Electronic Journal of Statistics
Vol. 10 (2016) 1829–1873
ISSN: 1935-7524
DOI: 10.1214/16-EJS1120

The benefit of group sparsity in group inference with de-biased scaled group Lasso

Ritwik Mitra

ORFE, Princeton University, NJ 08540
e-mail: rmitra@princeton.edu

and

Cun-Hui Zhang[*]

Dept. of Statistics & Biostatistics, Rutgers University, NJ 08854
e-mail: czhang@stat.rutgers.edu

Abstract: We study confidence regions and approximate chi-squared tests for variable groups in high-dimensional linear regression. When the size of the group is small, low-dimensional projection estimators for individual coefficients can be directly used to construct efficient confidence regions and p-values for the group. However, the existing analyses of low-dimensional projection estimators do not directly carry through for chi-squared-based inference of a large group of variables without inflating the sample size by a factor of the group size. We propose to de-bias a scaled group Lasso for chi-squared-based statistical inference for potentially very large groups of variables. We prove that the proposed methods capture the benefit of group sparsity under proper conditions, for statistical inference of the noise level and variable groups, large and small. Such benefit is especially strong when the group size is large.

Keywords and phrases: Group inference, asymptotic normality, relaxed projection, chi-squared distribution, bias correction, relaxed projection.

Received December 2014.

1. Introduction

We consider the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ with an unknown noise level σ, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T \in \mathbb{R}^p$ is the vector of unknown true regression coefficients. We are interested in making statistical inference about a group of coefficients $\boldsymbol{\beta}_G^* = (\beta_j^*, j \in G)^T$. For small

[*]Research was supported in part by the NSF Grants DMS-11-06753 and DMS-12-09014.

p, the F-distribution, which is approximately chi-squared with proper normalization, provides classical confidence regions for $\boldsymbol{\beta}^*_G$ and p-values for testing $\boldsymbol{\beta}^*_G$. We want to construct approximate versions of such procedures for potentially very large groups in high-dimensional models where p is large, possibly much larger than n.

The study of asymptotic inference for parameter estimates in high dimensional regression has experienced a flurry of research activities in recent years. Many attempts have been made to assess the model selected by high dimensional regularizers; for example, some early work was done in Knight and Fu (2000), sample splitting was considered in Wasserman and Roeder (2009) and Meinshausen, Meier and Bühlmann (2009), and subsampling was considered in Meinshausen and Bühlmann (2010) and Shah and Samworth (2013). See Bühlmann and van de Geer (2011) for more detailed account of some of these methods. Leeb and Potscher (2006) proved that the sampling distribution of statistics based on selected models is not estimable. Berk, Brown and Zhao (2010) proposed conservative approaches. Alternative approaches were proposed in Lockhart et al. (2014) and Meinshausen (2014).

Recent works in Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014a) among others are more relevant to the line of research we have adopted in the current work, which we describe in some detail. For the effect of a preconceived variable, Zhang and Zhang (2014) pointed out the feasibility of regular statistical inference at the parametric $n^{-1/2}$ rate by correcting the bias of a regularized estimator of the entire coefficient vector, such as the Lasso, and proposed a low-dimensional projection estimator (LDPE) to carry out the task. The basic idea is to project the residual of the regularized estimator to the direction of a certain score vector which is approximately orthogonal to all variables other than the preconceived one. Such bias correction, which has been called de-biasing, is parallel to correcting the bias of nonparametric estimators in semiparametric inference (Bickel et al., 1993). In a general setting, Zhang (2011) developed an alternative formulation of the LDPE and provided formulas for the direction of the least favorable submodel and the Fisher information bound for the asymptotic variance. In linear regression, the least favorable submodel more explicitly connects the Lasso estimator of the score vector to column-by-column estimation of the precision matrix for random designs (Cai, Liu and Luo, 2011; Sun and Zhang, 2013). Bühlmann (2013) developed and studied methods to correct the bias of ridge regression. Belloni, Chernozhukov and Hansen (2014) considered estimation of treatment effects with a large number of controls. van de Geer et al. (2014) proved that the LDPE attains the Fisher information bound under a sparsity condition on the precision matrix and made a connection between the Lasso estimation of the score vector and the inversion of the Karush-Kuhn-Tucker (KKT) conditions through the precision matrix. Moreover, van de Geer et al. (2014) extended their results to generalized linear models (GLMs) with an innovative way of analyzing such models. Javanmard and Montanari (2014a) proved that when a quadratic programming method of Zhang and Zhang (2014) is used to estimate the score vector, the LDPE attains the Fisher information bound for Gaussian designs without requiring sparsity condition on the precision matrix; see Subsection 2.2 for further discussion.

In a separate work, Javanmard and Montanari (2014b) considered inference with lower sample size requirements when the design is known to be standard Gaussian. Sun and Zhang (2012a), Ren et al. (2013) and Jankova and van de Geer (2014) considered extensions to graphical models and precision matrix estimation.

It is possible to directly extend the above described de-biasing method to the case of grouped variables. In fact, the LDPE provides

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G\big) = \mathsf{N}_{|G|}\big(\mathbf{0}, \sigma^2 \mathbf{V}_{G,G}\big) + \mathrm{Rem}_G \tag{1.2}$$

along with a known covariance structure $\mathbf{V}_{G,G}$ and a remainder term that satisfies $\|\mathrm{Rem}_G\|_\infty \lesssim \|\boldsymbol{\beta}^*\|_0(\log p)/\sqrt{n}$ (Zhang and Zhang, 2014). However, this does not directly provide a sharp error bound for the ℓ_2- or equivalently chi-squared-based group inference for large groups. As $\mathrm{Var}(\chi_{|G|}) \approx 1/2$, the trivial bound $\|\mathrm{Rem}_G\|_2 \lesssim |G|^{1/2}\|\boldsymbol{\beta}^*\|_0(\log p)/\sqrt{n} = o(1)$ for group inference leads to an extra factor $|G|$ in the sample size requirement. Thus, the group inference problem is unsolved when one is unwilling to impose such a strong condition on n. Our goal is to construct $\widehat{\boldsymbol{\beta}}_G$ satisfying $\|\mathrm{Rem}_G\|_2 = o(1)$ in an expansion of the form (1.2) with potentially very large $|G|$. The impact of such a result is certainly beyond the specific problem under consideration.

Our approach is based on the natural idea that group sparsity can be exploited in statistical inference of variable groups. To this end, we propose to use a linear estimator to correct the bias of a scaled group Lasso estimator. This combines and extends the ideas of the group Lasso (Yuan and Lin, 2006) and bias correction (Zhang and Zhang, 2014), and will be shown to capture the benefit of group sparsity in both high-dimensional estimation as in Huang and Zhang (2010) and in bias correction. We note that the type of statistical inference under consideration here is regular in the sense that it does not require model selection consistency, and that it attains asymptotic efficiency in the sense of Fisher information without being super-efficient. A characterization of such inference is that it does not require a uniform signal strength condition on informative features, e.g. a lower bound on the non-zero $|\beta_j|$ above an inflated noise level due to model uncertainly, known as the "beta-min" condition.

Since our proposed method relies upon a group regularized initial estimator, in the following we provide a brief discussion of the literature on the topic. The group Lasso (Yuan and Lin, 2006) can be defined as

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega}) = \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}), \quad \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}) = \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \sum_{j=1}^{M} \omega_j \|\boldsymbol{\beta}_{G_j}\|_2, \tag{1.3}$$

where $\{G_j, , 1 \le j \le M\}$ forms a partition of the index set $\{1, \dots, p\}$ of variables. It is worthwhile to note that when the group effects are being regularized, the choice of the basis $\mathbf{X}_{G_j} = (\boldsymbol{x}_k, k \in G_j)$ within the group may not play a prominent role, so that the design is often "pre-normalized" to satisfy $\mathbf{X}_{G_j}^T \mathbf{X}_{G_j}/n = \mathbf{I}_{G_j \times G_j}$ as in Yuan and Lin (2006). The group Lasso and its variants have been studied in Bach (2008), Koltchinskii and Yuan (2008), Obozinski, Wainwright and Jordan (2008), Nardi and Rinaldo (2008), Liu and Zhang (2009), Huang and Zhang (2010), and Lounici et al. (2011) among many others. Huang and Zhang (2010) characterized the benefit of group Lasso in ℓ_2 estimation, versus the Lasso (Tibshirani, 1996), under the assumption of *strong group sparsity*; see (2.1) in Section 2. Huang et al. (2009) and Breheny and Huang (2011) developed methodologies for concave group and bi-level regularization. We refer to Huang, Breheny and Ma (2012) for further discussion and additional references

Estimation of the scale parameter, or the noise level σ, is also an important aspect of high dimensional regularized regression. Due to scale invariance, it is natural to let the groupwise weights in (1.3) be proportional to the scale parameter σ. Thus, a consistent estimate of σ also becomes necessary for truly adaptive estimation of the parameters. For the Lasso problem, Antoniadis (2010) and Sun and Zhang (2010, 2012b) proposed a scaled Lasso that estimates both the scale parameter σ and coefficient vector $\boldsymbol{\beta}^*$, which is closely related to the earlier proposals of Zhang (2010) and Städler, Bühlmann and Geer (2010). It turns out that this scaled Lasso and the square-root Lasso (Belloni, Chernozhukov and Wang, 2011) yield the same estimator of $\boldsymbol{\beta}$ although the estimation of σ is not considered in Belloni, Chernozhukov and Wang (2011). For group regularization, Bunea, Lederer and She (2014) proposed a square-root group Lasso for adaptive estimation of the coefficient vector $\boldsymbol{\beta}$. In this paper, we study a scaled group Lasso for simultaneous estimation of both $\boldsymbol{\beta}$ and σ with a different weighted $\ell_{2,1}$ penalty and prove the benefit of grouping in the estimation of the scale parameter in terms of convergence rates.

This paper is organized as follows. In Section 2, we describe a general procedure for statistical inference of groups of variables and provide theoretical guarantees for our results. In Section 3, we study the scaled group Lasso needed for the construction of estimators in Section 2. In Section 4, we present some simulation results to demonstrate the feasibility and performance of the proposed methods. In Section 5 we provide a brief summary of our results and discuss future directions of research. Proofs of some technical results are relegated to the Appendix.

We use the following notation throughout the paper. For vectors $\boldsymbol{u} \in \mathbb{R}^d$, the ℓ_p norm is denoted by $\|\boldsymbol{u}\|_p = (\sum_{k=1}^d |u_k|^p)^{1/p}$, with $\|\boldsymbol{u}\|_\infty = \max_{1\le k\le d} |u_k|$ and $\|\boldsymbol{u}\|_0 = \#\{j : u_j \neq 0\}$. For matrices $\mathbf{A}$, the Moore-Penrose pseudo inverse is denoted by $\mathbf{A}^\dagger$, the spectrum norm is denoted by $\|\mathbf{A}\|_S = \max_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \boldsymbol{u}^T\mathbf{A}\boldsymbol{v}$, the Frobenius norm by $\|\mathbf{A}\|_F = \{\mathrm{trace}(\mathbf{A}^T\mathbf{A})\}^{1/2}$, and the nuclear norm by $\|\mathbf{A}\|_N = \max_{\|\mathbf{B}\|_S=1} \mathrm{trace}(\mathbf{B}^T\mathbf{A})$. Given $A \subset \{1, \cdots, p\}$, for any vector $\boldsymbol{u} \in \mathbb{R}^p$, $\boldsymbol{u}_A \in \mathbb{R}^{|A|}$ denotes a vector with corresponding components from $\boldsymbol{u}$, $\mathbf{X}_A \in \mathbb{R}^{n\times|A|}$ denotes the sub-matrix of $\mathbf{X}$ with corresponding columns as indicated by the set A, $\mathbf{X}_{-A}$ denotes the sub-matrix of $\mathbf{X}$ with column indices belonging to the complement of A, $\mathcal{R}(\mathbf{X}_A)$ denotes the column space spanned by columns of $\mathbf{X}_A$, $\mathbf{Q}_A = \mathbf{X}_A(\mathbf{X}_A^T\mathbf{X}_A)^\dagger\mathbf{X}_A^T$ denotes the orthogonal projection to $\mathcal{R}(\mathbf{X}_A)$, and $\mathbf{Q}_A^\perp = \mathbf{I}_{p\times p} - \mathbf{Q}_A$. Additionally, $\mathbb{E}$ and $\mathbb{P}$ denote respectively the expectation and probability measure.

2. Group inference

We present our results in seven subsections. Subsection 2.1 describes the group structure of the regression problem in detail and the notion of strong group sparsity. Subsection 2.2 provides a brief account of the bias correction procedure for statistical inference of a single variable. Subsection 2.3 proposes an extension of the bias correction idea to group inference. Subsection 2.4 justifies the proposed group inference methodology in an ideal setting and states a working assumption for more general settings. Subsection 2.5 provides optimization methods for construction of group inference procedures under the working assumption. Subsection 2.6 provides sufficient conditions for the feasibility of the optimization scheme considered in Subsection 2.5. Subsection 2.7 discusses convexations of the optimization problem and summarizes the overall scheme.

2.1. Group structure and strong group sparsity

We assume an inherent and pre-specified non overlapping group structure of the feature set. Put precisely, assume that $\{1, \cdots, p\} = \cup_{j=1}^{M} G_j$ such that $G_j \cap G_k = \varnothing$. Define $d_j = |G_j|$ for all j so that $\sum_{j=1}^{M} d_j = p$. For any index set $T \subset \{1, \cdots, M\}$, we define $G_T = \cup_{j \in T} G_j$. In the following, we allow the quantities n, p, M, d_j's etc. to all grow to infinity.

In light of this group structure, further results on consistency of group regularized estimators of $\boldsymbol{\beta}^*$ will be based on a weighted mixed $\ell_{2,1}$ norm, defined as $\sum_{j=1}^{M} \omega_j \|\boldsymbol{u}_{G_j}\|_2$ for $\boldsymbol{u} = (\boldsymbol{u}_{G_j}; 1 \le j \le M) \in \mathbb{R}^p$ with $\boldsymbol{u}_{G_j} \in \mathbb{R}^{|G_j|}$, where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_M) \in \mathbb{R}^M$ with $\omega_j > 0$ for all j. This norm will be used both as penalty and as a key loss function. Weighted mixture norm of this type provides suitable description of the complexity of the unknown $\boldsymbol{\beta}$ when the following strong group sparsity condition of Huang and Zhang (2010) holds.

Strong group sparsity: *With the given group structure $\{G_j, j = 1, \ldots, M\}$ as a partition of $\{1, \ldots, p\}$, there exists a group-index set, $S^* \subset \{1, \cdots, M\}$, such that*

$$|S^*| \le g, \quad |G_{S^*}| \le s, \quad \mathit{supp}(\boldsymbol{\beta}^*) \subset G_{S^*} = \cup_{j \in S^*} G_j. \tag{2.1}$$

In this case, we say that the true coefficient vector $\boldsymbol{\beta}^*$ is (g, s) strongly group sparse with group support S^*.

Our aim is to make chi-squared-type statistical inference about the effect of a group G of variables, including confidence regions and p-values for $\mathbf{X}_G \boldsymbol{\beta}^*_G$ and $\boldsymbol{\beta}^*_G$. As will be clear from our analysis, the methodologies proposed in this paper will allow the size of the group G to grow unboundedly up to $|G| = o(n)$. Moreover, the group G of interest does not have to be congruent with the group structure $\{G_j, j = 1, \ldots, M\}$. In fact, each of the $|G|$ variables in G could belong to any of the M different pre-specified groups of variables so that

$$\mathbf{X}_G \boldsymbol{\beta}^*_G = \sum_{k: G_k \cap G \neq \emptyset} \mathbf{X}_{G_k \cap G} \boldsymbol{\beta}^*_{G_k \cap G}.$$

Thus we can rewrite the regression problem (1.1) as

$$\boldsymbol{y} = \mathbf{X}_G\boldsymbol{\beta}^*_G + \sum_{G_k \not\subseteq G} \mathbf{X}_{G_k\setminus G}\boldsymbol{\beta}^*_{G_k\setminus G} + \boldsymbol{\varepsilon} = \boldsymbol{\mu}^*_G + \sum_{G_k \not\subseteq G} \boldsymbol{\mu}^*_{G_k\setminus G} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where for any $A \subset \{1, \cdots, p\}$, $\boldsymbol{\mu}^*_A = \mathbf{X}_A\boldsymbol{\beta}^*_A$. In the simplest case, when the variable group of interest G matches the group structure in the sense that,

$$\mathbf{X}_G\boldsymbol{\beta}^*_G = \sum_{G_k \cap G \neq \emptyset} \mathbf{X}_{G_k}\boldsymbol{\beta}^*_{G_k}, \tag{2.3}$$

(e.g. $G = G_{j_0}$ for some $1 \leq j_0 \leq M$), (2.2) could be simplified as,

$$\boldsymbol{y} = \mathbf{X}_G\boldsymbol{\beta}^*_G + \sum_{G_k \cap G = \emptyset} \mathbf{X}_{G_k}\boldsymbol{\beta}^*_{G_k} + \boldsymbol{\varepsilon} = \boldsymbol{\mu}^*_G + \sum_{G_k \cap G = \emptyset} \boldsymbol{\mu}^*_{G_k} + \boldsymbol{\varepsilon}.$$

2.2. Bias correction for a single coefficient

In high-dimensional regression, regularized estimators have been extensively studied and proven to be consistent for the estimation of the entire mean vector $\mathbf{X}\boldsymbol{\beta}$ and coefficient vector $\boldsymbol{\beta}$ under various loss functions. However, since such estimators are typically nonlinear and biased, their sampling distribution is typically intractable. Zhang and Zhang (2014) proposed to correct the bias of a regularized estimator $\widehat{\boldsymbol{\beta}}^{(init)}$ with an LDPE of the following form:

$$\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + \boldsymbol{z}_j^T\big(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}\big)/\boldsymbol{z}_j^T\boldsymbol{x}_j, \tag{2.4}$$

where $\boldsymbol{z}_j$ is a certain score vector depending on $\mathbf{X}$ only. Here we provide a brief review of some ideas involved in this methodology to prepare their extension to group inference.

The basic idea of the LDPE can be briefly explained as follows. In the low-dimensional regime where $\text{rank}(\mathbf{X}) = p \leq n$, we may pick $\boldsymbol{z}_j = \boldsymbol{x}_j^\perp$ as the projection of $\boldsymbol{x}_j$ to the orthogonal complement of the column space of $\mathbf{X}_{-j} = (\boldsymbol{x}_k, k \neq j)$, i.e. $\boldsymbol{z}_j^T\mathbf{X}_{-j} = \mathbf{0}$ and $\boldsymbol{z}_j^T\boldsymbol{x}_j = \|\boldsymbol{z}_j^\perp\|_2^2 > 0$. For this choice $\boldsymbol{z}_j = \boldsymbol{x}_j^\perp$, the $\widehat{\beta}_j$ in (2.4) is identical to the least squares estimator $(\boldsymbol{x}_j^\perp)^T\boldsymbol{y}/(\boldsymbol{x}_j^\perp)^T\boldsymbol{x}_j$, and thus is unbiased regardless of the choice of the initial estimator. In the high dimensional case where $p > n$, $\boldsymbol{x}_j^\perp$ is no longer a valid choice of $\boldsymbol{z}_j$ as the condition $\boldsymbol{z}_j^T\mathbf{X}_{-j} = \mathbf{0}$ forces $\boldsymbol{z}_j = \mathbf{0}$ when $\mathbf{X}$ is in general position. When $\boldsymbol{z}_j^T\mathbf{X}_{-j} \neq \mathbf{0}$, the linear estimator $\widehat{\beta}_j^{(lin)} = \boldsymbol{z}_j^T\boldsymbol{y}/\boldsymbol{z}_j^T\boldsymbol{x}_j$ has unbounded bias for the estimation of β_j even if we assume the sparsity condition $\|\boldsymbol{\beta}\|_0 = 1$. However, the linear estimator is used in (2.4) to project the residual $\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}$ to the direction of $\boldsymbol{z}_j$ for the purpose of bias correction, and the full strength of the unbiasedness property $\boldsymbol{z}_j^T\mathbf{X}_{-j} = \mathbf{0}$ is not necessary to reduce the bias of $\widehat{\boldsymbol{\beta}}^{(init)}$ to an acceptable level.

The performance of a score vector $\boldsymbol{z}_j$ can be measured by a bias factor η_j and a noise factor τ_j defined as follows,

$$\eta_j = \|\boldsymbol{z}_j^T \mathbf{X}_{-j}\|_\infty / \|\boldsymbol{z}_j\|_2, \quad \tau_j = \|\boldsymbol{z}_j\|_2 / |\boldsymbol{z}_j^T \boldsymbol{x}_j|.$$

This can be seen from the following error decomposition for the LDPE in (2.4),

$$\widehat{\beta}_j - \beta_j = \boldsymbol{z}_j^T \boldsymbol{\varepsilon} / \boldsymbol{z}_j^T \boldsymbol{x}_j + \tau_j \mathrm{Rem}_j, \tag{2.5}$$

in which $\boldsymbol{z}_j^T \boldsymbol{\varepsilon} / \boldsymbol{z}_j^T \boldsymbol{x}_j \sim N(0, \tau_j^2 \sigma^2)$ and an ℓ_∞-ℓ_1 split leads to

$$\big|\mathrm{Rem}_j\big| = \big|\boldsymbol{z}_j^T \mathbf{X}_{-j} (\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}^*)_{-j}\big| / \|\boldsymbol{z}_j\|_2 \le \eta_j \big\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}^*\big\|_1. \tag{2.6}$$

Thus, when $|\mathrm{Rem}_j| = o_{\mathbb{P}}(1)$, statistical inference for β_j can be carried out with a consistent estimate of σ. For example, when $\eta_j \lesssim \sqrt{\log p}$ and $\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}^*\big\|_1 \lesssim \|\boldsymbol{\beta}^*\|_0 \sqrt{(\log p)/n}$,

$$n \gg (\|\boldsymbol{\beta}\|_0 \log p)^2 \;\Rightarrow\; (\widehat{\beta}_j - \beta_j^*)/(\widehat{\sigma}\tau_j) \approx (\widehat{\beta}_j - \beta_j^*)/(\sigma\tau_j) \approx \mathsf{N}(0,1)$$

It is worthwhile to mention here that τ_j and η_j are both explicitly available given $\boldsymbol{z}_j$, so that the validity of the above scheme requires no stronger assumptions than an ℓ_1 error bound for the estimation of $\boldsymbol{\beta}$ and a consistent estimate of σ. A scaled Lasso estimator can be used as $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$, which satisfies

$$\left|\frac{\widehat{\sigma}}{\sigma^*} - 1\right| + \left(\frac{\log p}{n}\right)^{1/2} \|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}\left(\frac{\|\boldsymbol{\beta}^*\|_0 \log p}{n}\right), \tag{2.7}$$

with $\sigma^* = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2/\sqrt{n}$ and $s = \|\boldsymbol{\beta}^*\|_0$ (Sun and Zhang, 2012b), provided an ℓ_1 restricted eigenvalue or compatibility condition on the design (Bickel, Ritov and Tsybakov, 2009; van de Geer and Bühlmann, 2009). Thus, the remaining issue is to find a score vector $\boldsymbol{z}_j$ with sufficiently small a bias factor η_j and a noise factor τ_j.

