Z)\right| > \delta\right) \le C_2 \exp\bigl(-C_K (n\delta)^{1/3}\bigr),$$

where $C_2 > 0$ and C_K is a constant depending only on K.

Remark B.1. Similar to the subexponential distribution, we still denote $\|Z\|_{\psi_1} = K$. From Lemma B.4, we see that, for any fixed $\delta > 0$, as $n > C_1/\delta$, the inequality in Lemma B.4 holds. Equivalently, for fixed n, if $\delta > C_1/n$, the above inequality holds.

Proof of Lemma B.4. Without loss of generality, we assume $\operatorname{var}(Z) = 1$ in the following argument. Let $A_n = n^{-1}\sum_{i=1}^{n} Z_i^2 - 1$. For any fixed large constant $M > \max(C_0, 1)$, let $Y_i^M = (Z_i^2 - 1)I_{\{|Z_i|<M\}}$ and $X_i^M = (Z_i^2 - 1)I_{\{|Z_i|>M\}}$. Then we have

$$A_n = n^{-1}\sum_{i=1}^{n} Y_i^M + n^{-1}\sum_{i=1}^{n} X_i^M := A_{1n} + A_{2n}.$$

Note that $E(A_n) = E(A_{1n}) + E(A_{2n}) = 0$. It holds that

$$A_n = A_{1n} - E(A_{1n}) + A_{2n} - E(A_{2n}).$$

By the Bernstein inequality, due to the fact that $|Y_i^M| \le M^2 - 1$, we have, for any $\delta > 0$

$$\begin{aligned} P\bigl(\bigl|A_{1n} - E(A_{1n})\bigr| > \delta/2\bigr) &= P\left(\left|n^{-1}\sum_{i=1}^{n} Y_i^M - E\bigl(Y_i^M\bigr)\right| > \delta/2\right) \\ &\le 2\exp\left(-\frac{(n\delta/2)^2/2}{nE(Y_1^M)^2 + (M^2-1)n\delta/6}\right) \\ &\le 2\exp\left(-\frac{n\delta^2}{8E(Y_1^M)^2 + 4(M^2-1)\delta/3}\right) \\ &\le 2\exp\left(-\frac{n\delta^2}{8c_0 + 4(M^2-1)\delta/3}\right), \end{aligned} \tag{B.4}$$

where $c_0 = \operatorname{var}(Z^2) < \infty$. In addition,

$$P\bigl(\bigl|A_{2n} - E(A_{2n})\bigr| > \delta/2\bigr) \le \frac{4}{n\delta^2}\operatorname{var}\bigl(X_1^M\bigr) \le \frac{4}{n\delta^2}E\bigl(Z_1^4 I_{\{|Z_1|>M\}}\bigr).$$

Let $V = |Z_1|$ and note that V is subexponential. By integration by parts, it is easy to see

$$E\big(V^4 I_{\{V>M\}}\big) = P(V > M)M^4 + \int_{x>M} 4P(V > x)x^3 \,\mathrm{d}x.$$

Because V is subexponential, it follows that $P(V > s) \le C\exp(-s/K)$ for any $s > M$. Inserting this inequality into the above equation and applying integration by part to the second term repeatedly, for large M, we have

$$E\big(V^4 I_{\{V>M\}}\big) \le c_1 M^4 \exp(-C_K M),$$

for some $c_1 > 0$ where C_K is a constant depending only on K. Therefore, we have

$$P\big(\big|A_{2n} - E(A_{2n})\big| > \delta/2\big) \le \frac{c_1 M^4}{n\delta^2}\exp(-C_K M). \tag{B.5}$$

For $M = c_2(n\delta)^{1/3}$ for some constant c_2, the right-hand side of (B.4) has the same order as that of (B.5). Also, we remind that we need $M > \max(C_0, 1)$. Therefore, we have $n\delta > [c_2^{-1}\max(C_0, 1)]^3 := C_1$, that is δ cannot be too small. Consequently, for any $\delta > C_1/n$, we have

$$P\big(|A_n| > \delta\big) \le P\big(\big|A_{1n} - E(A_{1n})\big| > \delta/2\big) + P\big(\big|A_{2n} - E(A_{2n})\big| > \delta/2\big) \le C_2 \exp\big[-C_K(n\delta)^{1/3}\big]$$

for some constant $C_2 > 0$ and C_K, where C_K depends only on K. The proof is completed. □

Appendix C: Some properties of sub-Gaussian and subexponential distributions

The following two propositions on sub-Gaussian and subexponential distributions are from Vershynin [26]. They are listed here for completeness.