For random designs with a Gram matrix $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}^T\mathbf{X}/n)$ that is invertible, Zhang (2011) provided the direction of the least favorable submodel $\boldsymbol{\beta} = \beta_j \boldsymbol{u}$ as

$$\boldsymbol{u}_j^o = \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j / (\boldsymbol{\Sigma}^{-1})_{j,j} = \arg\min_{\boldsymbol{u}} \Big\{\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} : \boldsymbol{e}_j^T\boldsymbol{u} = 1\Big\},$$

with $\boldsymbol{e}_j$ being the j-th canonical unit vector, and defined an ideal, efficient $\boldsymbol{z}_j$ as

$$\boldsymbol{z}_j^o = \mathbf{X}\boldsymbol{u}_j^o.$$

As the j-th element of $\boldsymbol{u}_j^o$ equals 1, this can be written as a linear regression model

$$\boldsymbol{x}_j = \mathbf{X}_{-j}\boldsymbol{\gamma}_{-j} + \boldsymbol{z}_j^o \tag{2.8}$$

with $\boldsymbol{\gamma}_{-j} = (\gamma_{1,j}, \cdots, \gamma_{j-1,-j}, \gamma_{j+1,j}, \cdots, \gamma_{p,j})^T = (-\boldsymbol{u}_j^o)_{-j} \in \mathbb{R}^{p-1}$.

Given a design matrix $\mathbf{X}$, Zhang and Zhang (2014) proposed two choices of $\boldsymbol{z}_j$ for the LDPE in (2.4). The first proposal of $\boldsymbol{z}_j$ takes a point in the Lasso path in the linear regression of $\boldsymbol{x}_j$ against $\mathbf{X}_{-j}$:

$$\boldsymbol{z}_j = \boldsymbol{x}_j - \mathbf{X}_{-j}\widehat{\gamma}_{-j}, \quad \widehat{\gamma}_{-j} = \arg\min_{\boldsymbol{b}} \Big\{ \|\boldsymbol{x}_j - \mathbf{X}_{-j}\boldsymbol{b}\|_2^2/2n + \lambda_j\|\boldsymbol{b}\|_1 \Big\}. \tag{2.9}$$

For $p \le n$, we may take $\lambda_j = 0$, so that $\boldsymbol{z}_j = \boldsymbol{x}_j^{\perp}$ and the $\widehat{\beta}_j$ in (2.4) is the least squares estimator of β_j. For $p > n$, (2.9) provides a relaxed projection of $\boldsymbol{x}_j$ via the Lasso, and the KKT conditions for $\boldsymbol{z}_j$ automatically provides

$$\tau_j \le 1/\|\boldsymbol{z}_j\|_2, \quad \eta_j = \|\boldsymbol{z}_j^T\mathbf{X}_{-j}\|_\infty/\|\boldsymbol{z}_j\|_2 = n\lambda_j/\|\boldsymbol{z}_j\|_2,$$

which implies $\eta_j = \sqrt{2\log p}$ with a scaled λ_j satisfying $\lambda_j = \sqrt{\|\boldsymbol{z}_j\|_2^2(2\log p)/n^2}$.

The second proposal of $\boldsymbol{z}_j$, closely related to the first one in (2.9) and given in the discussion section of Zhang and Zhang (2014), was a constrained variance minimization scheme

$$\boldsymbol{z}_j = \arg\min_{\boldsymbol{z}} \Big\{ \|\boldsymbol{z}\|_2^2 : |\boldsymbol{z}^T\boldsymbol{x}_j/n| = 1, \|\boldsymbol{z}^T\mathbf{X}_{-j}/n\|_\infty \le \lambda'_j \Big\}. \tag{2.10}$$

This quadratic program, which provides $\tau_j = \|\boldsymbol{z}_j\|_2/n$, can be understood as

$$\text{minimize} \quad \tau_j^2 \quad \text{subject to} \quad \eta_j \le \lambda'_j/\tau_j \approx \sqrt{2\log p}.$$

A variant of the optimization in (2.10), studied in Javanmard and Montanari (2014a) is

$$\widetilde{\boldsymbol{z}}_j = \mathbf{X}\widehat{\boldsymbol{m}}, \qquad \widehat{\boldsymbol{m}} = \arg\min_{\boldsymbol{m}} \Big\{ \boldsymbol{m}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{m} : \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{m} - \boldsymbol{e}_j\|_\infty \le \lambda''_j \Big\}. \tag{2.11}$$

Since $\widetilde{\boldsymbol{z}}_j^T\boldsymbol{x}_j/n = 1 - \lambda''_j$ and (2.10) is neutral in the sign of $\boldsymbol{z}$, (2.11) and (2.10) are equivalent with $\widetilde{\boldsymbol{z}}_j/(1-\lambda''_j) = \boldsymbol{z}_j$ when $\lambda_j = \lambda''_j/(1-\lambda''_j)$ and $\boldsymbol{z}_j$ is the solution with $\boldsymbol{z}_j^T\boldsymbol{x}_j = n$.

2.3. *Bias correction for a group of variables*

In this subsection we propose a multivariate extension of the methodologies described in Subsection 2.2.

The algebraic extension of (2.4) to the grouped variable scenario is straightforward. For the estimation of $\boldsymbol{\beta}_G^*$, a formal vectorization of the estimator is

$$\widehat{\boldsymbol{\beta}}_G = \widehat{\boldsymbol{\beta}}_G^{(init)} + (\mathbf{Z}_G^T\mathbf{X}_G)^\dagger\mathbf{Z}_G^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}), \tag{2.12}$$

where $\mathbf{Z}_G \in \mathbb{R}^{n\times|G|}$, depending on $\mathbf{X}$ only, can be viewed as a "score matrix". Recall that for any matrix $\mathbf{A}$, $\mathbf{A}^\dagger$ is its Moore-Penrose pseudo inverse. For the estimation of $\boldsymbol{\mu}_G^* = \mathbf{X}_G\boldsymbol{\beta}_G^*$, a variation of (2.12) is

$$\widehat{\boldsymbol{\mu}}_G = \widehat{\boldsymbol{\mu}}_G^{(init)} + (\mathbf{Z}_G\mathbf{Q}_G)^\dagger\mathbf{Z}_G^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}), \tag{2.13}$$

where $\widehat{\boldsymbol{\mu}}_G^{(init)} = \mathbf{X}_G\widehat{\boldsymbol{\beta}}_G^{(init)}$ and $\mathbf{Q}_G$ is the orthogonal projection to the column space $\mathbf{X}_G$.

The extension of the error decomposition (2.5) to (2.12) and (2.13) is also algebraic but requires a mild condition due to the need to factorize out a multivariate version of the noise factor. We carry out this task in the following proposition.

Proposition 1. *Let $\boldsymbol{Z}_G \in \mathbb{R}^{n\times|G|}$, $\boldsymbol{Q}_A$ and $\boldsymbol{P}_{G,0}$ be the orthogonal projections to $\mathcal{R}(\boldsymbol{X}_A)$ and $\mathcal{R}(\boldsymbol{Z}_G)$ respectively, $\boldsymbol{P}_G$ be the orthogonal projection to $\mathcal{R}(\boldsymbol{P}_{G,0}\boldsymbol{Q}_G)$, $\widehat{\boldsymbol{\beta}}_G$ be as in (2.12), $\widehat{\boldsymbol{\mu}}_G = \boldsymbol{X}_G\widehat{\boldsymbol{\beta}}_G$, $\boldsymbol{\mu}_A^* = \boldsymbol{X}_A\boldsymbol{\beta}_A^*$, $\widehat{\boldsymbol{\mu}}_A^{(init)} = \boldsymbol{X}_A\boldsymbol{\beta}_A^{(init)}$, and*

$$Rem_G = \sum_{G_k \not\subseteq G} \boldsymbol{P}_G\Big(\widehat{\boldsymbol{\mu}}_{G_k\setminus G}^{(init)} - \boldsymbol{\mu}_{G_k\setminus G}^*\Big) = \sum_{G_k \not\subseteq G}\Big(\boldsymbol{P}_G\boldsymbol{Q}_{G_k\setminus G}\Big)\Big(\widehat{\boldsymbol{\mu}}_{G_k\setminus G}^{(init)} - \boldsymbol{\mu}_{G_k\setminus G}^*\Big). \tag{2.14}$$

(i) Suppose rank$(\boldsymbol{Z}_G^T\boldsymbol{X}_G) = |G|$. Then, rank$(\boldsymbol{P}_G\boldsymbol{X}_G) = |G|$, $\boldsymbol{P}_G = \boldsymbol{P}_{G,0}$, and

$$\widehat{\boldsymbol{\beta}}_G = \widehat{\boldsymbol{\beta}}_G^{(init)} + (\boldsymbol{P}_G\boldsymbol{X}_G)^\dagger\boldsymbol{P}_G\Big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}\Big) = \boldsymbol{\beta}_G^* + (\boldsymbol{P}_G\boldsymbol{X}_G)^\dagger\left(\boldsymbol{P}_G\boldsymbol{\varepsilon} - Rem_G\right). \tag{2.15}$$

(ii) Suppose rank$(\boldsymbol{P}_G) =$ rank$(\boldsymbol{X}_G)$. Then, (2.13) holds and

$$\widehat{\boldsymbol{\mu}}_G = \widehat{\boldsymbol{\mu}}_G^{(init)} + (\boldsymbol{P}_G\boldsymbol{Q}_G)^\dagger\boldsymbol{P}_G\Big(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}\Big) = \boldsymbol{\mu}_G^* + (\boldsymbol{P}_G\boldsymbol{Q}_G)^\dagger\left(\boldsymbol{P}_G\boldsymbol{\varepsilon} - Rem_G\right). \tag{2.16}$$

Consequently,

$$(\boldsymbol{P}_G\boldsymbol{Q}_G)(\widehat{\boldsymbol{\mu}}_G - \boldsymbol{\mu}_G^*) = (\boldsymbol{P}_G\boldsymbol{X}_G)(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}_G^*) = \boldsymbol{P}_G\boldsymbol{\varepsilon} - Rem_G. \tag{2.17}$$

In particular, when $\boldsymbol{\mu}_G^* = \mathbf{0}$,

$$\boldsymbol{P}_G\boldsymbol{\varepsilon} - Rem_G = \boldsymbol{P}_G\widehat{\boldsymbol{\mu}}_G = \boldsymbol{P}_G\Big(\boldsymbol{y} - \sum_{G_k\not\subseteq G}\widehat{\boldsymbol{\mu}}_{G_k\setminus G}^{(init)}\Big). \tag{2.18}$$

The first equations of (2.15) and (2.16) assert the scale invariance of the proposed estimator in the choice of $\boldsymbol{Z}_G$ in the sense that it depends in $\mathbf{Z}_G$ only through the projection $\mathbf{P}_G$.

The condition rank$(\mathbf{P}_G)$ = rank$(\mathbf{X}_G)$, slightly weaker than the condition rank$(\mathbf{Z}_G^T\mathbf{X}_G) = |G|$, requires $\mathbf{Z}_G^T\mathbf{X}_G$ to have the same kernel as $\mathbf{X}_G$. If this condition fails to hold, there will be no bias correction in a certain direction $\boldsymbol{a} = \mathbf{X}_G\boldsymbol{b}_G \neq \mathbf{0}$ in the sense that $\boldsymbol{a}^T\widehat{\boldsymbol{\mu}}_G = \boldsymbol{a}^T\widehat{\boldsymbol{\mu}}_G^{(init)}$.

In Proposition 1, the matrices $(\mathbf{P}_G\mathbf{X}_G)^\dagger$ and $(\mathbf{P}_G\mathbf{Q}_G)^\dagger$ and can be viewed as multivariate noise factors respectively for statistical inference of $\boldsymbol{\beta}_G^*$ and $\boldsymbol{\mu}_G^*$, and the remainder term Rem$_G$ can be viewed as standardized bias.

For any estimator $\widehat{\sigma}$ for the noise level and measurable function $h : \mathcal{R}(\mathbf{P}_G) \to \mathbb{R}$,

$$h\big((\mathbf{P}_G\mathbf{Q}_G)(\widehat{\boldsymbol{\mu}}_G - \boldsymbol{\mu}^*_G)/\widehat{\sigma}\big) = h\big((\mathbf{P}_G\mathbf{X}_G)(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G)/\widehat{\sigma}\big) \tag{2.19}$$

is an approximate pivotal quantity with approximate distribution $h(\mathbf{P}_G\boldsymbol{\varepsilon}/\sigma)$ whenever

$$\sup_{-\infty<t<\infty}\Big|\mathbb{P}\Big\{h\big((\mathbf{P}_G\boldsymbol{\varepsilon} - \mathrm{Rem}_G)/\widehat{\sigma}\big) \le t\Big\} - \mathbb{P}\Big\{h\big(\mathbf{P}_G\boldsymbol{\varepsilon}/\sigma\big) \le t\Big\}\Big| = o(1). \tag{2.20}$$

From this point of view, the proposed method is generic. If a pivotal quantity (2.19) with a specific $h(\cdot)$ suits the aim of a statistical experiment, statistical inference can be carried out if certain estimator $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ and score matrix $\mathbf{Z}_G$ can be found to satisfy (2.20).

As we are interested in chi-squared type inference, the right choice of $h(\cdot)$ is $h(\boldsymbol{v}) = \|\boldsymbol{v}\|_2$. This choice yields elliptical confidence regions for $\boldsymbol{\beta}^*_G$ and $\boldsymbol{\mu}^*_G$ via (2.19). For testing the hypothesis $H_0 : \boldsymbol{\beta}_G = \mathbf{0}$, (2.18) provides the test statistic

$$T_G = \frac{1}{\widehat{\sigma}}\left\|\mathbf{P}_G\left(\boldsymbol{y} - \sum_{G_k \not\subseteq G}\widehat{\boldsymbol{\mu}}^{(init)}_{G_k\setminus G}\right)\right\|_2 \tag{2.21}$$

as an approximation of $\|\mathbf{P}_G\boldsymbol{\varepsilon}/\sigma\|_2$. Let $k_G = \mathrm{rank}(\mathbf{P}_G)$. It is worthwhile to note that

$$\|\mathbf{P}_G\boldsymbol{\varepsilon}\|_2/\sigma - \sqrt{k_G} \to \mathsf{N}(0, 1/2) \tag{2.22}$$

when $k_G \to \infty$. Thus, without further investigation of possible stochastical cancellation between $\mathbf{P}_G\boldsymbol{\varepsilon}$ and Rem_G, (2.20) for $h(\boldsymbol{v}) = \|\boldsymbol{v}\|_2$ and $k_G \ge 1$ amounts to

$$\sqrt{k_G}\big|\widehat{\sigma}/\sigma - 1\big| + \big\|\mathrm{Rem}_G/\sigma\big\|_2 = o_{\mathbb{P}}(1). \tag{2.23}$$

As $\|\mathbf{P}_G\boldsymbol{\varepsilon}\|_2^2/\sigma^2$ has the $\chi^2_{k_G}$ distribution, (2.23) implies

$$\begin{cases} \sup_t\Big|\mathbb{P}\Big\{\|(\mathbf{P}_G\mathbf{X}_G)(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G)\|_2^2 \le \widehat{\sigma}t\Big\} - \mathbb{P}\Big\{\chi^2_{k_G} \le t\Big\}\Big| \to 0, \\ \sup_t\Big|\mathbb{P}\Big\{\|(\mathbf{P}_G\mathbf{Q}_G)(\widehat{\boldsymbol{\mu}}_G - \boldsymbol{\mu}^*_G)\|_2^2 \le \widehat{\sigma}t\Big\} - \mathbb{P}\Big\{\chi^2_{k_G} \le t\Big\}\Big| \to 0, \\ \boldsymbol{\mu}^*_G = \mathbf{0} \;\Rightarrow\; \sup_t\Big|\mathbb{P}\Big\{T_G^2 \le t\Big\} - \mathbb{P}\Big\{\chi^2_{k_G} \le t\Big\}\Big| \to 0. \end{cases} \tag{2.24}$$

When $k_G = \mathrm{rank}(\mathbf{P}_G) \to \infty$, we can apply central limit theorem (2.22) to approximate the $\chi^2_{k_G}$ distribution.

The problem, as before, is to choose $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ and $\mathbf{Z}_G$ to guarantee (2.23) for the given $h(\cdot)$. For definiteness, we will pick in the sequel the following scaled version of the group Lasso estimator (1.3):

$$\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\} = \mathop{\arg\min}_{\boldsymbol{\beta},\sigma}\left\{\frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \sum_{j=1}^{M}\omega_j\|\boldsymbol{\beta}_{G_j}\|_2\right\}. \tag{2.25}$$

This estimator, which aims to take advantage of the group sparsity (2.1), will be considered carefully in Section 3, so that we can move on to the more pressing issue of finding a proper $\mathbf{Z}_G$. Still, we would like to mention that this choice of $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ and $h(\cdot)$ will in no way confine the scope of the proposed method, as Proposition 1 and (2.20) are completely general.

2.4. An ideal solution and a working assumption

To study the feasibility of the approach outlined above in Subsection 2.3, we first consider, parallel to (2.8), an ideal $\mathbf{Z}_G$ as the noise matrix in the following multivariate regression model,

$$\mathbf{X}_G = \mathbf{X}_{-G}\boldsymbol{\Gamma}_{-G,G} + \mathbf{Z}^o_G. \tag{2.26}$$

This regression model is best explained in the context of random design where

$$\boldsymbol{\Gamma}_{-G,G} = \left\{\mathbb{E}(\mathbf{X}^T_{-G}\mathbf{X}_{-G})\right\}^{-1}\mathbb{E}(\mathbf{X}^T_{-G}\mathbf{X}_G). \tag{2.27}$$

To this end, we consider in the following theorem random design matrices $\mathbf{X}$ having iid sub-Gaussian rows satisfying $\mathbb{E}\mathbf{X} = \mathbf{0}$, $\mathbb{E}(\mathbf{X}^T\mathbf{X}/n) = \boldsymbol{\Sigma}$ with a positive-definite $\boldsymbol{\Sigma}$, and

$$(\textbf{Sub-Gaussianity}) \qquad \sup_{\boldsymbol{b}\neq\mathbf{0}} \mathbb{E}\exp\left(\frac{(\boldsymbol{e}_i^T\mathbf{X}\boldsymbol{b})^2}{v_0\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b}} + \frac{1}{v_0}\right) \leq 2 \tag{2.28}$$

with a certain constant $v_0 > 1$, where $\boldsymbol{e}_i \in \mathbb{R}^n$ is the i^{th} canonical unit vector in $\mathbb{R}^n$.

Theorem 1. *Let $0 < c_* \leq c^*$ and $1 < A_* < A^*$ be fixed constants and $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ be a solution of (2.25) with $\omega_j/A^* \leq \|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S\omega_{*,j} \leq \omega_j/A_*$, where $\omega_{*,j} = n^{-1/2}(\sqrt{|G_j|} + \sqrt{2\log M})$. Suppose $\boldsymbol{X}$ satisfies condition (2.28) with $c_* \leq$eigenvalues$(\boldsymbol{\Sigma}) \leq c^*$. Let $\boldsymbol{Z}^o_G$ be as in (2.26) with the $\boldsymbol{\Gamma}_{-G,G}$ in (2.27) and $\widehat{\boldsymbol{\beta}}_G$ be as in (2.12) with $\boldsymbol{Z}_G = \boldsymbol{Z}^o_G$. Suppose $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$ and $\boldsymbol{\beta}^*$ satisfies the (g,s) strong group sparsity condition (2.1) with*

$$\frac{\max_{j\leq M}|G_j|}{n} + \frac{|G|}{n} \to 0, \qquad \frac{s + g\log M}{n^{1/2}}\left(\frac{|G|^{1/2}}{n^{1/2}} + \max_{G_k\nsubseteq G}\frac{\omega'_k}{\omega_{*,k}}\right) \to 0, \tag{2.29}$$

where $\omega'_k = n^{-1/2}\big(\sqrt{|G| + |G_k\setminus G|} + \sqrt{\log M}\big)$. Then, $\mathbb{P}\{rank(\boldsymbol{P}_G) = |G|\} \to 1$, (2.24) holds, and

$$(\boldsymbol{P}_G\boldsymbol{Q}_G)(\widehat{\boldsymbol{\mu}}_G - \boldsymbol{\mu}^*_G)/\widehat{\sigma} = (\boldsymbol{P}_G\boldsymbol{X}_G)(\widehat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}^*_G)/\widehat{\sigma} = \mathsf{N}_n(\mathbf{0}, \boldsymbol{P}_G) + o_{\mathbb{P}}(1). \tag{2.30}$$

Theorem 1, whose proof is merged with that of Theorem 4 and provided in Subsection 2.6, asserts that with a combination of the $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ in (2.25) and the ideal $\mathbf{Z}_G = \mathbf{Z}^o_G$ in (2.26), bias correction provides valid asymptotic chi-squared-type statistical inference for the group effect $\boldsymbol{\mu}^*_G \in \mathbb{R}^n$ and the coefficient group $\boldsymbol{\beta}^*_G \in \mathbb{R}^{|G|}$. However, this theorem requires a sub-Gaussian design and the knowledge of $\mathbf{Z}^o_G$.

To extend this approach to more general settings with unknown $\mathbf{Z}^o_G$ or even deterministic $\mathbf{X}$, we follow a strategy parallel to the one described in Subsection 2.2: We may directly approximate $\mathbf{Z}^o_G$ via a regularized multivariate regression in (2.26) or mimic properties of $\mathbf{Z}^o_G$ with a regularized optimization scheme. The question is to make a right choice of the regularization on $\mathbf{Z}_G$ to match properties one can reasonably expect from $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$. To this end, we extract, as the following working assumption, some properties of $\{\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma}\}$ which are proven and used in our analysis under the conditions of Theorem 1.

Working assumption: *Suppose that we have estimators* $\widehat{\boldsymbol{\beta}}^{(init)}$ *and* $\widehat{\sigma}$ *of a* (g, s) *strong group sparse signal* $\boldsymbol{\beta}^*$ *and scale parameter* σ *respectively satisfying*

$$\left|\frac{\widehat{\sigma}}{\sigma^*} - 1\right| + \frac{1}{n^{1/2}} \sum_{j=1}^{M} \frac{\omega_{*,j}}{\sigma} \left\| \boldsymbol{X}_{G_j} \widehat{\boldsymbol{\beta}}^{(init)}_{G_j} - \boldsymbol{X}_{G_j} \boldsymbol{\beta}^*_{G_j} \right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{s + g\log M}{n}\right), \tag{2.31}$$

where $\omega_{*,j} = \sqrt{|G_j|/n} + \sqrt{(2/n)\log M}$, $\sigma^* = \|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{y}\|_2/\sqrt{n}$ *is an oracle estimate of the noise level* σ*, and* G_j*,* s *and* g *are as in (2.1).*

The above working assumption still aims to take advantage of the group sparsity (2.1) as the mixed prediction error and the complexity measure $s + g\log M$ dictate. However, compared with the more specific (2.25), it provides a direction for regularizing a proper $\mathbf{Z}_G$ for any estimator satisfying (2.31), possibly with deterministic designs.

Under the strong group sparsity (2.1), error bounds in the ℓ_2 and mixed $\ell_{2,1}$ norms for group regularized methods have been established in the literature as we reviewed in the introduction. In Section 3, we contribute to this literature by obtaining ℓ_2 as well as weighted mixed ℓ_2 norm error bounds of the group Lasso and its scaled version (2.25). We will also provide a faster rate of convergence of the scale parameter σ under strong group sparsity, which is crucial to our analysis. In particular, we will prove in Section 3 that the error bound for $\widehat{\boldsymbol{\beta}}^{(init)}$ in (2.31) is attainable under proper conditions on the design matrix if the group Lasso is used with a proper estimate of σ, and the error bounds for both $\widehat{\boldsymbol{\beta}}^{(init)}$ and $\widehat{\sigma}$ in (2.31) are attainable if the scaled group Lasso is used; see Corollaries 1 and 2 and Theorem 7.

It is worthwhile to point out that the working assumption exhibits the benefit of strong group sparsity, compared with a reasonable working assumption based on the ℓ_0 sparsity condition $\|\boldsymbol{\beta}^*\|_0 \le s$ as given in (2.7). In general, the error bounds in (2.31) and those in (2.7) do not strictly dominate each other. However, if in both the scenarios, s is of similar order and $g \ll s$, then (2.31) dominates the rates necessary for univariate inference as given in (2.7).

An alternative possibility is to use an ℓ_1 regularized estimate of $\mathbf{\Gamma}_{-G,j}$ in the univariate regression of $\boldsymbol{x}_j$ against $\mathbf{X}_{-G}$ for all individual $j \in G$. This has been considered in van de Geer (2014). However, the advantage of such a scheme is unclear compared with directly using $(\widehat{\beta}_j, j \in G)^T$ with the $\widehat{\beta}_j$ in (2.4). It is worthwhile to mention that the central limit theorem for (2.4) came with large deviation bounds to justify Bonferroni adjustments (Zhang and Zhang, 2014), so that (2.4) and its variations can be used to test $H_0 : \boldsymbol{\beta}^*_G = \mathbf{0}$ versus an alternative hypothesis on $\|\boldsymbol{\beta}^*_G\|_\infty$, especially when an ℓ_1 regularized $\widehat{\boldsymbol{\beta}}^{(init)}$ is used as in van de Geer et al. (2014). However, we are interested in extensions of traditional F- or chi-squared tests for ℓ_2 alternatives and taking advantage of the group sparsity of $\boldsymbol{\beta}^*$. Such methods require control of ℓ_2 and groupwise weighted ℓ_2 error and accordingly, a proper choice $\mathbf{Z}_G$ to match the working assumption.

2.5. An optimization strategy

In this subsection we propose a multivariate extension of the optimization strategy (2.10) to match an initial estimator satisfying the working assumption (2.31) in the bias correction scheme (2.12).

It follows from Proposition 1 that the estimator (2.12) depends on the resulting $\mathbf{Z}_G$ only through the orthogonal projection $\mathbf{P}_G$ to the range of $\mathbf{Z}_G$ under a necessary assumption for the bias correction scheme to work, as we commented below Proposition 1. Moreover, it follows from (2.14) and (2.17) that the desired $\mathbf{P}_G$, which depends on $\mathbf{X}$ only, must be close to $\mathbf{Q}_G$ and approximately orthogonal to $\mathbf{Q}_{G_k\setminus G}$ for all k with $G_k \not\subseteq G$.

Let $\mathbf{Q}$ be the projection to $\mathcal{R}(\mathbf{X})$. In the low-dimensional case of $\mathrm{rank}(\mathbf{X}) = p < n$, we may set $\mathbf{P}_G = \mathbf{Q}\prod_{G_k\not\subseteq G}\mathbf{Q}^\perp_{G_k\setminus G}$, so that (2.12) is the least squares estimator of $\boldsymbol{\beta}_G$ with $\mathrm{Rem}_G = \mathbf{0}$ in (2.14) and (2.15), and $T^2_G/|G|$ is the F-statistic for testing $H_0 : \boldsymbol{\beta}_G = \mathbf{0}$ when $\widehat{\sigma}$ is the degree adjusted estimate of noise level based on the residuals of the least squares estimator. Of course, we need to relax the requirement of the orthogonality condition $\mathbf{P}_G\mathbf{Q}_{G_k\setminus G} = \mathbf{0}$ for all $G_k \not\subseteq G$ in the high-dimensional case.

Analytically, the key is to prove the upper bound $\|\mathrm{Rem}_G/\sigma\|_2 = o_{\mathbb{P}}(1)$ in (2.23). To this end we use the formula in (2.14) and the working assumption in (2.31) to obtain

$$\begin{aligned} \|\mathrm{Rem}_G\|_2 &\le \Big(\max_{G_k\not\subseteq G} M_k\omega^{-1}_{*,k}\|\mathbf{P}_G\mathbf{Q}_{G_k}\|_S\Big)\sum_{G_k\not\subseteq G}\omega_{*,k}\|\widehat{\boldsymbol{\mu}}^{(init)}_{G_k} - \boldsymbol{\mu}_{G_k}\|_2 \\ &= \mathcal{O}_{\mathbb{P}}\Big(\frac{s + g\log M}{n^{1/2}}\Big)\Big(\max_{G_k\not\subseteq G} M_k\omega^{-1}_{*,k}\|\mathbf{P}_G\mathbf{Q}_{G_k}\|_S\Big), \end{aligned} \tag{2.32}$$

where

$$\omega_{*,k} = \sqrt{|G_k|/n} + \sqrt{(2/n)\log M} \text{ and } M_k = \max_{\|\mathbf{X}_{G_k}\boldsymbol{u}_{G_k}\|_2=1}\|\mathbf{X}_{G_k\setminus G}\boldsymbol{u}_{G_k\setminus G}\|_2.$$

We note that $M_k = 1$ when $\mathbf{X}_{G_k}^T\mathbf{X}_{G_k}/n = \mathbf{I}_{d_k\times d_k}$. Since $(s+g\log M)/n$ is the order of the mixed $\ell_{2,1}$ error bound for $\widehat{\boldsymbol{\beta}}$, we may treat

$$\eta_G = \max_{G_k\not\subseteq G} M_k\omega_{*,k}^{-1}\|\mathbf{P}_G\mathbf{Q}_{G_k}\|_S$$

as a scalar bias factor. The error bound in (2.32) motivates the following extension of (2.10):

$$\mathbf{P}_G = \arg\min_{\mathbf{P}}\Big\{\|\mathbf{P}\mathbf{Q}_G^{\perp}\|_S : \mathbf{P}=\mathbf{P}^2=\mathbf{P}^T,\ \|\mathbf{P}_G\mathbf{Q}_{G_k\setminus G}\|_S \le \omega_k'\ \forall\ G_k\not\subseteq G\Big\}. \tag{2.33}$$

We say that $\mathbf{P}_G$ is a feasible solution of (2.33) if it satisfies all the constraints. The optimization problem (2.33) is a generalization of (2.10) and provides geometric insights. As $(\mathbf{P}_G\mathbf{Q}_G)^{\dagger}$ is a multivariate noise factor for the inference of $\boldsymbol{\mu}_G^*$, we may define $\tau_G = \|(\mathbf{P}_G\mathbf{Q}_G)^{\dagger}\|_S$ as a scalar noise factor. The quantity $\|\mathbf{P}_G\mathbf{Q}_G^{\perp}\|_S$, which is the so-called 'gap' between the subspaces spanned by $\mathbf{P}_G$ and $\mathbf{Q}_G$, equals $(1-\tau_G^{-2})^{1/2}$. Thus, minimizing $\|\mathbf{P}_G\mathbf{Q}_G^{\perp}\|_S$ is equivalent to minimizing the noise factor τ_G. This minimization is done subject to upper-bounds on the components $\|\mathbf{P}_G\mathbf{Q}_{G_k\setminus G}\|_S$ of the bias factor. Thus, (2.33) is an extension of (2.10) as we discussed immediately after (2.10). When $p<n$ and $\omega_k'=0$, $\boldsymbol{P}_G$ in (2.33) is the projection to the orthogonal complement of $\sum_{G_k\not\subseteq G}\mathcal{R}(\mathbf{X}_{G_k\setminus G})$ in $\mathcal{R}(\mathbf{X})$, or equivalently the linear space $\big(\prod_{G_k\not\subseteq G}\mathbf{Q}_{G_k\setminus G}^{\perp}\big)\mathcal{R}(\mathbf{X})$.

In the following theorem, we provide a summary of the analysis we have carried out above.

Theorem 2. *Let $\boldsymbol{P}_G$ be a feasible solution of (2.33) satisfying $\|\boldsymbol{P}_G\boldsymbol{Q}_G^{\perp}\|_S<1$, and $\widehat{\boldsymbol{\beta}}_G$ be as in (2.12) with $\boldsymbol{Z}_G=\boldsymbol{P}_G$ and certain $\{\widehat{\boldsymbol{\beta}}^{(init)},\widehat{\sigma}\}$ satisfying (2.31). Suppose $\boldsymbol{\varepsilon}\sim \mathsf{N}_n(\mathbf{0},\sigma^2\boldsymbol{I}_n)$, rank$(\boldsymbol{X}_G)=|G|$, and*

$$\frac{|G|}{n}\to 0,\qquad \frac{s+g\log M}{n^{1/2}}\left(\frac{|G|^{1/2}}{n^{1/2}}+\max_{G_k\not\subseteq G}M_k\frac{\omega_k'}{\omega_{*,k}}\right)\to 0, \tag{2.34}$$

with the M_k in (2.32). Then, (2.24) and (2.30) hold.

Proof of Theorem 2. Since $\|\mathbf{P}_G\mathbf{Q}_G^{\perp}\|_S<1$, we have rank$(\mathbf{P}_G\mathbf{X}_G)=$ rank$(\mathbf{X}_G)=|G|$, so that the condition of Proposition 1 (i) holds, which implies the condition of Proposition 1 (ii) with $k_G=|G|$. It follows from (2.31), (2.32), (2.34) and the feasibility of $\mathbf{P}_G$ in (2.33) that (2.23) holds, which implies (2.24) and (2.30). Note that (2.31) and (2.34) imply $|\sigma/\widehat{\sigma}-1| = o_{\mathbb{P}}(|G|^{-1/2})+O_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(|G|^{-1/2})$ in the proof for the first component of (2.23). □

A modification of (2.33), which removes the factors M_k in condition (2.34), is to re-parameterize the effect of the k-th group by writing

$$\mathbf{X}_{G_k}\boldsymbol{\beta}_{G_k} = \widetilde{\mathbf{X}}_{G_k\cap G}\boldsymbol{\beta}_{G_k\cap G}+\mathbf{X}_{G_k\setminus G}\widetilde{\boldsymbol{\beta}}_{G_k\setminus G},$$

where $\widetilde{\mathbf{X}}_{G_k\cap G} = \mathbf{Q}^{\perp}_{G_k\setminus G}\mathbf{X}_{G_k\cap G}$ and $\widetilde{\boldsymbol{\beta}}_{G_k\setminus G}$ is a solution of $\mathbf{X}_{G_k\setminus G}\widetilde{\boldsymbol{\beta}}_{G_k\setminus G} = \mathbf{Q}_{G_k\setminus G}\mathbf{X}_{G_k}\boldsymbol{\beta}_{G_k}$. We recall that $\mathbf{Q}_{G_k\setminus G}$ is the orthogonal projection to the column space of $\mathbf{X}_{G_k\setminus G}$. As this within-group re-parameterization retains $\boldsymbol{\beta}_{G_k\cap G}$ and $\mathbf{X}_{G_k\setminus G}$,

$$\boldsymbol{y} = \widetilde{\mathbf{X}}_G\boldsymbol{\beta}_G + \sum_{G_k\not\subseteq G}\mathbf{Q}_{G_k\setminus G}\boldsymbol{\mu}_{G_k} + \boldsymbol{\varepsilon} = \widetilde{\mathbf{X}}_G\boldsymbol{\beta}_G + \sum_{G_k\not\subseteq G}\mathbf{X}_{G_k\setminus G}\widetilde{\boldsymbol{\beta}}_{G_k\setminus G} + \boldsymbol{\varepsilon},$$

where $\widetilde{\mathbf{X}}_G$ is the $n\times|G|$ matrix given by $\widetilde{\mathbf{X}}_G\boldsymbol{v}_G = \sum_{k=1}^M\big(\mathbf{Q}^{\perp}_{G_k\setminus G}\mathbf{X}_{G_k\cap G}\big)\boldsymbol{v}_{G\cap G_k}$. As $\widetilde{\mathbf{X}}_{G_k\cap G}$ is orthogonal to $\mathbf{X}_{G_k\setminus G}$, we have $M_k = 1$ after re-parametrization. Moreover, the strong group sparsity condition $\mathrm{supp}(\boldsymbol{\beta}^*)\subset G_{S^*}$ and the working assumption (2.31) are invariant under the re-parameterization. We note that $\widetilde{\mathbf{X}}_G = \mathbf{X}_G$ when $\mathbf{X}^T_{G_k}\mathbf{X}_{G_k}/n = \mathbf{I}_{G_k\times G_k}$ for all k with $0 < |G_k\setminus G| < |G_k|$. Let $\widetilde{\mathbf{Q}}_G$ be the projection to the column space of $\widetilde{\mathbf{X}}_G$. The optimization scheme and statistical methods are changed accordingly as follows:

$$\begin{aligned}
&\mathbf{P}_G = \arg\min_{\mathbf{P}}\Big\{\|\mathbf{P}\widetilde{\mathbf{Q}}^{\perp}_G\|_S : \mathbf{P} = \mathbf{P}^2 = \mathbf{P}^T,\ \|\mathbf{P}_G\mathbf{Q}_{G_k\setminus G}\|_S \le \omega'_k\ \forall\ k\Big\},\\
&\widehat{\boldsymbol{\beta}}_G = (\mathbf{P}_G\widetilde{\mathbf{X}}_G)^{\dagger}\mathbf{P}_G\left(\boldsymbol{y} - \sum_{G_k\not\subseteq G}\mathbf{Q}_{G_k\setminus G}\widehat{\boldsymbol{\mu}}^{(init)}_{G_k}\right),\ \text{when }\mathrm{rank}(\mathbf{P}_G\widetilde{\mathbf{X}}_G) = |G|,\\
&\hspace{30em}(2.35)\\
&T_G = \frac{1}{\widehat{\sigma}}\left\|\mathbf{P}_G\left(\boldsymbol{y} - \sum_{G_k\not\subseteq G}\mathbf{Q}_{G_k\setminus G}\widehat{\boldsymbol{\mu}}^{(init)}_{G_k}\right)\right\|_2.
\end{aligned}$$

With $\{\mathbf{X}_G, \mathbf{Q}_G\}$ replaced by $\{\widetilde{\mathbf{X}}_G, \widetilde{\mathbf{Q}}_G\}$, our analysis yields the following theorem.

Theorem 3. *Let $\boldsymbol{P}_G$, $\widehat{\boldsymbol{\beta}}_G$ and T_G be given by (2.35) with $\|\boldsymbol{P}_G\widetilde{\boldsymbol{Q}}^{\perp}_G\|_S < 1$. Suppose $\boldsymbol{\varepsilon}\sim\mathsf{N}_n(\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$, rank$(\boldsymbol{X}_G) = |G|$, and (2.31) and (2.34) hold with $M_k = 1$. Then, (2.24) and (2.30) hold with $\{\boldsymbol{X}_G, \boldsymbol{Q}_G\}$ replaced by $\{\widetilde{\boldsymbol{X}}_G, \widetilde{\boldsymbol{Q}}_G\}$.*

Remark 1. It is worthwhile to note that Theorems 2 and 3 only require a feasible solution satisfying $\|\mathbf{P}_G\mathbf{Q}^{\perp}_G\|_S < 1$ and $\|\mathbf{P}_G\widetilde{\mathbf{Q}}^{\perp}_G\|_S < 1$ respectively, which can be directly verified for any given $\mathbf{P}_G$. Still, the optimality criterion on $\mathbf{P}_G$ aims to have smaller confidence regions and more powerful tests through (2.24). In practice, it suffices to find a feasible solution with $\|\mathbf{P}_G\mathbf{Q}^{\perp}_G\|_S$ or $\|\mathbf{P}_G\widetilde{\mathbf{Q}}^{\perp}_G\|_S$ reasonably bounded away from 1. As the optimization problems in (2.33) and (2.35) are still somewhat abstract for the moment, in the following we prove the feasibility of $\mathbf{P}_G$ in (2.33) for sub-Gaussian designs and describe penalized regression methods to find feasible solutions of (2.33) and (2.35).

2.6. Feasibility of relaxed orthogonal projection for random designs

In this subsection, we discuss the existence of feasible solutions of the optimization in (2.33) for a sub-Gaussian design matrix satisfying (2.28) with $\mathbb{E}\mathbf{X} = \mathbf{0}$

and a positive-definite population Gram matrix $\mathbb{E}(\mathbf{X}^T\mathbf{X}/n) = \mathbf{\Sigma}$. The feasibility is established under the assumption of the groupwise regression model as described in (2.26).

We group the effects in the linear regression model (2.26) as follows:

$$\mathbf{X}_G = \mathbf{X}_{-G}\mathbf{\Gamma}_{-G,G} + \mathbf{Z}_G^o = \sum_{k=1}^{M} \mathbf{X}_{G_k\setminus G}\mathbf{\Gamma}_{G_k\setminus G,G} + \mathbf{Z}_G^o, \tag{2.36}$$

where $\mathbf{\Gamma}_{-G,G} = \mathbf{\Sigma}_{-G,-G}^{-1}\mathbf{\Sigma}_{-G,G}$. Under this model assumption, $\mathbf{Z}_G^o$ is the true residual after projection of $\mathbf{X}_G$ onto the range of $\mathbf{X}_{-G}$. Let $\mathbf{P}_G^o$ be the orthogonal projection to the column space of $\mathbf{Z}_G^o$,

$$\mathbf{P}_G^o = \mathbf{Z}_G^o\Big((\mathbf{Z}_G^o)^T\mathbf{Z}_G^o\Big)^{\dagger}(\mathbf{Z}_G^o)^T. \tag{2.37}$$

The following theorem establishes the distributional convergence results in (2.24) and (2.30) for $\widehat{\boldsymbol{\beta}}_G$ by establishing the feasibility of $\mathbf{P}_G^o$ as a solution of the optimization scheme in (2.33).

Theorem 4. *Suppose the sub-Gaussian condition (2.28) holds with*

$$0 < c_* \le eigen(\mathbf{\Sigma}) \le c^* \text{ and fixed } \{v_0, c_*, c^*\}.$$

Let $\omega'_k = \xi n^{-1/2}\big(\sqrt{|G| + |G_k \setminus G|} + \sqrt{\log(M/\delta)}\big)$.
(i) Let $\lambda_{\min}$ *be the smallest eigenvalue of* $\{\mathbf{\Sigma}_{G,G}^{-1/2}(\mathbf{\Sigma}^{-1})_{G,G}\mathbf{\Sigma}_{G,G}^{-1/2}\}^{1/2}$*, and let* $\xi n^{-1/2}\big(\sqrt{|G|} + \sqrt{\log(M/\delta)}\big) \le \eta_n$*, and* $a_n = \lambda_{\min}(1-\eta_n)/(1+\eta_n)$*. Then, there exist numerical constants* $\epsilon_0 \in (0,1)$ *and* $\xi_0 < \infty$ *such that when* $\xi \ge \xi_0 v_0$ *and* $\eta_n \le \epsilon_0$*,*

$$\mathbb{P}\left\{\begin{array}{l} (2.33) \textit{ has a feasible solution } \boldsymbol{P}_G \textit{ with} \\ rank(\boldsymbol{P}_G) = rank(\boldsymbol{P}_G\boldsymbol{X}_G) = |G| \textit{ and } \|\boldsymbol{P}_G\boldsymbol{Q}_G^{\perp}\|_S \le \sqrt{1-a_n^2} \end{array}\right\} \ge 1-\delta. \tag{2.38}$$

(ii) Suppose the strong sparsity condition the sample size condition (2.29) hold and that $\{\widehat{\boldsymbol{\beta}}^{(init)}, \sigma\}$ *is as in Theorem 1. Then, the working assumption (2.31) holds.*
(iii) Suppose the working assumption (2.31) and the sample size condition (2.29) hold. Then, (2.24) and (2.30) hold.