Proposition C.1 (Bernstein-type inequality). *Let $X_1, \dots, X_n$ be independent centered subexponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \dots, a_n) \in R^n$ and every $t \ge 0$, we have*

$$P\left(\left|\sum_{i=1}^n a_i X_i\right| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right),$$

where $c > 0$ is an absolute constant.

Proposition C.2. *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover, $\|X\|_{\psi_2}^2 \le \|X^2\|_{\psi_1} \le 2\|X\|_{\psi_2}^2$.*

Corollary 1. *If X, Y are centered sub-Gaussian with $\max\{\|X\|_{\psi_2}, \|Y\|_{\psi_2}\} < K$. Then XY is subexponential with $\|XY\|_{\psi_1} \le 2K^2$.*

Proof. For any $t > 0$, $P(|XY| > t) \leq P(X^2 + Y^2 > 2t) \leq P(X^2 > t) + P(Y^2 > t)$. By Proposition C.2, we have that X^2 and Y^2 are both subexponential with ψ_1 norm less than $2K^2$; that is, $P(X^2 > t) \leq c\exp(-t/2K^2)$ for some constant $c > 0$ and the same is true for Y^2. Consequently, we have $P(|XY| > t) \leq 2c\exp(-t/2K^2)$. The conclusion holds. $\square$

Acknowledgements

We are grateful to the Associate Editor and an anonymous referee for their helpful comments. The researches of Zhao is supported by National Science Foundation of China (Nos. 11101022, 11471030) and Fundamental Research Funds for the Central Universities.

References

[1] Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statist. Sci.* **27** 450–468. MR3025128
[2] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469
[3] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2010). Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat. Collect.* **6** 56–69. Beachwood, OH: IMS. MR2798511
[4] Bien, J., Taylor, J. and Tibshirani, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805
[5] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644
[6] Choi, N.H., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105** 354–364. MR2656056
[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581
[8] Fang, Y.G., Loparo, K.A. and Feng, X. (1994). Inequalities for the trace of matrix product. *IEEE Trans. Automat. Control* **39** 2489–2490. MR1337578
[9] Hall, P. and Xue, J.-H. (2014). On selecting interacting features from high-dimensional data. *Comput. Statist. Data Anal.* **71** 694–708. MR3132000
[10] Hao, N. and Zhang, H.H. (2012). Interaction selection under marginality principle in high dimensional regression. Manuscript.
[11] Hao, N. and Zhang, H.H. (2012). A note on regression models with interactions. Manuscript.
[12] Hao, N. and Zhang, H.H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697
[13] Lin, Y. and Zhang, H.H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500
[14] Magnus, J.R. and Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7** 381–394. MR0520247
[15] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351
[16] Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.* **28** 277–289. MR2453368

[17] Negahban, S.N., Ravikumar, P., Wainwright, M.J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist*. *Sci*. **27** 538–557. MR3025133

[18] Radchenko, P. and James, G.M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *J*. *Amer*. *Statist*. *Assoc*. **105** 1541–1553. MR2796570

[19] Raskutti, G., Wainwright, M.J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J*. *Mach*. *Learn*. *Res*. **11** 2241–2259. MR2719855

[20] Raskutti, G., Wainwright, M.J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans*. *Inform*. *Theory* **57** 6976–6994. MR2882274

[21] Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans*. *Inform*. *Theory* **59** 3434–3447. MR3061256

[22] Shah, R.D. (2012). Modelling interactions in high-dimensional data with backtracking. Manuscript.

[23] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **58** 267–288. MR1379242

[24] van de Geer, S. (2007). The deterministic lasso. Techinical Report 140, ETH, Zurich.

[25] van de Geer, S.A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron*. *J*. *Stat*. **3** 1360–1392. MR2576316

[26] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170

[27] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J*. *Amer*. *Statist*. *Assoc*. **104** 1512–1524. MR2750576

[28] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J*. *Mach*. *Learn*. *Res*. **11** 3519–3540. MR2756192

[29] Yuan, M., Joseph, V.R. and Zou, H. (2009). Structured variable selection and estimation. *Ann*. *Appl*. *Stat*. **3** 1738–1757. MR2752156

[30] 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. MR2212574

[31] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann*. *Statist*. **36** 1567–1594. MR2435448

[32] Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann*. *Statist*. **37** 3468–3497. MR2549566

[33] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J*. *Mach*. *Learn*. *Res*. **7** 2541–2563. MR2274449

Received April 2014 and revised November 2014