Theorem 4 removes the requirement of the knowledge of $\mathbf{Z}_G^o$ in Theorem 1. It shows the existence of at least one feasible solution of (2.33) and that for such a choice of $\mathbf{P}_G$, the χ^2 based inference can be carried out as in (2.24) and (2.30). However, (2.33) is not a convex program. In Subsection 2.7 we will describe group Lasso programs as convexation of (2.33).

The proof of Theorem 4 requires the following lemma on the probabilistic control of the spectral norm of the product of two random matrices with sub-Gaussian rows. As an extension of that result, spectral norm control of the product of two orthogonal projection matrices is also obtained. These probabilistic bounds in Lemma 1 are of independent interest. See Remark 2 for more details.

Lemma 1. *Let $\boldsymbol{B}_k$ be deterministic matrices with with p rows and rank$(\boldsymbol{B}_k) = r_k$ for $k = \{1, 2\}$. Let $\boldsymbol{P}_k$ be the projection to the range of $\boldsymbol{X}\boldsymbol{B}_k$ and*

$$\boldsymbol{\Omega}_{1,2} = ((\boldsymbol{B}_1^T \boldsymbol{\Sigma} \boldsymbol{B}_1)^\dagger)^{1/2} \boldsymbol{B}_1^T \boldsymbol{\Sigma} \boldsymbol{B}_2 ((\boldsymbol{B}_2^T \boldsymbol{\Sigma} \boldsymbol{B}_2)^\dagger)^{1/2}.$$

Let $r = \text{rank}(\boldsymbol{\Omega}_{1,2})$ and $1 \ge \lambda_1 \ge \cdots \ge \lambda_r > 0$ be the nonzero singular values of $\boldsymbol{\Omega}_{1,2}$. Define $\lambda_{\min} = \lambda_r I\{r = r_1 = r_2\}$. Suppose (2.28) holds. Then, there exists a numerical constant $C_0 > 1$ such that when $C_0 v_0 \sqrt{t/n + (r_1 + r_2)/n} < \epsilon_0 < 1$,

$$\mathbb{P}\left\{ \|((\boldsymbol{B}_1^T \boldsymbol{\Sigma} \boldsymbol{B}_1)^\dagger)^{1/2} \boldsymbol{B}_1^T (\boldsymbol{X}^T \boldsymbol{X}/n) \boldsymbol{B}_2 ((\boldsymbol{B}_2^T \boldsymbol{\Sigma} \boldsymbol{B}_2)^\dagger)^{1/2} - \boldsymbol{\Omega}_{1,2}\|_S \le \epsilon_0 \right\} \ge 1 - e^{-t}, \tag{2.39}$$

and

$$\mathbb{P}\left\{ \|\boldsymbol{P}_1 \boldsymbol{P}_2\|_S \le \frac{\lambda_1(1+\epsilon_0)}{1-\epsilon_0}, \|\boldsymbol{P}_1 \boldsymbol{P}_2^\perp\|_S^2 \le 1 - \left(\frac{\lambda_{\min}(1-\epsilon_0)}{1+\epsilon_0}\right)^2 \right\} \ge 1 - e^{-t}. \tag{2.40}$$

Moreover, $\lambda_1 < 1$ iff rank$(\boldsymbol{B}_1, \boldsymbol{B}_2) = r_1 + r_2$ and $\lambda_{\min} > 0$ iff rank$(\boldsymbol{B}_1^T \boldsymbol{B}_2) = r_1 = r_2$.

We have moved the proof of Lemma 1 to the Appendix to avoid a distraction from the main results of this section. Based on Lemma 1, we prove Theorems 1 and 4 as follows.

Proofs of Theorems 1 and 4. By (2.37), $\mathbf{P}_G^o$ is the orthogonal projection to the range of $\mathbf{Z}_G^o = \mathbf{X}\mathbf{B}_G^o$ with $\mathbf{B}_G^o = (\boldsymbol{\Sigma}^{-1})_{*,G}(\boldsymbol{\Sigma}^{-1})_{G,G}^{-1}$. By definition, $\mathbf{Q}_{G_k \setminus G}$ is the projection to the range of $\mathbf{X}_{G_k\setminus G} = \mathbf{X}\mathbf{B}_{G_k \setminus G}$ and $\mathbf{Q}_G$ to the range of $\mathbf{X}_G = \mathbf{X}\mathbf{B}_G$, where $\mathbf{B}_{G_k\setminus G}$ and $\mathbf{B}_G$ are 0-1 diagonal matrices projecting to the indicated spaces. Define $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{G,G}^{-1/2}\{(\boldsymbol{\Sigma}^{-1})_{G,G}\}^{1/2}$. We have $\mathbf{B}_{G_k\setminus G}^T \boldsymbol{\Sigma} \mathbf{B}_G^o = \boldsymbol{\Sigma}_{G_k\setminus G,*}\mathbf{B}_G^o = 0$, $\mathbf{B}_G^T \boldsymbol{\Sigma}\mathbf{B}_G^o = \boldsymbol{\Sigma}_{G,*}\mathbf{B}_G^o = (\boldsymbol{\Sigma}^{-1})_{G,G} = (\mathbf{B}_G^o)^T \boldsymbol{\Sigma} \mathbf{B}_G^o$ and

$$\begin{aligned}(\mathbf{B}_G^T \boldsymbol{\Sigma} \mathbf{B}_G)^{-1/2} \mathbf{B}_G^T \boldsymbol{\Sigma} \mathbf{B}_G^o \big((\mathbf{B}_G^o)^T \boldsymbol{\Sigma} \mathbf{B}_G^o\big)^{-1/2} &= \boldsymbol{\Sigma}_{G,G}^{-1/2}\{(\boldsymbol{\Sigma}^{-1})_{G,G}\}^{1/2} \\ &= \boldsymbol{\Omega} \in \mathbb{R}^{|G|\times|G|}.\end{aligned}$$

Moreover, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{G,G}^{-1/2}\{(\boldsymbol{\Sigma}^{-1})_{G,G}\}^{1/2}$ is a $|G| \times |G|$ matrix of rank $|G|$ and the smallest singular value of $\boldsymbol{\Omega}$ is $\lambda_{\min}$. Thus, by (2.40) of Lemma 1 and the definition of ω_k' and a_n,

$$\mathbb{P}\Big\{ \|\boldsymbol{P}_G \mathbf{Q}_{G_k\setminus G}\|_S \le \omega_k' \ \forall k \le M, \ \|\boldsymbol{P}_G \mathbf{Q}_G^\perp\|_S \le \sqrt{1 - a_n^2} \Big\} \ge 1 - \delta.$$

This yields (2.38). Moreover, (2.38) also holds when $\mathbf{P}_G = \mathbf{P}_G^o$ or equivalently $\mathbf{Z}_G = \mathbf{Z}_G^o$ is used as in Theorem 1. As part (ii) of Theorem 4 restates Theorem 7 in Section 3, it remains to prove $\max_{G_k \setminus G \neq \emptyset} M_k = O_{\mathbb{P}}(1)$ in view of Theorem 2. To this end, we notice that due to the condition $|G_k| + g \log M \ll n$, (2.39) of Lemma 1 with $\mathbf{B}_1 = \mathbf{B}_2$ implies $\|\mathbf{X}_A^T \mathbf{X}_A/n - \mathbf{\Sigma}_{A,A}\|_S = o_{\mathbb{P}}(1)$ for both $A = G_k$ and $A = G_k \setminus G$ and all k with $G_k \setminus G \neq \emptyset$, so that $\max_{G_k \setminus G \neq \emptyset} M_k = o_{\mathbb{P}}(1) + O(1)$. □

Remark 2. Since Lemma 1 is a crucial ingredient for Theorems 1 and 4, we highlight a few key points. Let us write $p = p_1 + p_2$ and $\mathbf{I}_p = [\mathbf{I}_{p\times p_1}\, \mathbf{I}_{p\times p_2}]$. Consider the choices: $\mathbf{B}_1 = \mathbf{I}_{p\times p_1}$ and $\mathbf{B}_2 = \mathbf{I}_{p\times p_2}$. Also consider the partition $\mathbf{X} = [\mathbf{X}_1\, \mathbf{X}_2]$ so that $\mathbf{X}_i = \mathbf{X}\mathbf{B}_i$. Writing

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \text{ where } \mathbf{\Sigma}_{11} \in \mathbb{R}^{p_1\times p_1}, \mathbf{\Sigma}_{12} \in \mathbb{R}^{p_1\times p_2}, \mathbf{\Sigma}_{22} \in \mathbb{R}^{p_2\times p_2},$$

it follows that $\text{cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{\Sigma}_{12}$. For such choices, Lemma 1 gives,

$$\|\mathbf{\Sigma}_{11}^{-1/2}\Big(\mathbf{X}_1^T \mathbf{X}_2/n - \mathbf{\Sigma}_{12}\Big)\mathbf{\Sigma}_{22}^{-1/2}\|_S \leq C\sqrt{t/n + (p_1+p_2)/n} \tag{2.41}$$

with probability at least $1 - e^{-t}$. This result provides a spectral norm bound on the cross-product of two correlated random matrices with sub-Gaussian rows. The probability bound in (2.41) is a generalization of a similar result for product of two mutually independent random matrices with iid $\mathsf{N}(0,1)$ entries, given in Proposition D.1 in the supplement to Ma (2013). Control of spectral norm of product of random and deterministic matrices have been studied as well; see Vershynin (2011), Rudelson and Vershynin (2013) etc. In particular, spectral norm concentration of product of a fixed projection matrix and a random matrix have been derived in (Rudelson and Vershynin, 2013, Remark 3.3). In comparison, our results in (2.40) studies product of two projection matrices with their range being column spaces of correlated random matrices with sub-Gaussian rows.

2.7. Finding feasible solutions and construction of tests

While (2.38) of Theorem 4 guarantees a feasible solution of (2.33), the practicality of the optimization scheme (2.33) has not yet been addressed. We discuss here penalized multivariate regression methods for finding feasible solutions of (2.33) and (2.35). As the only difference between (2.33) and (2.35) is the respective use of $\mathbf{X}_G$ and $\widetilde{\mathbf{X}}_G$, we provide formulas here only for (2.33), with the understanding that formulas for (2.35) can be generated in the same way with $\mathbf{X}_G$ replaced by $\widetilde{\mathbf{X}}_G$.

The optimization problem in (2.33) is carried out over the non-convex space of orthogonal projection matrices. In the following, we provide a convex program for obtaining such orthogonal projection matrices under the linear regression framework of (2.36). In model (2.36), a general formulation of the penalized multivariate regression is

$$\widehat{\mathbf{\Gamma}}_{-G,G} = \underset{\mathbf{\Gamma}_{-G,G}}{\arg\min} \left\{ \frac{1}{2n} \Big\| \mathbf{X}_G - \sum_{G_k \not\subseteq G} \mathbf{X}_{G_k\setminus G} \mathbf{\Gamma}_{G_k\setminus G, G} \Big\|_F^2 + R(\mathbf{\Gamma}_{-G,G}) \right\}, \tag{2.42}$$

where $\|\cdot\|_F$ is the Frobenius norm and $R(\mathbf{\Gamma}_{-G,G})$ is a penalty function. Define

$$\mathbf{Z}_G = \mathbf{X}_G - \sum_{G_k \not\subseteq G} \mathbf{X}_{G_k \setminus G} \widehat{\mathbf{\Gamma}}_{G_k \setminus G, G}, \quad \mathbf{P}_G = \mathbf{Z}_G(\mathbf{Z}_G^T \mathbf{Z}_G)^{-1}\mathbf{Z}_G^T. \tag{2.43}$$

Our main interest is to find a feasible solution of (2.33) and (2.35), not to estimate $\mathbf{\Gamma}_{-G,G}$. The following weighted group nuclear penalty matches the dual of the constraint in (2.33) and (2.35):

$$R(\mathbf{\Gamma}_{-G,G}) = \sum_{G_k \not\subseteq G} \frac{\xi\omega_k''}{n^{1/2}} \left\| \mathbf{X}_{G_k \setminus G} \mathbf{\Gamma}_{G_k \setminus G, G} \right\|_N. \tag{2.44}$$

Recall that nuclear norm of a matrix $\mathbf{A}$, denoted $\|\mathbf{A}\|_N$, is the sum of absolute values of the singular values of $\mathbf{A}$. It follows from the KKT conditions for (2.42) with (2.44) that

$$\left\| \mathbf{Q}_{G_k \setminus G} \mathbf{Z}_G / \sqrt{n} \right\|_S \le \xi\omega_k''. \tag{2.45}$$

If we set $\omega_k'' = \omega_k$ in (2.44), condition (2.34) follows from

$$\frac{|G|}{n} \to 0, \quad \frac{s + g\log M}{n^{1/2}} \left(\frac{|G|^{1/2}}{n^{1/2}} + \xi \|(\mathbf{Z}_G^T \mathbf{Z}_G / n)^{-1/2}\|_S \right) \to 0, \tag{2.46}$$

provided $\max_{G_k \not\subseteq G} M_k = O(1)$ in the case of Theorem 2. Moreover, as in van de Geer (2014), under the assumption $\lambda_{\min}(\mathbf{Z}_G) > c > 0$, only $(s + g\log M)/n^{1/2} + |G|/n \to 0$ suffices.

When the group sizes are not too large, one may consider replacing the weighted group nuclear penalty with a weighted group Frobenius penalty:

$$R(\mathbf{\Gamma}_{-G,G}) = \sum_{G_k \not\subseteq G} \frac{\xi\omega_k''}{n^{1/2}} \left\| \mathbf{X}_{G_k \setminus G} \mathbf{\Gamma}_{G_k \setminus G, G} \right\|_F. \tag{2.47}$$

The KKT conditions for (2.42) with (2.47) yield

$$\left\| \mathbf{Q}_{G_k \setminus G} \mathbf{Z}_G / \sqrt{n} \right\|_S \le \left\| \mathbf{Q}_{G_k \setminus G} \mathbf{Z}_G / \sqrt{n} \right\|_F \le \xi\omega_k'',$$

so that (2.46) is still valid. However, this second layer of inequality indicates that the resulting procedure may not be as efficient as the (2.44) penalty. In any case, as discussed in Remark 1, it is reasonable to proceed with the computed $\boldsymbol{Z}_G$ as long as the resulting $\|\mathbf{P}_G \mathbf{Q}_G^{\perp}\|_S$ is not too close to 1. One important benefit of the formulation of the groupwise penalty as in (2.47) is that it can be conveniently computed using the standard group Lasso algorithms; see Yuan and Lin (2006), Huang, Breheny and Ma (2012) etc. As we will show in Section 4, group Lasso performs well for empirical studies. We summarize our proposal and main results as follows.

Summary: Statistical inference for groups of variables can be carried out as follows:

- Given $(\boldsymbol{y}, \mathbf{X})$ and a group structure $\{G_j : 1 \le j \le M\}$, construct the initial estimates $(\widehat{\boldsymbol{\beta}}^{(init)}, \widehat{\sigma})$ via the scaled group Lasso (2.25) or any alternative leading to (2.31).
- Given a variable group G of interest, construct relaxed projection estimate $\mathbf{P}_G = \mathbf{Z}_G(\mathbf{Z}_G^T\mathbf{Z}_g)^{-1}\mathbf{Z}_G^T$ by the penalized procedure (2.42) and (2.43) with the penalty function (2.44) or (2.47).
- Carry out statistical inference according to (2.24) and (2.30)

Benefit of group sparsity: Existing sample size condition for statistical inference of a univariate parameter at $n^{-1/2}$ rate requires,

$$n \gg \|\boldsymbol{\beta}^*\|_0^2(\log p)^2.$$

See for exampe Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014a). As discussed below (1.2), direct application of these results to approximate chi-square group inference requires an extra factor $|G|$:

$$n \gg |G| \times \|\boldsymbol{\beta}^*\|_0^2(\log p)^2.$$

If the true parameter $\boldsymbol{\beta}^*$ is (g, s) strong group sparse with $s \asymp \|\boldsymbol{\beta}^*\|_0$, the sample size conditions in (2.34), (2.29) and (2.46) clearly demonstrate the benefit of group sparsity by incorporating the smaller estimation error bound as in Huang and Zhang (2010) and removing the extra $|G|$. In particular, our sample size requirement becomes the much weaker

$$n \gg \big(s + g\log p\big)^2$$

for approximate chi-square inference when $|G| \lesssim \min_{G_k \not\subseteq G}\{|G_k| + \log(M/\delta)\}$ in (2.29) or $\xi\|(\mathbf{Z}_G^T\mathbf{Z}_G/n)^{-1/2}\|_S = O(1)$ in (2.46).

3. Verification of working assumption

The analysis in the preceding section established the benefits of grouping in constructing ℓ_2 type statistical inference procedures for variable groups. One key aspect of our analysis was the working assumption in (2.31). These results showed a faster convergence rate for the scale parameter estimate and the coefficient parameter estimate. As promised, in this section we will establish the bona fides of (2.31) under the strong group sparsity assumption in (2.1).

Generally, for high dimensional regression problems, certain regularity conditions on the the design matrix is required for estimation as well as prediction consistency. In the following Subsection 3.1, we discuss similar assumptions on the design matrix $\mathbf{X}$ that ensure the consistency results in (2.31). We also derive estimation and prediction consistency result for the non-scaled group Lasso problem in (1.3) in Theorem 5 as an illustration. The main result of this section is Theorem 6 and Corollary 1 in Subsection 3.2 and Theorem 7 in Subsection 3.3 that establish the working assumption (2.31).

3.1. Group Lasso and conditions on the design matrix

In the Lasso problem, performance bounds of the estimator are derived based on various conditions on the design matrix, for example, the restricted isometry property (Candes and Tao, 2005), the sparse Riesz condition (Zhang and Huang, 2008), the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009; Koltchinskii, 2009), the compatibility condition (van de Geer, 2007; van de Geer and Bühlmann, 2009), and cone invertibility conditions (Ye and Zhang, 2010). van de Geer and Bühlmann (2009) showed that the compatibility condition is weaker than the restricted eigenvalues condition for the prediction and ℓ_1 loss, while Ye and Zhang (2010) showed that both conditions can be weakened by cone invertibility conditions. In the following, we define grouped versions of such conditions, which will be used in our study.

Let us first define a groupwise mixed norm cone for $T \subset \{1\cdots,M\}$ and $\xi \geq 0$ as

$$\mathscr{C}^{(G)}(\xi,\boldsymbol{\omega},T)=\Big\{\boldsymbol{u}:\textstyle\sum_{j\in T^c}\omega_j\|\boldsymbol{u}_{G_j}\|_2\leq\xi\sum_{j\in T}\omega_j\|\boldsymbol{u}_{G_j}\|_2\neq 0\Big\}. \tag{3.1}$$

Let $T^* = \{1\cdots,M\}$ and $T \subseteq T' \subseteq T^*$. Following Nardi and Rinaldo (2008) and Lounici et al. (2011), the restricted eigenvalue (RE) is defined as

$$\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T')=\inf_{\boldsymbol{u}}\left\{\frac{\|\mathbf{X}\boldsymbol{u}\|_2}{\sqrt{n}\|\boldsymbol{u}_{G_{T'}}\|_2}:\boldsymbol{u}\in\mathscr{C}^{(G)}(\xi,\boldsymbol{\omega},T)\right\}. \tag{3.2}$$

For the weighted $\ell_{2,1}$ norm, the groupwise compatibility constant (CC) can be defined as

$$\mathrm{CC}^{(G)}(\xi,\boldsymbol{\omega},T)=\inf_{\boldsymbol{u}}\left\{\frac{\|\mathbf{X}\boldsymbol{u}\|_2\big(\sum_{j\in T}\omega_j^2\big)^{1/2}}{\sqrt{n}\sum_{j\in T}\omega_j\|\boldsymbol{u}_{G_j}\|_2}:\boldsymbol{u}\in\mathscr{C}^{(G)}(\xi,\boldsymbol{\omega},T)\right\}. \tag{3.3}$$

We note that $\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T)$ and the somewhat larger $\mathrm{CC}^{(G)}(\xi,\boldsymbol{\omega},T)$ are aimed at the prediction and the weighed $\ell_{2,1}$ estimation errors, while the smaller $\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T^*)$ is aimed at the ℓ_2 estimation error.

We also introduce the notion of groupwise cone invertibility factor and its sign-restricted version. For $q \geq 1$, the cone invertibility factor (CIF) is defined as

$$\mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T')=\inf_{\boldsymbol{u}\in\mathscr{C}^{(G)}(\xi,\boldsymbol{\omega},T)}\frac{\max_j\Big[\omega_j^{-1}\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{u}\|_2\Big]\big(\sum_{j\in T}\omega_j^2\big)^{1/q}}{n\big(\sum_{j\in T'}\omega_j^2(\|\boldsymbol{u}_{G_j}\|_2/\omega_j)^q\big)^{1/q}}. \tag{3.4}$$

We note that $\big(\sum_{j\in T'}\omega_j^2(\|\boldsymbol{u}_{G_j}\|_2/\omega_j)^q\big)^{1/q}=\|\boldsymbol{u}\|_2$ when $T'=T^*$ and $q=2$. Define

$$\mathscr{C}_-^{(G)}(\xi,\boldsymbol{\omega},T)=\Big\{\boldsymbol{u}:\boldsymbol{u}\in\mathscr{C}^{(G)}(\xi,\boldsymbol{\omega},T),\ \boldsymbol{u}_{G_j}^T\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{u}\leq 0\ \forall j\in T^c\Big\}, \tag{3.5}$$

as a sign-restricted cone. We extend the CIF to the groupwise sign-restricted cone invertibility factor (SCIF) as

$$\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T') = \inf_{\boldsymbol{u}\in\mathscr{C}_-^{(G)}(\xi,\boldsymbol{\omega},T)} \frac{\max_j \left[\omega_j^{-1}\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{u}\|_2\right]\big(\sum_{j\in T}\omega_j^2\big)^{1/q}}{n\big(\sum_{j\in T'}\omega_j^2(\|\boldsymbol{u}_{G_j}\|_2/\omega_j)^q\big)^{1/q}}. \tag{3.6}$$

Similar to the RE and CC, $\mathrm{CIF}_1^{(G)}(\xi,\boldsymbol{\omega},T,T)$ and $\mathrm{SCIF}_1^{(G)}(\xi,\boldsymbol{\omega},T,T)$ are aimed at the prediction and weighted $\ell_{2,1}$ losses, while $\mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T^*)$ and $\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T^*)$ is aimed at the weighted loss $\big(\sum_{j=1}^M\omega_j^2(\|\boldsymbol{u}_{G_j}\|_2/\omega_j)^q\big)^{1/q}$. We note that the weighted $\ell_{2,q}$ norm is identical to the ℓ_2 norm for $q=2$. For $\boldsymbol{u}\in\mathscr{C}_-^{(G)}(\xi,\boldsymbol{\omega},T)$,

$$\|\mathbf{X}\boldsymbol{u}\|_2^2/\max_j(\omega_j^{-1}\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{u}\|_2) \le \textstyle\sum_{j\in T}\omega_j\|\boldsymbol{u}_{G_j}\|_2 \le \|\boldsymbol{u}_{G_T}\|_2\big(\sum_{j\in T}\omega_j^2\big)^{1/2}$$

by the sign restriction and the Cauchy-Schwarz inequality, so that

$$\begin{aligned}\{\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T)\}^2 \le \{\mathrm{CC}^{(G)}(\xi,\boldsymbol{\omega},T)\}^2 \le \mathrm{SCIF}_1^{(G)}(\xi,\boldsymbol{\omega},T,T),\\ \mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T^*)\mathrm{CC}^{(G)}(\xi,\boldsymbol{\omega},T) \le \mathrm{SCIF}_2^{(G)}(\xi,\boldsymbol{\omega},T,T^*).\end{aligned} \tag{3.7}$$

For $\boldsymbol{u}\in\mathscr{C}^{(G)}(\xi,\boldsymbol{\omega},T)$, $\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T')$ can be replaced by

$$(\xi+1)\ \mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T')$$

in (3.7), as

$$\|\mathbf{X}\boldsymbol{u}\|_2^2/\max_j(\omega_j^{-1}\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{u}\|_2) \le \sum_j\omega_j\|\boldsymbol{u}_{G_j}\|_2 \le (1+\xi)\sum_{j\in T}\omega_j\|\boldsymbol{u}_{G_j}\|_2.$$

Thus, if a restricted eigenvalue condition as in $\{\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T)\}^2 > \kappa_0$ holds with a fixed κ_0, then all the other quantities in (3.7) and $(\xi+1)\ \mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T)$ are bounded from below by κ_0, $q\in\{1,2\}$. It follows that the cone invertibility factors provide error bounds of sharper form than (3.2), in view of Theorem 5 below and Theorem 3.1 of Lounici et al. (2011).

In the following Theorem 5 we provide the prediction, ℓ_2 and mixed norm consistency results for the non-scaled group Lasso problem defined in (1.3) under the SCIF condition.

Theorem 5. *Let $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ be a solution of (1.3) with data $(\boldsymbol{X},\boldsymbol{y})$ and $\boldsymbol{\beta}^*$ be a vector with supp$(\boldsymbol{\beta}^*)\subseteq G_{S^*}$ for some $S^*\subset T^*=\{1,\cdots,M\}$. Let $\xi>1$ and define*

$$\mathcal{E} = \left\{\max_{1\le j\le M}\frac{\|\boldsymbol{X}_{G_j}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*)\|_2}{\omega_j n} \le \frac{\xi-1}{\xi+1}\right\}. \tag{3.8}$$

Then in the event $\mathcal{E}$, we have

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \frac{\{2\xi/(\xi+1)\}^2 \sum_{j\in S^*}\omega_j^2}{\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*)}, \tag{3.9}$$

and for all $q \ge 1$

$$\left\{\sum_{j=1}^M \omega_j^2 \left(\frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q} \le \frac{\{2\xi/(\xi+1)\}\big(\sum_{j\in S^*}\omega_j^2\big)^{1/q}}{\mathrm{SCIF}_q^{(G)}(\xi, \boldsymbol{\omega}, S^*, T^*)}. \tag{3.10}$$

Moreover, if $\boldsymbol{y} - \boldsymbol{X}\beta^* \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\boldsymbol{I}_n\,)$ and

$$\omega_j \ge A\sigma\|\boldsymbol{X}_{G_j}\|_S\big\{|G_j|^{1/2} + \sqrt{2\log(M/\delta)}\big\}/n \text{ for some } 0 < \delta < 1,$$

and $A \ge (\xi+1)/(\xi-1)$, then

$$\mathbb{P}(\mathcal{E}) > 1 - \delta. \tag{3.11}$$

Theorem 5 asserts that the prediction loss $\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2/n$, the ℓ_2 loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ and the mixed norm loss $\sum_{j=1}^M \omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2$ are all of the order

$$\textstyle\sum_{j\in S^*}\omega_j^2 \asymp (s + g\log M)/n$$

when the SCIF can be treated as constant and $\max_j \|\mathbf{X}_{G_j}/\sqrt{n}\|_S = O_{\mathbb{P}}(1)$. This result illustrates the benefit of the group Lasso as compared to Lasso. The results in Theorem 5 are not entirely new. In fact, for the group Lasso problem (1.3), the same convergence rate can be derived from the ℓ_2 consistency result in Huang and Zhang (2010). While the result of Huang and Zhang (2010) is derived under a sparse eigenvalue condition on the design matrix $\mathbf{X}$, our results are based on the weaker sign-restricted cone invertibility condition and cover the weighted $\ell_{2,q}$ loss for $q > 2$. The proof of Theorem 5 is relegated to the Appendix.

3.2. A scaled group Lasso

In the optimization problem (1.3), scale-invariance considerations have not been taken into account. Usually the individual penalty level ω_j's could be chosen proportional to the scale σ as a remedy. This issue has been discussed and studied, pertaining to the Lasso problem, in the literature. See Huber (2011), Städler, Bühlmann and Geer (2010), Antoniadis (2010), Sun and Zhang (2010), Belloni, Chernozhukov and Wang (2011), Sun and Zhang (2012b), Sun and Zhang (2013) and many more. For the group Lasso problems, this issue has been tackled via the square-root group Lasso formulation in Bunea, Lederer and She (2014). Here we follow the the prescription from Antoniadis (2010) and define an optimization problem,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = \mathop{\arg\min}_{\boldsymbol{\beta},\sigma}\ \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma), \tag{3.12}$$

$$\text{where } \ \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma) = \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n\sigma} + \frac{(1-a)\sigma}{2} + \sum_{j=1}^M \omega_j\|\boldsymbol{\beta}_{G_j}\|_2. \tag{3.13}$$

Following Sun and Zhang (2010) we define an iterative algorithm for the estimation of $\{\boldsymbol{\beta}, \sigma\}$,

$$\begin{array}{rcl} \widehat{\sigma}^{(k+1)} & \leftarrow & \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(k)}\|_2/\sqrt{(1-a)n}, \\ \boldsymbol{\omega}' & \leftarrow & \widehat{\sigma}^{(k+1)}\boldsymbol{\omega}, \\ \widehat{\boldsymbol{\beta}}^{(k+1)} & \leftarrow & \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\boldsymbol{\omega}'}(\boldsymbol{\beta}), \end{array} \tag{3.14}$$

where $\mathcal{L}_{\boldsymbol{\omega}'}(\boldsymbol{\beta})$ was as defined in (1.3). Due to the convexity of the joint loss function $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma)$, the solution of (3.12) and the limit of (3.14) give the same estimator. Moreover, if the minimization of σ is first taken with the unknown $\boldsymbol{\beta}$ in (3.12), the second minimization of $\min_\sigma \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma)$ over $\boldsymbol{\beta}$ becomes the square-root group Lasso problem of Bunea, Lederer and She (2014) when $\omega_j \propto |G_j|^{1/2}$. As the aim of this paper is statistical inference of group effects, the formulation in (3.12) explicitly provides a needed estimate of σ. Moreover, we use a different penalty $\omega_j \propto |G_j|^{1/2} + \sqrt{2\log(M/\delta)}$ to benefit from group sparsity in the estimation of both $\boldsymbol{\beta}$ and σ and in prediction as well.

The constant $a \geq 0$ provides control over the degrees of freedom adjustments. For simplicity, we take $a = 0$ for all subsequent discussions. It is clear that that with $a = 0$ and $\boldsymbol{\omega}' = \widehat{\sigma}\boldsymbol{\omega}$, one has $\widehat{\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \widehat{\sigma}) = \mathcal{L}_{\boldsymbol{\omega}'}(\boldsymbol{\beta}) + \widehat{\sigma}^2/2$. The algorithm in (3.14) suggests a profile optimization approach. The following lemma is similar to Proposition 1 in Sun and Zhang (2012b) and characterizes the solution via partial derivative of the profile objective.

Lemma 2. *Let $\widehat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ denote a solution of the optimization problem in (1.3). Then, $\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})$ is a minimizer of $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta}, \sigma)$ in (3.13) for given σ, and the profile loss function $\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}), \sigma)$ is convex and continuously differentiable in σ with*

$$\frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}), \sigma) = \frac{1}{2} - \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})\|_2^2}{2n\sigma^2}. \tag{3.15}$$

Moreover, the algorithm in (3.14) converges to a minimizer $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ in (3.12) satisfying $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\sigma}\boldsymbol{\omega})$, and the estimator $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}$ are scale equivariant in $\boldsymbol{y}$.

The proof of Lemma 2 is relegated to the Appendix. We now present the consistency theorem which extends Theorem 5 by providing convergence results for the estimate of scale. Define

$$\mu(\boldsymbol{\omega}, \xi) = \frac{2\xi\sum_{j\in S^*}\omega_j^2}{\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*)}, \quad \tau_- = \frac{2\mu(\boldsymbol{\omega}, \xi)(\xi - 1)}{\xi + 1}, \quad \tau_+ = \frac{\tau_-}{2} + \mu(\boldsymbol{\omega}, \xi).$$

Let $m_{d,n}$ be the median of the beta$(d/2, n/2 - d/2)$ distribution and define

$$\omega_{*,j} \geq \sqrt{m_{d_j,n}} + \sqrt{\frac{2\log(M/\delta)}{(n \vee 2) - 3/2}}, \; A_* = \frac{(\xi+1)/(\xi-1)}{\sqrt{\{1 - 2\mu(\boldsymbol{\omega}_*, \xi)(\xi+1)/(\xi-1)\}_+}},$$

where $\boldsymbol{\omega}_*$ is the vector with elements $\omega_{*,j}$ and $d_j = |G_j|$. We will show that $\sqrt{m_{d_j,n}} \leq (d_j/n)^{1/2} + n^{-1/2}$ in the proof of the following theorem.

Theorem 6. *Let $\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\}$ be a solution of the optimization problem (3.13) with data $(\boldsymbol{X}, \boldsymbol{y})$ and $\boldsymbol{\beta}^*$ be a vector with supp$(\boldsymbol{\beta}^*) \subset G_{S^*}$ for some $S^* \subset T^* = \{1, \cdots, M\}$. Let $\xi > 1$.*
(i) Suppose $\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*) > 0$ in (3.6) and $\tau_+ < 1$. Define the following event

$$\mathcal{E} = \left\{ \max_{1 \le j \le M} \frac{\|\boldsymbol{X}_{G_j}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_2}{\omega_j n \sigma^* / \sqrt{1+\tau_-}} < \frac{\xi - 1}{\xi + 1} \right\}, \tag{3.16}$$

where $\sigma^* = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2 / \sqrt{n}$ is the oracle noise level. Then in the event $\mathcal{E}$, we have

$$\frac{\sigma^*}{\sqrt{1+\tau_-}} \le \widehat{\sigma} \le \frac{\sigma^*}{\sqrt{1-\tau_+}}, \tag{3.17}$$

$$\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2/n \le \frac{(\sigma^*)^2 \{2\xi/(\xi+1)\}^2 \sum_{j \in S^*} \omega_j^2}{(1-\tau_+)\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*)}, \tag{3.18}$$

and for all $q \ge 1$

$$\left\{ \sum_{j=1}^M \omega_j^2 \left(\frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j} \right)^q \right\}^{1/q} \le \frac{\sigma^* \{2\xi/(\xi+1)\} \left(\sum_{j \in S^*} \omega_j^2 \right)^{1/q}}{\sqrt{1-\tau_+}\mathrm{SCIF}_q^{(G)}(\xi, \boldsymbol{\omega}, S^*, T^*)}. \tag{3.19}$$

(ii) Suppose the regression model in (1.1) holds with Gaussian error, $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Suppose $\omega_j \ge A\|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S \omega_{*,j}$ with $A \ge A_*$. Then,

$$\mathbb{P}(\mathcal{E}) \ge 1 - \delta \tag{3.20}$$

with the event $\mathcal{E}$ in (3.16). Moreover, if $\sqrt{n}\mu(\boldsymbol{\omega}, \xi) \to 0$, then

$$\sqrt{n}\,(\widehat{\sigma}/\sigma - 1) \xrightarrow{\mathrm{D}} \mathsf{N}(0, 1/2). \tag{3.21}$$

Theorem 6, whose proof is again relegated to the Appendix, provides explicit rates and constants for mixed ℓ_q norm estimation of $\boldsymbol{\beta}^*$ and estimation of scale parameter σ. When $\omega_j \asymp \omega_{*,j}$ and $\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*) \asymp 1$, we have

$$\textstyle\sum_{j \in T} \omega_j^2 \asymp \mu(\boldsymbol{\omega}, \xi) \asymp \{s + g\log(M/\delta)\}/n.$$

It also establishes the veracity of the working assumption in (2.31). The following Corollary 1 provides a more succinct summary to make clear the connection of Theorem 6 to (2.31).

Corollary 1 (Verification of working assumption for deterministic designs)**.** *Let $\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\}$ be as in (3.13) with a penalty level satisfying $\omega_j/A^* \le \|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S \omega_{*,j} \le \omega_j/A_*$. Suppose the design matrix $\boldsymbol{X}$ satisfy the condition $\|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S^2 \le c^*$ and that the sign-restricted cone invertibility condition holds in the sense of*

$\mathrm{SCIF}_q^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*) > c_*$ *for some fixed* $c_* > 0$. *Suppose* $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ *and* $supp(\boldsymbol{\beta}^*) \subseteq G_{S^*}$ *with* $|G_{S^*}| + |S^*|\log(M/\delta) \le a_0 n$. *Then, for certain constants* $\{a_*, C\}$ *depending on* $\{c_*, c^*, \xi, A^*\}$ *only,*

$$\max\left\{\left|1 - \frac{\widehat{\sigma}}{\sigma^*}\right|, \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2}{n\sigma^2}, \sum_{j=1}^M \frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\sigma/\omega_j}, \sum_{j=1}^M \frac{\|\boldsymbol{X}_{G_j}(\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j})\|_2}{n^{1/2}\sigma/\omega_j}\right\} \le C\left\{|G_{S^*}| + |S^*|\log(M/\delta)\right\}/n \tag{3.22}$$

with probability at least $1 - \delta$ *whenever* $a_0 \le a_*$.

Corollary 1 touches upon the mixed prediction loss $\sum_{j=1}^M \omega_j \|\mathbf{X}_{G_j}\widehat{\boldsymbol{\beta}}_{G_j} - \mathbf{X}_{G_j}\boldsymbol{\beta}^*_{G_j}\|_2$ the first time in this section. The reason for this omission is two fold. Firstly,

$$\left\{\sum_{j=1}^M \omega_j^2 \left(\frac{\|\mathbf{X}_{G_j}(\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j})\|_2}{n^{1/2}\omega_j}\right)^q\right\}^{1/q} \le \max_{j \le M}\left\|\frac{\mathbf{X}_{G_j}}{\sqrt{n}}\right\|_S \left\{\sum_{j=1}^M \omega_j^2 \left(\frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q}$$

so that (3.10) and (3.19) automatically generate the corresponding bounds for the mixed prediction error under the respective conditions. Secondly, upper bounds for the mixed prediction loss can be obtained by reparametrization within the given group structure as in the following corollary.

Corollary 2. *Let* $\boldsymbol{X}_{G_j} = \boldsymbol{U}_{G_j}\boldsymbol{\Lambda}_{G_j}\boldsymbol{V}^T_{G_j}$ *be the SVD of* $\boldsymbol{X}_{G_j}$ *with* $\boldsymbol{\Lambda}_{G_j} \in \mathbb{R}^{|G_j| \times |G_j|}$. *Define* $\boldsymbol{b}$ *by* $\boldsymbol{b}_{G_j} = \boldsymbol{\Lambda}_{G_j}\boldsymbol{V}^T_{G_j}\boldsymbol{\beta}_{G_j}$ *and* $\boldsymbol{U}$ *by* $\boldsymbol{U}\boldsymbol{b} = \sum_{j=1}^M \boldsymbol{U}_{G_j}\boldsymbol{b}_{G_j}$. *Then,*

$$\begin{aligned}\left\{\sum_{j=1}^M \omega_j^2 \left(\frac{\|\boldsymbol{X}_{G_j}\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{X}_{G_j}\boldsymbol{\beta}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q} &= \left\{\sum_{j=1}^M \omega_j^2 \left(\frac{\|\widehat{\boldsymbol{b}}_{G_j} - \boldsymbol{b}^*_{G_j}\|_2}{\omega_j}\right)^q\right\}^{1/q} \\ &\le \frac{2\sigma^*\xi\left(\sum_{j \in S^*}\omega_j^2\right)^{1/q}}{\sqrt{1-\tau_+}\mathrm{SCIF}_q^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*)}\end{aligned}$$

for all $q \ge 1$ *when the conditions for (3.19), including the definition of the estimator and the SCIF, hold with* $\boldsymbol{X}$, $\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^*$ *replaced by* $\boldsymbol{U}$, $\boldsymbol{b}$ *and* $\boldsymbol{b}^*$ *respectively.*

Remark 3. Corollary 1 can be viewed as a scaled version of the main results of Huang and Zhang (2010) although here the regularity condition of the design is of a weaker form and smaller penalty levels are allowed.

3.3. Random designs

In this subsection, we verify the working assumption for sub-Gaussian designs by checking the groupwise cone invertibility condition. Our analysis also provides lower bounds for the groupwise restricted eigenvalue and compatibility constant. We first state in the following theorem the main result for random designs.

Theorem 7 (Verification of working assumption for random designs)**.** *Let* $0 < c_* \le c^*$ *and* $0 < \delta < 1 < A_* < A^*$ *be fixed constants and* $\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\}$ *be a solution of (3.13) with*

$$\omega_j/A^* \le \|\boldsymbol{X}_{G_j}\|_S\big\{\sqrt{d_j} + \sqrt{2\log(M/\delta)}\big\}/n \le \omega_j/A_*.$$

Let $\sigma^* = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2/\sqrt{n}$*. Suppose* $\boldsymbol{X}$ *satisfies the sub-Gaussian condition (2.28) with* $c_* \le$*eigenvalues*$(\boldsymbol{\Sigma}) \le c^*$, $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\,)$*, and supp*$(\boldsymbol{\beta}^*) \subseteq G_{S^*}$ *with*

$$\max_{1 \le j \le M} \Big(|G_j| + \log(M/\delta)\Big) I_{\{|S^*|>0\}} + |G_{S^*}| + |S^*|\log(M/\delta) \le a_0 n. \tag{3.23}$$

Then, there exist constants a_* *and* C *depending on* $\{c_*, c^*, A_*, A^*\}$ *only such that*

$$\begin{aligned} &\max\Bigg\{\Big|1 - \frac{\widehat{\sigma}}{\sigma^*}\Big|, \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2}{n\sigma^2}, \sum_{j=1}^M \frac{\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j}\|_2}{\sigma/\omega_j}, \\ &\qquad\qquad\qquad\qquad \sum_{j=1}^M \frac{\|\boldsymbol{X}_{G_j}(\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}^*_{G_j})\|_2}{n^{1/2}\sigma/\omega_j}\Bigg\} \\ &\le C\left\{|G_{S^*}| + |S^*|\log(M/\delta)\right\}/n \end{aligned} \tag{3.24}$$

with probability at least $1 - \delta$ *whenever* $a_0 \le a_*$*.*

Theorem 7 justifies the working assumption for sub-Gaussian designs. It demonstrates the benefit of the strong group sparsity as the sample size condition (3.23) is typically weaker than the usual $\|\boldsymbol{\beta}^*\|_0\{1 + \log(p/\delta)\} \le a_0 n$ for the Lasso when $\text{supp}(\boldsymbol{\beta}) = G_{S^*}$. We omit its proof as it is a direct consequence of Theorem 6 and Proposition 2 below. We preface the presentation of Proposition 2 by first defining the following quantities.

Let $q > 1$ and $\boldsymbol{f} = (f_1, \ldots, f_M)^T$ with $f_j > 0$. Define

$$\begin{aligned} \rho_q(s) = \inf_{\boldsymbol{u}} \sup_{\boldsymbol{v}} \Bigg\{ &\frac{\boldsymbol{v}^T(\mathbf{X}^T\mathbf{X}/n)\boldsymbol{u}}{\|\boldsymbol{v}\|_{(q/(q-1))}\|\boldsymbol{u}\|_{(q)}} : \text{supp}(\boldsymbol{u}) = \text{supp}(\boldsymbol{v}) = G_B, \\ &\qquad\qquad \min_{|B\setminus S| \le 1} \|\boldsymbol{f}_S\|_2^2 < s \Bigg\} \end{aligned} \tag{3.25}$$

with weighted $\ell_{2,q}$ norm $\|\boldsymbol{v}\|_{(q)} = \Big(\sum_{j=1}^M f_j^2 \big(\|\boldsymbol{v}_{G_j}\|_2/f_j\big)^q\Big)^{1/q}$, and

$$\theta_q(s,t) = \sup\left\{ \frac{\boldsymbol{v}^T(\mathbf{X}^T\mathbf{X}/n)\boldsymbol{u}}{\|\boldsymbol{v}\|_{(q/(q-1))}\|\boldsymbol{u}\|_{(q)}} : \mathrm{supp}(\boldsymbol{u}) = G_{B_1}, \mathrm{supp}(\boldsymbol{v}) = G_{B_2}, \right.$$
$$\left. |B_k \setminus S_k| \le 1, \|\boldsymbol{f}_{S_1}\|_2^2 < s, \|\boldsymbol{f}_{S_2}\|_2^2 < t, B_1 \cap B_2 = \emptyset \right\}. \tag{3.26}$$

Under the norm $\|\cdot\|_{(q)}$, $1/\rho_q(s)$ is the maximum operator norm of $n(\mathbf{X}_{G_B}^T\mathbf{X}_{G_B})^{-1}$ in $\mathbb{R}^{|G_B|}$, and $\theta(s,t)$ is the maximum operator norm of $\mathbf{X}_{G_{B_2}}^T\mathbf{X}_{G_{B_1}}/n$. In particular, $\rho_2(s)$ is the smallest eigenvalue of $\mathbf{X}_{G_B}^T\mathbf{X}_{G_B}/n$ under the given constraints on the support set G_B. Let $a_q = (1-1/q)/q^{1/(q-1)}$. For $\xi > 0$, $T \subset \{1,\dots,M\}$, $t_0 = \sum_{j\in T} f_j^2$, $x_0 \ge 1$, $1 \le y_0 \le x_0/a_q$ and $m \in \{1,2\}$, define quantities $C_q(\xi,x_0,y_0) = \xi + \big(1 + a_q y_0 - x_0\big)_+ x_0^{-1/q}$ and

$$\kappa_{q,m}(\xi,t_0,x_0,y_0) = \rho_q(x_0t_0) - m\theta_q(x_0t_0,y_0t_0)y_0^{1/q-1}C_q(\xi,x_0,y_0). \tag{3.27}$$

Proposition 2. *(i) Suppose $\omega_j = C_n f_j$ for some constant C_n not depending on j. Then,*

$$\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T') \ge \kappa_{2,2}^{1/2}(\xi,t_0,x_0,y_0)/\{1+\delta'\big(1+\xi\big)/2\}, \tag{3.28}$$

$$\mathrm{CC}^{(G)}(\xi,\boldsymbol{\omega},T) \ge \kappa_{2,2}^{1/2}(\xi,t_0,x_0,y_0) \tag{3.29}$$

$$\mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T') \ge \frac{\kappa_{q,1}(\xi,t_0,x_0,y_0)}{(x_0 + \max_j f_j^2/t_0)^{1/q}\{1+\delta'\big(1+\xi\big)a_q^{1-1/q}\}}, \tag{3.30}$$

with $\delta' = 0$ for $T' = T$ and $\delta' = 1$ for $T' = T^*$, and for $1 \le q \le 2$

$$\min\left(\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T'), \frac{\mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T')}{(1+\xi)^{-1}}\right) \ge \frac{\kappa_{2,1}(\xi,t_0,x_0,y_0)}{1+\delta'\big(1+\xi\big)a_q^{1-1/q}}. \tag{3.31}$$

(ii) Suppose $\boldsymbol{X}$ satisfies the sub-Gaussian condition (2.28) with $c_* \le$eigen$(\boldsymbol{\Sigma}) \le c^*$ and

$$\omega_j/A^* \le C_n\|\boldsymbol{X}_{G_j}/\sqrt{n}\|_S\big\{\sqrt{|G_j|} + \sqrt{2\log(M/\delta)}\big\} \le \omega_j/A_*$$

where $\{c_*, c^*, A_*, A^*\}$ are positive constants. Let $\delta' = 0$ for $T' = T$ and $\delta' = 1$ for $T' = T^*$. For any $\epsilon_0 \in (0,1)$, there exists a_0 depending on $\{\epsilon_0, c_*, c^*, A_*, A^*\}$ only such that

$$\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T') \ge (1-\epsilon_0)\lambda_{\min}(\boldsymbol{\Sigma})/\big\{1+\delta'\big(1+\xi\big)a_q^{1-1/q}\big\}, \quad 1 \le q \le 2.$$

with at least probability $1-\delta$ whenever (3.23) holds. Moreover, the inequality also holds with $\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T')$ replaced by $\big\{\mathrm{RE}^{(G)}(\xi,\boldsymbol{\omega},T,T')\big\}^2$ for $q=2$, by $\big\{\mathrm{CC}^{(G)}(\xi,\boldsymbol{\omega},T)\big\}^2$ for $q=1$ and $T'=T$, or by $(1+\xi)\mathrm{CIF}_q^{(G)}(\xi,\boldsymbol{\omega},T,T')$.

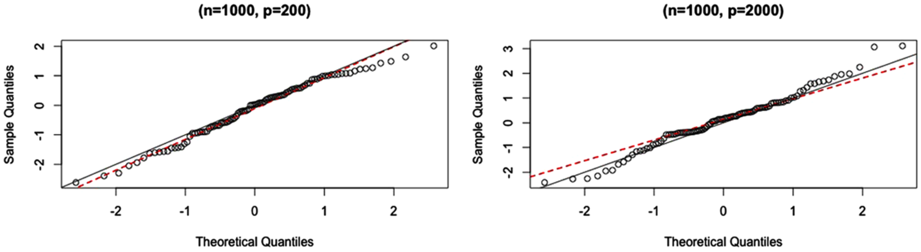

FIG 1. *Normal QQ plot for the test statistic for $\widehat{\sigma}$ in (3.21) in Theorem 6 with $n = 1000, p = \{200, 2000\}, g = 2, s = 8$. The results are produced with 100 replications of the scaled group Lasso. The red dotted line is fitted through 1^{st} and 3^{rd} sample quantiles.*

4. Simulation results

In this section we provide a few simulation results in support of our theory developed in Sections 2 and 3. As a prelude, we first show the performance of the scaled group Lasso procedure in a simulation experiment.

4.1. Normality of estimate of the scale parameter

We consider two simulation designs with $(n = 1000, p = 200)$ and $(n = 1000, p = 2000)$ design matrices with the elements of the design matrix generated independently from $\mathsf{N}(0, 1)$. We assume that the true parameter $\boldsymbol{\beta}^*$ has an inherent grouping with total set of p parameters divided into groups of size $d_j = 4$. In the design $(n = 1000, p = 200)$ we have total number of groups $M = 50$ and in $(n = 1000, p = 2000)$, $M = 500$. For both scenarios, the true parameter $\boldsymbol{\beta}^*$ is assumed to be $(g = 2, s = 8)$ strong group sparse with its non-zero coefficients in $\{-1, 1\}$. Both simulation designs have a $\mathsf{N}(0, \sigma^2)$ error added to the true regression model $\mathbf{X}\boldsymbol{\beta}^*$ with $\sigma = 1$. We also assume that the design matrix is groupwise orthogonalized in the sense of $\mathbf{X}_{G_j}^T \mathbf{X}_{G_j}/n = \mathbf{I}_{G_j \times G_j}$, $j = 1, \dots, M$.

In estimation of σ we employ the scaled group Lasso procedure as shown in (3.14). The groupwise penalty factors ω_j's are chosen to equal to $\lambda(\sqrt{d_j/n} + \sqrt{(2/n)\log(M)})$ for some fixed $\lambda > 0$. The implementation of group Lasso procedure is via the R package `grpreg`.

In the design setup with $(n = 1000, p = 200)$, the estimate of $\widehat{\sigma}$ averaged over a 100 replications is 0.997 with a standard deviation of 0.02. In the design setup with $(n = 1000, p = 2000)$, the estimate of $\widehat{\sigma}$ averaged over a 100 replications is 1.0002 with a standard deviation of 0.02. Additionally Figure 1 shows the Gaussian QQ plots of the test statistic $\sqrt{2n}\,(\widehat{\sigma}/\sigma - 1)$.

4.2. Asymptotic distribution of regression parameters

We also seek the empirical validation of the asymptotic convergence of the group $\boldsymbol{\beta}_{G_j}$ as described in our theoretical results. For bias correction we take the

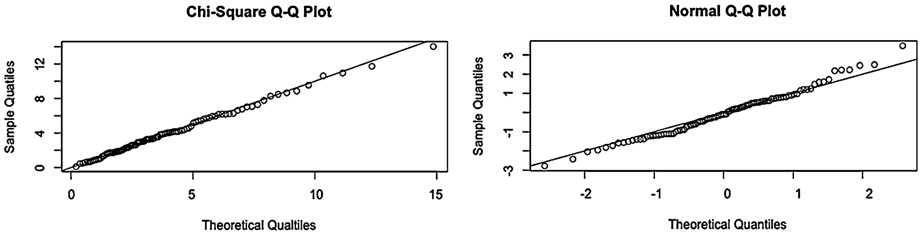

Fig 2. *The left panel considers test for a* ***Small group****. It shows chi-squared QQ plot for the test statistic T_G with $n = 1000, p = 200, g = 10, s = 40$. The theoretical quantiles were drawn from χ^2_4 random variable. The group being tested has size 4. The right panel considers test for a* ***Large group****. It shows normal QQ plot for the test statistic $(T^2_G - |G|)/\sqrt{2|G|}$ with $n = 1000, p = 200, g = 2, s = 40$. Here the group size of the test group is 20.*

penalty function in (2.42) to be the Frobenius norm and apply group Lasso based optimization. We also consider a new simulation design which is similar to the earlier design with $(n = 1000, p = 200)$ and $\sigma = 1$. We will consider two different schemes for empirical analysis for asymptotic convergence.

Small group sizes

The true parameter $\boldsymbol{\beta}^*$ is simulated to be $(s = 40, g = 10)$ strong group sparse with its nonzero values in the interval [2, 3]. More specifically, $\boldsymbol{\beta}^*$ is grouped into groups of sizes $d_j = 4$ for all j. We construct the test statistic of $\boldsymbol{\mu}_{G_j}$ as in (2.21) for one of the nonzero groups. The **left panel** of Figure 2 provides χ^2_4 based QQ plot for the sample quantiles of our test statistic.

Large group sizes

The true parameter $\boldsymbol{\beta}^*$ is simulated to be $(s = 40, g = 2)$ strong group sparse with its nonzero values between [2, 3]. More specifically, $\boldsymbol{\beta}^*$ is grouped into 10 groups each of sizes $d_j = 20$ for all j. We let the sparsity of the true parameter $\boldsymbol{\beta}^*$ to be $s = 40$ contained within 2 separate groups. Again, we construct the test statistic of $\boldsymbol{\mu}_{G_j}$ as in (2.21) for one of the nonzero groups. The **right panel** of Figure 2 shows the QQ plot for this group's size- normalized test statistic as defined in (2.22). As the figure suggests, for large group sizes asymptotic normality of the group test statistic is empirically supported.

4.3. Comparison with other methods

In this subsection we compare the performance of our group Lasso methods with other recent methods developed for inference in high dimensional models. In particular we consider three different classes of methods.

Table 1

Comparison of true positive and false positive rates for three different choices of block correlation ρ and three choices of signal parameter τ. The scale parameter $\sigma = 1$ in all cases. The results are based on 100 replications for testing the nonzero group (for TP) and first zero group (FP). Performance of all the tests are good for the strong signal (τ= 1). For the weak signal $\tau = 0.1$, group Lasso clearly out-performs other methods.

Design	**Proposed Method**				**Projection Based**				**Multi sample-split**				**Group Bound**	
	Chi-squared		**Normal**		**Lasso**		**Ridge**		**Lasso**		**Group Lasso**			
(g, s), (ρ, τ)	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**
(1, 5), (0, 0.1)	0.04	0.11	0.04	0.11	0	0.02	0	0	0	0	0	0	0	0
(1, 5), (0, 0.5)	0	1	0	1	0	1	0.01	0.2	0	0.72	0	0.23	0	0
(1, 5), (0, 1)	0	1	0	1	0	1	0	1	0	1	0	1	0	0
(1, 5), (0.5, 0.1)	0.03	0.3	0.03	0.3	0	0.06	0	0	0	0	0	0	0	0
(1, 5), (0.5, 0.5)	0	1	0	1	0	1	0	0.71	0	0.99	0	0.47	0	0.02
(1, 5), (0.5, 1)	0	1	0	1	0	1	0	1	0	1	0	1	0	0.97
(1, 5), (0.9, 0.1)	0.02	0.45	0.02	0.45	0	0.02	0.2	0.02	0	0.32	0	0	0	0
(1, 5), (0.9, 0.5)	0	1	0	1	0	1	0	0.07	0	0.22	0	0.01	0	0.12
(1, 5), (0.9, 1)	0	1	0	1	0	0.98	0	0.81	0	0.86	0	0.32	0	1
(1, 20), (0.9, 0.1)	0	1	0	1	0	0.38	0	0.01	0	0.05	0	0.00	0	0.04

Projection based: For the projection based methods, we consider two cases. 1) The Ridge estimation based testing with correction for projection bias that was developed in Bühlmann (2013). 2) The Lasso relaxed projection followed by bias correction idea developed in Zhang and Zhang (2014) which is similar to the de-sparsified Lasso in van de Geer et al. (2014). These methods are adapted for testing of groups of variables adjustment of individual p-values; see Dezeure et al. (2014).

Sample split based: The idea of single sample splitting was developed in Wasserman and Roeder (2009) which involves splitting the sample into two parts. The first part is used to select variables and the second to construct p-values for the selected variables in the first model. The final step is to adjust the p-values for control of the familywise error rate (FWER). Due to the variability of the p-values for different splittings, Meinshausen, Meier and Bühlmann (2009) proposed multi sample-splitting idea which involves running the single sample splitting B times and aggregating the B adjusted p-values. We employ the multi sample-splitting with two different variable selection procedures: Lasso and group Lasso. For Lasso, the groupwise p-value is obtained by Bonferroni adjustments.

Group bound: The final procedure we consider is the group bound method developed in Meinshausen (2014). One advantage of this method is that it doesn't require any assumptions on the design matrix.

Implementation of all the above methods are available in the **R** package `hdi`; see also Dezeure et al. (2014).

Simulation Design: We consider a very simple simulation design where the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ is assumed to have iid rows with each row following $\mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is assumed to be a correlation matrix having a block diagonal structure with block size $k = 5$. We take $n = 100$ and $p = 200$ so that $\mathbf{\Sigma}$ has M =40 blocks. Within each block, the correlation is assumed to be ρ. For our simulations, we consider three possible choices of ρ namely $\{0, 0.5, 0.9\}$.

The true parameter $\boldsymbol{\beta}$ is assumed to have the group structure as defined by the block structure of $\mathbf{X}$. Moreover we assume only the first group has nonzero signals with all of them having the same value $\tau > 0$. Thus $\boldsymbol{\beta}^*$ is of the form,

$$\boldsymbol{\beta}^* = (\underbrace{\tau, \tau, \tau, \tau, \tau}_{\text{group 1}}, \underbrace{0, 0, 0, 0, 0}_{\text{group 2}}, \cdots, \underbrace{0, 0, 0, 0, 0}_{\text{group 40}})$$

Thus in all these cases, the true signal $\boldsymbol{\beta}^*$ is $(g = 1, s = 5)$ strong group sparse. We consider three choices of the signal parameter τ: $\{0.1, 0.5, 1\}$.

We also consider an additional scenario, where we take $k = 20$ so that number of groups $M = 10$ (The last line of Table 1). For this case we only compare the performance for signal strength $\tau = 0.1$ which highlights the performance of group Lasso.

The responses are simulated by $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathsf{N}(0, \sigma)$. We take the true scale parameter $\sigma = 1$ in all simulation designs and estimate σ via scaled group Lasso. For application of the group Lasso based testing, we take the group weights equal to $\omega_j = 5(\sqrt{d_j/n} + \sqrt{(2/n)\log(M)})$ where $M = 40$ and $d_j = 5$ for group sizes=5 and $d_j = 20$ for group sizes=20.

In Table 1, we provide a comparison of the true positive (TP) and false positive (FP) rates for 100 replications. It is clear from the table that group Lasso performs comparably or better than all the other methods. The false positive rates of all the methods are either 0 or close to zero for most of the designs. The true positive (TP) rate (power) of group Lasso method clearly dominates those of the other methods especially when the signal is not strong: $\tau = 0.1$. One rationale for this would be the accumulation of small signals in the ℓ_2 norm for the group that is used for the group Lasso. For group bound method, clearly the performance becomes comparable to group Lasso as the blockwise correlation ρ is increased. This phenomenon is also observed for group Lasso procedure to a certain extent.

5. Summary and discussion

We have considered statistical inference of variable groups in a high-dimensional linear regression setup. In particular we show the benefit of grouping in constructing chi-squared-type procedures for group inference. We construct such procedures via bias correction and group Lasso based relaxed projection. We show the validity of such approximate chi-squared-type inference under sample size conditions that could be potentially much weaker than the requirements for Lasso based procedures. This particular scaling also offers us valid statistical inference for a group of possibly unbounded number of variables.

A key step of our methodology concerns the nonconvex optimization scheme (2.33) over the set of orthogonal projection matrices. To the best of our knowledge, solution of an optimization problem as in (2.33) is not yet well studied, either algorithmically or analytically. However, we have proposed a convexation of (2.33) via a multivariate group Lasso with a weighted nuclear or Frobenius norm penalty, which provides feasible solutions for the optimization problem. As discussed in Remark 1, our theoretical results only requires feasibility solutions of the optimization scheme. As the multivariate group Lasso with Frobenius norm penalty can be carried out using the group Lasso program, an interesting direction of research would be to develop efficient algorithm for the group nuclear norm penalty.

Since our results can be directly applied to statistical inference for groups of variables with possibly unbounded sizes, application of our procedures for sparse nonparametric additive models (Ravikumar et al., 2009) would be another future direction of research.

Appendix

This appendix provides proof of

Proof of Proposition 1. (i) Since both $\mathbf{Z}_G$ and $\mathbf{X}_G$ are $n \times |G|$ matrices,

$$|G| = \mathrm{rank}(\mathbf{Z}_G^T \mathbf{X}_G) \le \mathrm{rank}(\mathbf{X}_G) \wedge \mathrm{rank}(\mathbf{Z}_G) \le |G| \wedge n,$$

so that $\mathrm{rank}(\mathbf{P}_G) = \mathrm{rank}(\mathbf{P}_G\mathbf{X}_G) = |G|$ and $\mathbf{P}_G = \mathbf{P}_{G,0}$. It follows that $\mathbf{P}_G\mathbf{X}_G(\mathbf{P}_G\mathbf{X}_G)^\dagger\mathbf{P}_G = \mathbf{P}_G$. As $\mathbf{Z}_G^T\mathbf{X}_G$ is a $|G| \times |G|$ invertible matrix,

$$\mathbf{P}_G\mathbf{X}_G(\mathbf{Z}_G^T\mathbf{X}_G)^{-1}\mathbf{Z}_G^T = \mathbf{P}_G.$$

Since $\mathrm{rank}(\mathbf{P}_G\mathbf{X}_G) = |G|$, we are allowed to cancel $\mathbf{P}_G\mathbf{X}_G$ to obtain

$$(\mathbf{P}_G\mathbf{X}_G)^\dagger\mathbf{P}_G = (\mathbf{Z}_G^T\mathbf{X}_G)^\dagger\mathbf{Z}_G^T.$$

This proves the first equality in (2.15). The second equality in (2.15) then follows from

$$(\mathbf{P}_G\mathbf{X}_G)^\dagger\mathbf{P}_G\big(\mathbf{X}_G\boldsymbol{\beta}_G^* - \mathbf{X}_G\widehat{\boldsymbol{\beta}}_G^{(init)}\big) = \boldsymbol{\beta}_G^* - \widehat{\boldsymbol{\beta}}_G^{(init)},$$

(2.2) and its estimated version, and the definition of the remainder term.

(ii) Let $\mathbf{Z}_1 = \mathbf{P}_G\mathbf{Z}_G$. As $\mathbf{P}_G = \mathbf{P}_G\mathbf{P}_{G,0}$ is the orthogonal projection to $\mathcal{R}(\mathbf{Z}_1)$, $\mathbf{Z}_G^T\mathbf{X}_G = \mathbf{Z}_G^T\mathbf{P}_{G,0}\mathbf{Q}_G\mathbf{X}_G = \mathbf{Z}_1^T\mathbf{P}_G\mathbf{Q}_G\mathbf{X}_G$ and $\mathrm{rank}(\mathbf{X}_G) = \mathrm{rank}(\mathbf{P}_G\mathbf{Q}_G) = \mathrm{rank}(\mathbf{Z}_1^T\mathbf{X}_G)$, so that

$$(\mathbf{Z}_G^T\mathbf{X}_G)^\dagger = (\mathbf{Z}_1^T\mathbf{P}_G\mathbf{Q}_G\mathbf{X}_G)^\dagger = \mathbf{X}_G^\dagger(\mathbf{P}_G\mathbf{Q}_G)^\dagger(\mathbf{Z}_1^T)^\dagger.$$

Consequently, as $\mathbf{Q}_G(\mathbf{P}_G\mathbf{Q}_G)^\dagger = (\mathbf{P}_G\mathbf{Q}_G)^\dagger = (\mathbf{P}_G\mathbf{Q}_G)^\dagger\mathbf{P}_G$ and $(\mathbf{Z}_1^T)^\dagger\mathbf{Z}_G^T = \mathbf{P}_G$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_G - \widehat{\boldsymbol{\mu}}_G^{(init)} &= \mathbf{X}_G(\mathbf{Z}_G^T\mathbf{X}_G)^\dagger\mathbf{Z}_G^T\Big(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}\Big) \\
&= \mathbf{X}_G\mathbf{X}_G^\dagger(\mathbf{P}_G\mathbf{Q}_G)^\dagger(\mathbf{Z}_1^T)^\dagger\mathbf{Z}_G^T\Big(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}\Big) \\
&= (\mathbf{P}_G\mathbf{Q}_G)^\dagger\mathbf{P}_G\Big(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(init)}\Big).
\end{aligned}$$

This gives (2.16). As $\mathbf{Q}_G(\mathbf{Z}_G^T\mathbf{Q}_G)^\dagger\mathbf{Z}_G^T = (\mathbf{P}_G\mathbf{Q}_G)^\dagger\mathbf{P}_G^T$ by the same proof, (2.13) also holds. Finally, (2.18) follows from (2.14) and (2.2). □

Proof of Lemma 1. Let $\boldsymbol{u}_j, 1 \le j \le r_k$, be the eigenvectors of $\mathbf{B}_k^T\boldsymbol{\Sigma}\mathbf{B}_k$ corresponding to positive eigenvalues and $\mathbf{U}_k = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r_k})$. Let

$$\mathbf{Z}_k = \mathbf{X}\mathbf{B}_k((\mathbf{B}_k^T\boldsymbol{\Sigma}\mathbf{B}_k)^\dagger)^{1/2}\mathbf{U}_k \in \mathbb{R}^{n\times r_k}.$$

We have $\mathbb{E}\mathbf{Z}_k = \mathbf{0}$, $\mathbb{E}(\mathbf{Z}_k^T\mathbf{Z}_k/n) = \mathbf{I}_{r_k\times r_k}$, $\mathbb{E}(\mathbf{Z}_1^T\mathbf{Z}_2/n) = \mathbf{U}_1^T\boldsymbol{\Omega}_{1,2}\mathbf{U}_2$, and

$$\sup_{\|\boldsymbol{b}\|_2\le 1} \mathbb{E}\exp\left(\frac{(\boldsymbol{e}_i^T\mathbf{Z}_k\boldsymbol{b})^2}{v_0} + \frac{1}{v_0}\right) \le 2,\ k = 1, 2.$$

Moreover, $\mathbf{P}_k = \mathbf{Z}_k(\mathbf{Z}_k^T\mathbf{Z}_k)^\dagger\mathbf{Z}_k^T$ and $\|\mathbf{U}_1^T\boldsymbol{\Omega}_{1,2}\mathbf{U}_2\|_S = \|\boldsymbol{\Omega}_{1,2}\|_S \le 1$.

For $1 \le j \le k \le 2$ and any vectors $\boldsymbol{v}_k \in \mathbb{R}^{r_k}$ with $\|\boldsymbol{v}_k\|_2 = 1$,

$$\boldsymbol{v}_j^T\Big(\mathbf{Z}_j^T\mathbf{Z}_k/n - \mathbb{E}\mathbf{Z}_j^T\mathbf{Z}_k/n\Big)\boldsymbol{v}_k = \frac{1}{n}\sum_{i=1}^n\Big\{(\boldsymbol{e}_i^T\mathbf{Z}_j\boldsymbol{v}_j)(\boldsymbol{e}_i^T\mathbf{Z}_k\boldsymbol{v}_k) - \boldsymbol{v}_j^T\mathbb{E}(\mathbf{Z}_j^T\mathbf{Z}_k/n)\boldsymbol{v}_k\Big\}$$

is an average of iid variables with

$$\begin{aligned} &\mathbb{E}\exp\left(\frac{(\boldsymbol{e}_i^T\mathbf{Z}_j\boldsymbol{v}_j)(\boldsymbol{e}_i^T\mathbf{Z}_k\boldsymbol{v}_k) - \boldsymbol{v}_j^T\mathbb{E}(\mathbf{Z}_j^T\mathbf{Z}_k/n)\boldsymbol{v}_k}{v_0}\right) \\ \le\ & \left\{\prod_{k=1}^2\sqrt{\mathbb{E}\exp\left((\boldsymbol{e}_i^T\mathbf{Z}_k\boldsymbol{v}_k)^2/v_0\right)}\right\}e^{1/v_0} \\ \le\ & 2. \end{aligned}$$

Since the size of an ϵ-net of the unit ball in $\mathbb{R}^{r_k}$ is bounded by $(1 + 2/\epsilon)^{r_k}$, the Bernstein inequality implies that for $r^* = r_1 + r_2$ and a certain numerical constant C_0,

$$\mathbb{P}\Big\{\|\mathbf{Z}_j^T\mathbf{Z}_k/n - \mathbb{E}(\mathbf{Z}_j^T\mathbf{Z}_k/n)\|_S > C_0v_0\max\left(\sqrt{t/n + r^*/n}, t/n + r^*/n\right)\Big\} \le e^{-t}/3.$$

This yields (2.39) as $\|\mathbf{U}_1^T\boldsymbol{\Delta}\mathbf{U}_2\|_S = \|\boldsymbol{\Delta}\|_S$ for all $\boldsymbol{\Delta}$ of proper dimension.

Suppose $\text{rank}(\mathbf{P}_k) = r_k$. Let $r_0 = \text{rank}(\mathbf{P}_1\mathbf{P}_2)$ and $1 \ge \widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_{r_0} > 0$ be the (nonzero) singular values of $\mathbf{P}_1\mathbf{P}_2$. We have $\|\mathbf{P}_1\mathbf{P}_2\|_S = \widehat{\lambda}_1$ and $\|\mathbf{P}_1\mathbf{P}_2^\perp\|_S = \|\mathbf{P}_1 - \mathbf{P}_2\|_S = \sqrt{1 - \widehat{\lambda}_{\min}^2}$ with $\widehat{\lambda}_{\min} = \widehat{\lambda}_{r_0}I\{r_0 = r_1 = r_2\}$. By definition,

$$\mathbf{P}_1\mathbf{P}_2 = \mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T\mathbf{Z}_2(\mathbf{Z}_2^T\mathbf{Z}_2)^{-1}\mathbf{Z}_2^T.$$

Since $(\mathbf{Z}_k^T\mathbf{Z}_k)^{-1/2}\mathbf{Z}_k^T$ are unitary maps from the range of $\mathbf{P}_k$ to $\mathbb{R}^{r_k}$, the singular values of $\mathbf{P}_1\mathbf{P}_2$ is the same as those of

$$(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1/2}\mathbf{Z}_1^T\mathbf{Z}_2(\mathbf{Z}_2^T\mathbf{Z}_2)^{-1/2}.$$

Now suppose that $\|\mathbf{Z}_j^T\mathbf{Z}_k/n - \mathbb{E}(\mathbf{Z}_j^T\mathbf{Z}_k/n)\|_S \le C_0v_0\sqrt{t/n + r/n} \le \epsilon_0 < 1$ for $1 \le j \le k \le 2$. Recall that $1 \ge \lambda_1 \ge \cdots \ge \lambda_r > 0$ are the nonzero singular values of $\boldsymbol{\Omega}_{1,2}$ and $\lambda_{\min} = \lambda_r I\{r = r_1 = r_2\}$. As $\mathbb{E}(\mathbf{Z}_k^T\mathbf{Z}_k/n) = \mathbf{I}_{r_k\times r_k}$, we have

$\mathrm{rank}(\mathbf{P}_k) = r_k$. Moreover, as $\mathbb{E}(\mathbf{Z}_1^T\mathbf{Z}_2/n) = \mathbf{U}_1^T\mathbf{\Omega}_{1,2}\mathbf{U}_2$ with unitary maps $\mathbf{U}_1$ and $\mathbf{U}_2$, the Weyl inequality implies that

$$\widehat{\lambda}_1 \le \frac{\lambda_1(1+\epsilon_0)}{1-\epsilon_0}, \quad \widehat{\lambda}_{\min} \ge \frac{\lambda_{\min}(1-\epsilon_0)}{1+\epsilon_0}.$$

Thus, (2.40) holds. As the conditions for $\lambda_1 < 1$ and $\lambda_{\min} > 0$ follow from the positive-definiteness of $\mathbf{\Sigma}$, the proof is complete. □

Proof of Theorem 5. The KKT conditions for the group Lasso asserts that

$$\begin{aligned} &\frac{1}{n}\mathbf{X}_{G_j}^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \omega_j\widehat{\boldsymbol{\beta}}_{G_j}/\|\widehat{\boldsymbol{\beta}}_{G_j}\|_2, \quad && \widehat{\boldsymbol{\beta}}_{G_j} \ne \mathbf{0}, \\ &\frac{1}{n}\|\mathbf{X}_{G_j}^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\|_2 \le \omega_j, && \widehat{\boldsymbol{\beta}}_{G_j} = \mathbf{0}. \end{aligned} \tag{A.1}$$

Let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. It follows that in the event $\mathcal{E}$

$$\frac{\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{h}\|_2}{\omega_j n} = \frac{\|\mathbf{X}_{G_j}^T(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y} + \boldsymbol{\varepsilon})\|_2}{\omega_j n} \le 1 + \frac{\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2}{\omega_j n} \le \frac{2\xi}{\xi+1}. \tag{A.2}$$

It also follows from (A.1) that in the event $\mathcal{E}$

$$\begin{aligned} & \boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{h}/n \\ = \; & \boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y} + \boldsymbol{\varepsilon})/n \\ \le \; & \begin{cases} \omega_j\|\boldsymbol{h}_{G_j}\|_2 + |\boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}|/n, & j \in S^*, \\ -\omega_j\|\boldsymbol{h}_{G_j}\|_2 + |\boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}|/n, & j \notin S^*, \end{cases} \\ \le \; & \begin{cases} \omega_j\|\boldsymbol{h}_{G_j}\|_2 2\xi/(\xi+1), \ j \in S^*, \\ -\omega_j\|\boldsymbol{h}_{G_j}\|_2 2/(\xi+1), \ j \notin S^*. \end{cases} \end{aligned} \tag{A.3}$$

Summing the above inequality over j, we have

$$\|\mathbf{X}\boldsymbol{h}\|_2^2/n \le \frac{2\xi}{\xi+1}\sum_{j\in S^*}\omega_j\|\boldsymbol{u}_{G_j}\|_2 - \frac{2}{\xi+1}\sum_{j\notin S^*}\omega_j\|\boldsymbol{u}_{G_j}\|_2.$$

This and (A.3) implies $\boldsymbol{h} \in \mathscr{C}_-^{(G)}(\xi, \boldsymbol{\omega}, S^*)$. Thus, by (3.6) and (A.2)

$$\begin{aligned} & \|\mathbf{X}\boldsymbol{h}\|_2^2/n \\ & \le \{2\xi/(\xi+1)\}\sum_{j\in S^*}\omega_j\|\widehat{\boldsymbol{\beta}}_{G_j} - \boldsymbol{\beta}_{G_j}^*\|_2 \\ & \le \{2\xi/(\xi+1)\}\max_j \omega_j^{-1}\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{h}\|_2 \textstyle\sum_{j\in S^*}\omega_j^2/\{n\,\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*)\} \\ & \le \{2\xi/(\xi+1)\}^2 \textstyle\sum_{j\in S^*}\omega_j^2/\{n\,\mathrm{SCIF}_1^{(G)}(\xi, \boldsymbol{\omega}, S^*, S^*)\}. \end{aligned}$$

Similarly, (3.6) and (A.2) yield

$$
\begin{aligned}
&\Big(\textstyle\sum_{j=1}^{M}\omega_j^2(\|\boldsymbol{h}_{G_j}\|_2/\omega_j)^q\Big)^{1/q}\\
&\le \big(\textstyle\sum_{j\in S^*}\omega_j^2\big)^{1/q}\max_j \omega_j^{-1}\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{h}\|_2/\{\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},S^*)\}\\
&\le \{2\xi/(\xi+1)\}\big(\textstyle\sum_{j\in S^*}\omega_j^2\big)^{1/q}\{\mathrm{SCIF}_q^{(G)}(\xi,\boldsymbol{\omega},S^*,T^*)\}.
\end{aligned}
$$

Finally, we prove (3.11). Let $\mathbf{Q}_{G_j}$ be the orthogonal projection to the range of $\mathbf{X}_{G_j}$. As $\boldsymbol{\varepsilon}\sim \mathsf{N}_n(\mathbf{0},\sigma^2\mathbf{I}_n)$, $\|\mathbf{Q}_{G_j}\boldsymbol{\varepsilon}/\sigma\|_2^2\sim\chi^2_{d_j'}$ with $d_j'=\mathrm{rank}(\mathbf{Q}_{G_j})\le d_j$. Thus, it follows from the Gaussian concentration inequality that for any $0<\delta<1$, with probability at least $1-\delta$,

$$
\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2/(\sigma\|\mathbf{X}_{G_j}\|_S)\le\|\mathbf{Q}_{G_j}\boldsymbol{\varepsilon}/\sigma\|_2\le\sqrt{n}\left\{\sqrt{d_j}+\sqrt{2\log(1/\delta)}\right\}.
$$

The result in (3.11) follows by an application of the union bound. □

Proof of Lemma 2. For $\eta\ge 0$ define

$$
\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta},\sigma,\eta)=\frac{\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2}{2n\sigma}+\frac{\sigma}{2}+\sum_{j=1}^{M}\omega_j\|\boldsymbol{\beta}_{G_j}\|_2^{1+\eta}+\frac{\eta\sigma^2}{2}
$$

and $\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta)=\arg\min_{\boldsymbol{\beta}}\ \mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta},\sigma,\eta)$. As $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta},\sigma,\eta)$ is convex in $(\boldsymbol{\beta},\sigma)$, the profile loss $\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta),\sigma,\eta)$ is convex in σ for all $\eta\ge0$. Note that for $\eta>0$

$$
\begin{aligned}
&\frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta),\sigma,\eta)\\
=\ &\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\theta},\sigma,\eta)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta)}\right\}^T\frac{\partial\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta)}{\partial\sigma}+\frac{\partial}{\partial t}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega}),t,\eta)\Big|_{t=\sigma}\\
=\ &1/2-\|\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta)\|_2^2/(2n\sigma^2)+\eta\sigma
\end{aligned}
$$

as all derivatives involved are continuous. Moreover, as $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta},\sigma)=\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta},\sigma,0)$ is strictly convex in $\mathbf{X}\boldsymbol{\beta}$,

$$
\lim_{\eta\to0+}\frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta),\sigma,\eta)\to1/2-\|\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})\|_2^2/(2n\sigma^2).
$$

Consequently,

$$
\begin{aligned}
\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma_2\boldsymbol{\omega}),\sigma_2)-\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma_1\boldsymbol{\omega}),\sigma_1)&=\lim_{\eta\to0+}\int_{\sigma_1}^{\sigma_2}\left\{\frac{\partial}{\partial\sigma}\mathcal{L}_{\boldsymbol{\omega}}(\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega},\eta),\sigma,\eta)\right\}d\sigma\\
&=\int_{\sigma_1}^{\sigma_2}\left\{1/2-\|\boldsymbol{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}(\sigma\boldsymbol{\omega})\|_2^2/(2n\sigma^2)\right\}d\sigma.
\end{aligned}
$$

All other claims follow from the joint convexity of $\mathcal{L}_{\boldsymbol{\omega}}(\boldsymbol{\beta},\sigma)$ and the strict convexity of the loss function in $\mathbf{X}\boldsymbol{\beta}$. □

Proof of Theorem 6. We follow the proof in Sun and Zhang (2012b). Let $t \ge \sigma^*/\sqrt{1+\tau_-}$ and $\boldsymbol{h}_{G_j} = \widehat{\boldsymbol{\beta}}_{G_j}(t\boldsymbol{\omega}) - \boldsymbol{\beta}^*_{G_j}$. As the oracle noise level is $(\sigma^*)^2 = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2/n$, we have

$$\begin{aligned}(\sigma^*)^2 - \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n &= (\mathbf{X}\boldsymbol{h})^T(2\boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{h})/n \\ &= (\mathbf{X}\boldsymbol{h})^T(\boldsymbol{\varepsilon} + \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega}))/n. \qquad \text{(A.4)}\end{aligned}$$

Suppose $\mathcal{E}$ happens so that $\|\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}\|_2/n \le t\omega_j(\xi-1)/(\xi+1)$. It follows that

$$\left|(\mathbf{X}\boldsymbol{h})^T\boldsymbol{\varepsilon}/n\right| = \left|\sum_{j=1}^M \boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T\boldsymbol{\varepsilon}/n\right| \le \frac{\xi-1}{\xi+1}\sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2.$$

Moreover, the KKT condition implies

$$\left|\boldsymbol{h}^T\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega}))/n\right| = \left|\sum_{j=1}^M \boldsymbol{h}_{G_j}^T\mathbf{X}_{G_j}^T(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega}))/n\right| \le \sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2.$$

As $(\mathbf{X}\boldsymbol{h})^T(2\boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{h})/n \le 2(\mathbf{X}\boldsymbol{h})^T\boldsymbol{\varepsilon}/n$, inserting these inequalities to (A.4) yields

$$\begin{aligned}-\left(\frac{\xi-1}{\xi+1}+1\right)\sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2 &\le {\sigma^*}^2 - \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n \\ &\le 2\frac{\xi-1}{\xi+1}\sum_{j=1}^M t\omega_j\|\boldsymbol{h}_{G_j}\|_2.\end{aligned}$$

A rescaled version $\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})$ can be written as

$$\frac{\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})}{t} = \arg\min_{\boldsymbol{b}}\left\{\frac{\|\boldsymbol{y}/t - \mathbf{X}\boldsymbol{b}\|_2^2}{2n} + \sum_{j=1}^M \omega_j\|\boldsymbol{b}_{G_j}\|_2\right\}$$

as the group Lasso estimator with target $\boldsymbol{\beta}^*/t$ and noise vector $\boldsymbol{\varepsilon}/t$. As $t \ge \sigma^*/\sqrt{1+\tau_-}$, the condition of Theorem 5 is satisfied with the rescaled noise $\boldsymbol{\varepsilon}/t$, so that

$$t^{-1}\sum_{j=1}^M \omega_j\|\boldsymbol{h}_{G_j}\|_2 = \sum_{j=1}^M \omega_j\|\widehat{\boldsymbol{\beta}}_{G_j}(t\boldsymbol{\omega})/t - \boldsymbol{\beta}^*_{G_j}/t\|_2 < \mu(\boldsymbol{\omega},\xi).$$

As $\tau_- = 2\mu(\boldsymbol{\omega},\xi)(\xi-1)/(\xi+1)$ and $\tau_+ = \mu(\boldsymbol{\omega},\xi)\{(\xi-1)/(\xi+1)+1\}$, we have

$$\begin{aligned}-\tau_+t^2 = -\left(\frac{\xi-1}{\xi+1}+1\right)t^2\mu(\boldsymbol{\omega},\xi) &< {\sigma^*}^2 - \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n \\ &< 2\frac{\xi-1}{\xi+1}t^2\mu(\boldsymbol{\omega},\xi) = \tau_-t^2.\end{aligned}$$

The upper bound above for $t = \sigma^*/\sqrt{1+\tau_-}$ implies

$$t^2 - \|y - \mathbf{X}\widehat{\boldsymbol{\beta}}(t\boldsymbol{\omega})\|_2^2/n < t^2 - {\sigma^*}^2 + \tau_- t^2 = 0,$$

so that $\widehat{\sigma} > t = \sigma^*/\sqrt{1+\tau_-}$ by Lemma 2. Similarly, the lower bound yields $\widehat{\sigma} < \sigma^*/\sqrt{1-\tau_+}$.

As $\widehat{\sigma} > \sigma^*/\sqrt{1+\tau_-}$, the error bounds in Theorem 5 holds for $\{\boldsymbol{y}/\widehat{\sigma}, \boldsymbol{\beta}^*/\widehat{\sigma}, \widehat{\boldsymbol{\beta}}/\widehat{\sigma}\}$, which implies (3.18) and (3.19) due to $\widehat{\sigma} < \sigma^*/\sqrt{1-\tau_+}$. When (1.1) holds with Gaussian error, $|\widehat{\sigma}/\sigma^* - 1| = o_P(\mu(\boldsymbol{\omega}, \xi)) = o_P(n^{-1/2})$ by (3.17) and the condition on $\mu(\boldsymbol{\omega}, \xi)$, so that (3.21) follows from the central limit theorem for $\sigma^*/\sigma \sim \chi_n/\sqrt{n}$.

It remains to prove (3.20). Let $\boldsymbol{u}^* = \boldsymbol{\varepsilon}/\|\boldsymbol{\varepsilon}\|_2$, $\mathbf{Q}_{G_j}$ be the orthogonal projection to the range of $\mathbf{X}_{G_j}$, $d_j' = \text{rank}(\mathbf{Q}_{G_j})$, and $f(\boldsymbol{u}^*) = \|\mathbf{Q}_{G_j}\boldsymbol{u}^*\|_2$. As $f(\boldsymbol{u}^*) = 1$ for $n = 1$, we assume $n \geq 2$ without loss of generality. The vector $\boldsymbol{u}^*$ is uniformly distributed in the sphere $\mathbb{S}^{n-1}$ and $f(\boldsymbol{u}^*)$ is a unit Lipschitz function of $\boldsymbol{u}^*$ with median $\sqrt{m_{d_j',n}} \leq \sqrt{m_{d_j,n}}$. As $\sigma^* = \|\boldsymbol{\varepsilon}\|_2/\sqrt{n}$, $\|\mathbf{X}_{G_j}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^*)/(n\sigma^*)\|_2/\|\mathbf{X}_{G_j}/\sqrt{n}\|_S \leq f(\boldsymbol{u}^*)$. Thus, for $t > 0$ and $n \geq 2$,

$$\mathbb{P}\left\{\|\boldsymbol{Q}_{G_j}\boldsymbol{u}^*\|_2 \geq \sqrt{m_{d_j,n}} + \frac{t}{\sqrt{n-3/2}}\right\} \leq e^{(4n-6)^{-2}}\mathbb{P}\Big\{\mathsf{N}(0,1) > t\Big\} \leq e^{-t^2/2}$$

by the Lévy concentration inequality as in Lemma 17 of Sun and Zhang (2013). It follows that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ by the union bound when $(\xi-1)\omega_j/\{(\xi+1)\sqrt{1+\tau_-}\} \geq \|\mathbf{X}_{G_j}/\sqrt{n}\|_S\omega_{*,j}$. Now, consider $\omega_j = A\|\mathbf{X}_{G_j}/\sqrt{n}\|_S\omega_{*,j}$. Let $\tau_* = 2\mu(\boldsymbol{\omega}_*, \xi)(\xi-1)/(\xi+1)$. It follows from (3.1) and (3.6) that $\mu(\boldsymbol{\omega}, \xi) = A^2\mu(\boldsymbol{\omega}^*, \xi)$, so that $\tau_- = A^2\tau_*$. Consequently,

$$\frac{(\xi-1)\omega_j}{(\xi+1)\sqrt{1+\tau_-}\|\mathbf{X}_{G_j}/\sqrt{n}\|_S\omega_{*,j}} = \frac{(\xi-1)A}{(\xi+1)\sqrt{1+A^2\tau_*}} \geq 1$$

if and only if $A \geq \{(\xi+1)/(\xi-1)\}/\{1 - \{(\xi+1)/(\xi-1)\}^2\tau_*\}^{1/2} = A_*$. Finally, we note that $\sqrt{m_{d_j,n}} \leq \mathbb{E}f(\boldsymbol{u}^*) + e^{(4n-6)^{-2}}\mathbb{E}|\mathsf{N}(0, 1/(n-3/2))|/2 \leq (d_j/n)^{1/2} + n^{-1/2}$. □

Proof of Proposition 2. (i) We prove that for every $\boldsymbol{u} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T)$, there exists a non-increasing nonnegative function $h(x)$ and $x_0t_0 \leq t_1 < x_0t_0 + \max_j f_j^2$ such that

$$\|\boldsymbol{u}_{G_T}\|_{(q)}^q = \textstyle\sum_{j\in T} f_j^2(\|\boldsymbol{u}_{G_j}\|_2/f_j)^q \leq \int_0^{t_0} h^q(x)dx, \tag{A.5}$$

$$\|\boldsymbol{u}\|_{(q)}^q = \textstyle\int_0^\infty h^q(x)dx \leq \big\{1 + \big(1+\xi\big)a_q^{1-1/q}\big\}\big(\int_0^{t_0} h^q(x)dx\big)^{1/q}, \tag{A.6}$$

$$\max_{j\leq M}\big[\big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j)\big]t_1^{1/q} \geq \kappa_{q,1}(\xi, t_0, x_0, y_0)\big(\textstyle\int_0^{t_1} h^q(x)dx\big)^{1/q}, \tag{A.7}$$

$$\max_{j\leq M}\big[\|\mathbf{X}_{G_j}^T\mathbf{X}\boldsymbol{u}\|_2/(nf_j)\big]\textstyle\int_0^{t_1} h(x)dx \geq \kappa_{2,1}(\xi, t_0, x_0, y_0)\int_0^{t_1} h^2(x)dx, \tag{A.8}$$

$$\|\mathbf{X}\boldsymbol{u}\|_2^2/n \geq \kappa_{2,2}(\xi, t_0, x_0, y_0)\textstyle\int_0^{t_1} h^2(x)dx. \tag{A.9}$$

Moreover, for $\boldsymbol{u} \in \mathscr{C}_{-}^{(G)}(\xi, \boldsymbol{\omega}, T)$,

$$\max_{j \le M}\big[\|\mathbf{X}_{G_j}^T \mathbf{X}\boldsymbol{u}\|_2/(nf_j)\big] \int_0^{t_0} h(x)dx \ge \kappa_{2,1}(\xi, t_0, x_0, y_0) \int_0^{t_1} h^2(x)dx. \tag{A.10}$$

In fact, as $\omega_j \propto f_j$, (3.28) and (3.29) follow from (3.2), (3.3), (A.5), (A.6) and (A.9), (3.30) follows from (3.4), (A.5), (A.6) and (A.7), and (3.31) follows from (3.6), (A.5), (A.6) and (A.10). As these steps of the proof are similar, we only provide the following example:

$$\mathrm{SCIF}_q^{(G)}(\xi, \boldsymbol{\omega}, T, T^*) \ge \inf_h \frac{t_0^{1/q}\kappa_{2,1}(\xi, t_0, x_0, y_0)\int_0^{t_1} h^2(x)dx}{\int_0^{t_0} h(x)dx\big(\int_0^\infty h^q(x)dx\big)^{1/q}} \ge \frac{\kappa_{2,1}(\xi, t_0, x_0, y_0)}{1 + \big(1+\xi\big)a_q^{1-1/q}}$$

for $1 \le q \le 2$ with an application of the Hölder inequality.

Let us prove (A.5)-(A.10) for a fixed $\boldsymbol{u} \in \mathscr{C}^{(G)}(\xi, \boldsymbol{\omega}, T)$. Relabelling the groups if necessary, we assume without loss of generality that $\|\boldsymbol{u}_{G_j}\|_2/f_j \ge \|\boldsymbol{u}_{G_{j+1}}\|_2/f_{j+1}$ for all $1 \le j < M$. Let $s_0 = 0$ and $s_j = \sum_{\ell=1}^j f_\ell^2$ for $1 \le j \le M$. Define $h(x) = \|\boldsymbol{u}_{G_j}\|_2/f_j$ for $s_{j-1} < x \le s_j$, $1 \le j \le M$, and $h(x) = 0$ for $x > s_M$. The identities in (A.5) and (A.6) follow from

$$\int_{s_{j-1}}^{s_j} h^q(x)dx = f_j^2(\|\boldsymbol{u}_{G_j}\|_2/f_j)^q. \tag{A.11}$$

As $t_0 = \sum_{j\in T} f_j^2$ and $h(x)$ is nondecreasing in $(0,\infty)$, $\sum_{j\in T} f_j^2(\|\boldsymbol{u}_{G_j}\|_2/f_j)^q \le \int_0^{t_0} h^q(x)dx$. This gives the inequality in (A.5). It follows from (A.5) and the identity in (A.6) that $\int_0^\infty h(x)dx \le (1+\xi)\int_0^{t_0} h(x)dx$, so that by the shifting inequality (Cai, Wang and Xu, 2010; Ye and Zhang, 2010, Eq. (62))

$$\Big(\int_{t_0}^\infty h^q(x)dx\Big)^{1/q} \le (a_q/t_0)^{1-1/q}\int_0^\infty h(x)dx \le \big(1+\xi\big)(a_q/t_0)^{1-1/q}\int_0^{t_0} h(x)dx.$$

Thus, the inequality in (A.6) follows with an application of the Hölder inequality.

The proof of (A.7) is a discrete version that of (A.6). Let

$$g_1 = \inf\Big\{j \ge 0 : s_j \ge x_0 t_0\Big\},\ t_1 = s_{g_1},$$

with the convention $\inf\emptyset = M+1$, and for $k > 1$,

$$g_k = \inf\Big\{j \ge g_{k-1} : s_j \ge t_{k-1} + y_0 t_0\Big\},\ t_k = s_{g_k}.$$

Recall that $t_0 = \sum_{j\in T} f_j^2$, $x_0 \ge 1$ and $y_0 \le x_0/a_q$. It follows from (A.11) that

$$\sum_{j=1}^{g_k} f_j^2(\|\boldsymbol{u}_{G_j}\|_2/f_j)^q = \int_0^{t_k} h^q(x)dx,\ k \ge 1. \tag{A.12}$$

As $h(x)$ is non-increasing in x and $(t_k - t_{k-1}) \wedge (t_1/a_q) \geq y_0 t_0$, another application of the shifting inequality (Cai, Wang and Xu, 2010; Ye and Zhang, 2010, Eq. (63)) yields

$$\begin{aligned}
& \sum_{k\geq 2}\Big(\int_{t_{k-1}}^{t_k} h^q(x)dx\Big)^{1/q} \\
\leq \;& \sum_{k\geq 2}(y_0t_0)^{1/q-1}\int_{t_{k-1}-a_qy_0t_0}^{t_k-a_qy_0t_0} h(x\vee t_1)dx \\
= \;& (y_0t_0)^{1/q-1}\int_{t_1-a_qy_0t_0}^{\infty} h(x\vee t_1)dx \\
\leq \;& (y_0t_0)^{1/q-1}\Big(\xi\int_0^{t_0} h(x)dx + \big(t_0-(t_1-a_qy_0t_0)\big)_+ h(t_1)\Big) \\
\leq \;& \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\Big(\xi y_0^{1/q-1} + \big(t_0+a_qy_0t_0-t_1\big)_+ (y_0t_0)^{1/q-1}t_1^{-1/q}\Big) \\
\leq \;& \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\Big(\xi y_0^{1/q-1} + \big(1+a_qy_0-x_0\big)_+ y_0^{1/q-1}x_0^{-1/q}\Big) \\
= \;& \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\big(\rho_q - \kappa_{q,1}(\xi,t_0,x_0,y_0)\big)/\theta_q(x_0t_0,y_0t_0). \qquad \text{(A.13)}
\end{aligned}$$

Let $B_1 = \{1,\ldots,g_1\}$ and $B_k = \{g_{k-1}+1,\ldots,g_k\}$ for $k \geq 2$. Let

$$\boldsymbol{v} = \arg\max_{\boldsymbol{w}}\Big\{\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}_{G_{B_1}}\boldsymbol{u}_{G_{B_1}}/n : \mathrm{supp}(\boldsymbol{w}) \subseteq G_{B_1}, \|\boldsymbol{w}\|_{(q/(q-1))} = 1\Big\}.$$

As $\sum_{j=1}^{g_1-1} f_j^2 \leq x_0t_0$, it follows from (3.25) and (A.12) that

$$\boldsymbol{v}^T\mathbf{X}^T\mathbf{X}_{G_{B_1}}\boldsymbol{u}_{G_{B_1}}/n \geq \rho_q(x_0t_0)\|\boldsymbol{u}_{G_{B_1}}\|_{(q)} = \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\rho_q(x_0t_0).$$

By (3.26), $\big|\boldsymbol{v}^T\mathbf{X}^T\mathbf{X}_{G_{B_k}}\boldsymbol{u}_{G_{B_k}})\big| \leq \theta_q(x_0t_0,y_0t_0)\big\|\boldsymbol{u}_{G_{B_k}}\big\|_{(q)}$, so that by (A.12) and (A.13),

$$\begin{aligned}
& \boldsymbol{v}^T(\mathbf{X}^T\mathbf{X}/n)\boldsymbol{u} \geq \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\rho_q(x_0t_0) - \sum_{k>1}\theta_q(x_0t_0,y_0t_0)\big\|\boldsymbol{u}_{G_{B_k}}\big\|_{(q)} \\
& = \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\rho_q(x_0t_0) - \sum_{k>1}\theta_q(x_0t_0,y_0t_0)\Big(\int_{t_{k-1}}^{t_k} h^q(x)dx\Big)^{1/q} \\
& \geq \Big(\int_0^{t_1} h^q(x)dx\Big)^{1/q}\kappa_{q,1}(\xi,t_0,x_0,y_0).
\end{aligned}$$

This yields (A.7) via

$$\begin{aligned}
\boldsymbol{v}^T(\mathbf{X}^T\mathbf{X}/n)\boldsymbol{u} \;&\leq\; \textstyle\sum_{j\in B_1} f_j^2\big(\|\boldsymbol{v}_{G_j}\|_2/f_j\big)\max_{j\leq M}\big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j) \\
&\leq\; \|\boldsymbol{v}\|_{(q/(q-1))}\big(\textstyle\sum_{j\in B_1} f_j^2\big)^{1/q}\max_{j\leq M}\big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j) \\
&=\; t_1^{1/q}\max_{j\leq M}\big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j).
\end{aligned}$$

For $q = 2$, $\rho_q(s)$ is the group-sparse eigenvalue of the Gram matrix as explained below (3.26), so that $\rho_2(s)$ is attained with $\boldsymbol{v}_{G_B} = \boldsymbol{u}_{G_B}/\|\boldsymbol{u}_{G_B}\|_2$. This gives (A.8) with the following modification of the proof of (A.7):

$$\begin{aligned}\boldsymbol{v}^T(\mathbf{X}^T\mathbf{X}/n)\boldsymbol{u} &\le \textstyle\sum_{j\in B_1} f_j^2\big(\|\boldsymbol{u}_{G_j}\|_2/f_j\big)\big\{\max_{j\le M}\big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j)\big\}/\|\boldsymbol{u}_{G_{B_1}}\|_2 \\ &\le \int_0^{t_1} h(x)dx\Big(\int_0^{t_1} h^2(x)dx\Big)^{-1/2}\max_{j\le M}\big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j).\end{aligned}$$

Similarly, (A.9) follows from

$$\begin{aligned}\|\mathbf{X}\boldsymbol{u}\|_2^2/n &\ge \|\mathbf{X}_{G_{B_1}}\boldsymbol{u}_{G_{B_1}}\|_2^2/n + 2\boldsymbol{u}_{G_{B_1}}^T\mathbf{X}_{G_{B_1}}^T\big(\mathbf{X}\boldsymbol{u} - \mathbf{X}_{G_{B_1}}\boldsymbol{u}_{G_{B_1}}\big)/n \\ &\ge \kappa_{2,2}(\xi, t_0, x_0, y_0)\int_0^{t_1} h^2(x)dx.\end{aligned}$$

Finally, for $\boldsymbol{u} \in \mathscr{C}_-^{(G)}(\xi, \boldsymbol{\omega}, T)$, we have (A.10) via (A.5) and

$$\begin{aligned}(\mathbf{X}_{G_{B_1}}\boldsymbol{u}_{G_{B_1}})^T(\mathbf{X}\boldsymbol{u})/n &\le \textstyle\sum_{j\in T}\boldsymbol{u}_{G_j}^T\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}/n \\ &\le \textstyle\sum_{j\in T} f_j\|\boldsymbol{u}_{G_j}\|_2 \max\limits_j \big\|\mathbf{X}_{G_j}\mathbf{X}\boldsymbol{u}\big\|_2/(nf_j).\end{aligned}$$

(ii) Let $f_j = \omega_j/C_n$. Consider the event $c_*(1-\epsilon_0) \le \|\mathbf{X}_{G_j}/\sqrt{n}\|_S \le (1+\epsilon_0)c^*$ for all $j \le M$, in which $f_j \asymp |G_j|^{1/2} + \sqrt{2\log(M/\delta)}$. Let $g^* = \max\big\{|B| : |B \setminus S| \le 2, \|\boldsymbol{f}_S\|_2^2 < (x_0 \vee y_0)t_0\big\}$ be the largest number of groups involved in the definition of $\rho_-(x_0t_0)$ and $\theta(x_0t_0, y_0t_0)$, and $s^* = \max\big\{|G_B| : |B \setminus S| \le 2, \|\boldsymbol{f}_S\|_2^2 < (x_0 \vee y_0)t_0\big\}$ be the largest number of variables involved. As $f_j \asymp |G_j|^{1/2} + \sqrt{2\log(M/\delta)}$ and (x_0, y_0) is fixed, we have

$$s^* + 2g^*\log(M/\delta) \lesssim t_0 + \max_j f_j^2 \lesssim n_*$$

with $n_* = \max_{j\le M}\big\{|G_j| + \log(M/\delta)\big\} + |G_T| + |T|\log(M/\delta)$.

The conclusion follows from part (i) and Lemma 1. Let

$$\begin{aligned}\Omega_n \;=\; \Big\{&c_*(1-\epsilon_0) \le \|\mathbf{X}_{G_j}/\sqrt{n}\|_S \le (1+\epsilon_0)c^* \ \forall j, \\ &\rho_2(x_0t_0) \le (1-\epsilon_0/2)\lambda_{\min}(\boldsymbol{\Sigma}), \theta_2(x_0t_0, y_0t_0) \ge (1+\epsilon_0)c^*\Big\}.\end{aligned}$$

Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be the orthogonal projections to the subspace of vectors $\boldsymbol{v} \in \mathbb{R}^p$ with support sets G_{B_1} and G_{B_2} respectively, $t = (2g^*+2)\log(M/\delta)$ and $\epsilon_0 = C_1\sqrt{t/n + s^*/n}$ with a sufficiently large C_1. Since $\{s^* + 2g^*\log(M/\delta)\}/n$ is small for small a_0, Lemma 1 yields $\mathbb{P}\big\{\Omega_n\big\} \le \binom{M}{g^*}^2 e^{-t} \le (\delta/M)^2$. For $x_0 = a_qy_0$ and sufficiently large y_0, $\kappa_{2,m}(\xi, t_0, x_0, y_0) \ge \rho_2(x_0t_0)(1-\epsilon_0/2)$ in Ω_n. The conclusions of part (ii) then follow from part (i). □

References

Antoniadis, A. (2010). Comments on: ℓ_1-penalization for Mixture Regression Models. *Test* **19** 257–258. MR2677723

Bach, F. R. (2008). Consistency of the Group Lasso and Multiple Kernel Learning. *The Journal of Machine Learning Research* **9** 1179–1225. MR2417268

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika* **98** 791–806.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies* **81** 608–650. MR3207983

Berk, R., Brown, L. B. and Zhao, L. (2010). Statistical Inference After Model Selection. *Journal of Quantitative Criminology* **26** 217–236.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *The Annals of Statistics* **37** 1705–1732. MR2533469

Bickel, P. J., Klaassen, J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore. MR1245941

Breheny, P. and Huang, J. (2011). Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396

Bühlmann, P. (2013). Statistical Significance in High-Dimensional Linear Models. *Bernoulli* **19** 1212–1242. MR3102549

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer. MR2807761

Bunea, F., Lederer, J. and She, Y. (2014). The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms. *Information Theory, IEEE Transactions on* **60** 1313–1325. MR3164977

Cai, T., Liu, W. and Luo, X. (2011). A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association* **106** 594–607. MR2847973

Cai, T., Wang, L. and Xu, G. (2010). Shifting Inequality and Recovery of Sparse Signals. *IEEE Transactions on Signal Processing* **58** 1300–1308. MR2730209

Candes, E. J. and Tao, T. (2005). Decoding by Linear Programming. *IEEE Trans. on Information Theory* **51** 4203–4215. MR2243152

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2014). High-Dimensional Inference: Confidence Intervals, p-values and R-Software `hdi`. *arXiv preprint arXiv:1408.4026*.

Huang, J., Breheny, P. and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science* **27** 481–499. MR3025130

Huang, J. and Zhang, T. (2010). The Benefit of Group Sparsity. *The Annals of Statistics* **38** 1978–2004. MR2676881

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A Group Bridge Approach for Variable Selection. *Biometrika* **96** 339–355. MR2507147

Huber, P. J. (2011). *Robust Statistics*. Springer.

Jankova, J. and van de Geer, S. (2014). Confidence Intervals for High-

Dimensional Inverse Covariance Estimation. *arXiv preprint arXiv:1403.6752*. MR3354336

Javanmard, A. and Montanari, A. (2014a). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *The Journal of Machine Learning Research* **15** 2869–2909. MR3277152

Javanmard, A. and Montanari, A. (2014b). Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory. *IEEE Transactions on Information Theory* **60** 6522–6554. MR3265038

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *The Annals of Statistics* **28** 1356–1378. MR1805787

Koltchinskii, V. (2009). The Dantzig Selector and Sparsity Oracle Inequalities. *Bernoulli* **15** 799–828. MR2555200

Koltchinskii, V. and Yuan, M. (2008). Sparse Recovery in Large Ensembles of Kernel Machines. In *Proceedings of COLT*.

Leeb, H. and Potscher, B. M. (2006). Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators? *The Annals of Statistics* **34** 2554–2591. MR2291510

Liu, H. and Zhang, J. (2009). Estimation Consistency of the Group Lasso and Its Applications. *Journal of Machine Learning Research-Proceedings Track* **5** 376–383.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A Significance Test for the Lasso. *The Annals of Statistics* **42** 413–468. MR3210970

Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A. B. (2011). Oracle Inequalities and Optimal Inference Under Group Sparsity. *The Annals of Statistics* **39** 2164–2204. MR2893865

Ma, Z. (2013). Sparse Principal Component Analysis and Iterative Thresholding. *The Annals of Statistics* **41** 772–801. MR3099121

Meinshausen, N. (2014). Group Bound: Confidence Intervals for Groups of Variables in Sparse High Dimensional Regression without Assumptions on the Design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. MR3414134

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 417–473. MR2758523

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for High-Dimensional Regression. *Journal of the American Statistical Association* **104**.

Nardi, Y. and Rinaldo, A. (2008). On the Asymptotic Properties of the Group Lasso Estimator for Linear Models. *Electronic Journal of Statistics* **2** 605–633. MR2426104

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2008). Union support recovery in high-dimensional multivariate regression. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on* 21–26. IEEE.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse

Additive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 1009–1030. MR2750255

Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H. H. (2013). Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Model. *arXiv preprint arXiv:1309.6024*. MR3346695

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright Inequality and Sub-Gaussian Concentration. *Electronic Communications in Probability* 1–9. MR3125258

Shah, R. D. and Samworth, R. J. (2013). Variable Selection with Error Control: Another Look at Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 55–80. MR3008271

Städler, N., Bühlmann, P. and Geer, S. (2010). ℓ_1-Penalization for Mixture Regression Models. *TEST* **19** 209–256. MR2677722

Sun, T. and Zhang, C.-H. (2010). Comments on: ℓ_1-Penalization for Mixture Regression Models. *Test* **19** 270–275. MR2677726

Sun, T. and Zhang, C.-H. (2012a). Comments on: Optimal Rates of Convergence for Sparse Covariance Matrix Estimation. *Statistica Sinica* **22** 1354–1358. MR3027086

Sun, T. and Zhang, C.-H. (2012b). Scaled Sparse Linear Regression. *Biometrika* **99** 879–898. MR2999166

Sun, T. and Zhang, C.-H. (2013). Sparse Matrix Inversion with Scaled Lasso. *Journal of Machine Learning Research* **14** 3385–3418. MR3144466

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288. MR1379242

van de Geer, S. (2007). The Deterministic Lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

van de Geer, S. (2014). Worst Possible Sub-Directions in High-Dimensional Models. *Contributions in infinite-dimensional statistics and related topics* 131.

van de Geer, S. and Bühlmann, P. (2009). On the Conditions Used to Prove Oracle Results for the Lasso. *Electronic Journal of Statistics* **3** 1360–1392. MR2576316

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *The Annals of Statistics* **42** 1166–1202. MR3224285

Vershynin, R. (2011). Spectral Norm of Products of Random and Deterministic Matrices. *Probability theory and related fields* **150** 471–509. MR2824864

Wasserman, L. and Roeder, K. (2009). High Dimensional Variable Selection. *Annals of statistics* **37** 2178. MR2543689

Ye, F. and Zhang, C.-H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. *The Journal of Machine Learning Research* **11** 3519–3540. MR2756192

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574

Zhang, C.-H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of Statistics* 894–942. MR2604701

Zhang, C.-H. (2011). Statistical inference for high-dimensional data. In *Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report No. 48/2011* 28–31.

Zhang, C.-H. and Huang, J. (2008). The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *Annals of Statistics* **36** 1567–1594. MR2435448

Zhang, C.-H. and Zhang, S. S. (2014). Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242. MR3153940