Electronic Journal of Statistics
Vol. 11 (2017) 3039–3080
ISSN: 1935-7524
DOI: 10.1214/17-EJS1309

Estimator augmentation with applications in high-dimensional group inference

Qing Zhou* **and Seunghyun Min**

Department of Statistics, University of California, Los Angeles, CA 90095
e-mail: zhou@stat.ucla.edu; seunghyun@ucla.edu

Abstract: To make statistical inference about a group of parameters on high-dimensional data, we develop the method of estimator augmentation for the block lasso, which is defined via block norm regularization. By augmenting a block lasso estimator $\hat{\beta}$ with the subgradient S of the block norm evaluated at $\hat{\beta}$, we derive a closed-form density for the joint distribution of $(\hat{\beta}, S)$ under a high-dimensional setting. This allows us to draw from an estimated sampling distribution of $\hat{\beta}$, or more generally any function of $(\hat{\beta}, S)$, by Monte Carlo algorithms. We demonstrate the application of estimator augmentation in group inference with the group lasso and a debiased group lasso constructed as a function of $(\hat{\beta}, S)$. Our numerical results show that importance sampling via estimator augmentation can be orders of magnitude more efficient than parametric bootstrap in estimating tail probabilities for significance tests. This work also brings new insights into the geometry of the sample space and the solution uniqueness of the block lasso. To broaden its application, we generalize our method to a scaled block lasso, which estimates the error variance simultaneously.

MSC 2010 subject classifications: Primary 62J07, 62F40; secondary 65C05.
Keywords and phrases: Estimator augmentation, group inference, group lasso, importance sampling, parametric bootstrap, sampling distribution, scaled block lasso.

Received December 2016.

1. Introduction

There has been a fast growth of high-dimensional data in many areas, such as genomics and the social sciences. Statistical inference for high-dimensional models becomes a necessary tool for scientific discoveries from such data. For example, significance tests have been performed to screen millions of genomic loci for disease markers. These applications have motivated the recent development in high-dimensional statistical inference. Some methods make use of sample splitting and subsampling to quantify estimation errors and significance [10, 11, 25], while others rely on the bootstrap to approximate the sampling distributions of lasso-type estimators [3, 29]. For Gaussian linear models, an interesting idea of

*Research supported by NSF grants DMS-1055286 and IIS-1546098.

de-biasing the lasso [20] has been developed by a few groups [6, 22, 27]. In addition, there are various other inferential methods for high-dimensional models [7, 8, 15, 16, 24], some of which are reviewed in [4].

1.1. Group inference

In this article, we consider a linear model

$$y = X\beta_0 + \varepsilon, \tag{1.1}$$

where $\beta_0 \in \mathbb{R}^p$ is the unknown parameter of interest, $y \in \mathbb{R}^n$ is a response vector, $X = [X_1 \mid \cdots \mid X_p] \in \mathbb{R}^{n\times p}$ is a design matrix and $\varepsilon \in \mathbb{R}^n$ is an error vector with mean zero and variance σ^2. Define $\mathbb{N}_k = \{1, \ldots, k\}$ for an integer $k \geq 1$. We are interested in making inference about a group of the parameters, $\beta_{0G} = (\beta_{0j})_{j\in G}$, for $G \subset \mathbb{N}_p$ under a high-dimensional setting that $p > n$. To be specific, the goal is to test the null hypothesis $H_{0G} : \beta_{0G} = 0$ and construct confidence regions for β_{0G}. These are arguably the most general inference problems, including individual inference about β_{0j} as a special case when we choose G to be a singleton.

Group inference arises naturally in applications where predictors have a block structure. For instance, inference about a group of genomic loci within the same gene for its association with a disease can identify responsive genes for the disease. Even if there is no application-driven block structure among the predictors, group inference may still be useful. By grouping variables, one can detect signals that are too small to detect individually. High correlation among predictors is a well-known difficulty for the lasso and related individual inference approaches. In this situation, grouping highly correlated predictors with the group lasso [26] will greatly stabilize the inference and increase detection power. Due to these advantages and practical usage, a few methods have been developed in recent papers for group inference. A de-biased group lasso is proposed by Mitra and Zhang [12] as a generalization of the de-biased lasso for more efficient group inference. van de Geer and Stucky [23] define a de-sparsified estimator for β_{0G} with a surrogate Fisher information matrix constructed by a multivariate square-root lasso. Meinshausen [9] develops the group-bound method to construct a one-sided confidence interval for $\|\beta_{0G}\|_1$ and shows that it is possible to detect the joint contribution of a group of highly correlated predictors even when each has no significant individual effect. Zhou and Min [30] establish that a modified parametric bootstrap is asymptotically valid for the group lasso and demonstrate the advantages of grouping in finite-sample inference.

A large portion of the above methods perform statistical inference based on the sampling distribution of an estimator $\hat{b} = \hat{b}(\hat{\beta}, y, X)$ constructed as a function of $\hat{\beta}$, which is either the lasso or the group lasso depending on whether group structure is used. Examples of such an estimator $\hat{b}$ include the de-biased lasso, the de-biased group lasso, and the trivial case $\hat{b} = \hat{\beta}$ in those methods that directly estimate the distribution of $\hat{\beta}$. There are two big challenges in these approaches. First, the finite-sample distribution of $\hat{b}$ is not well-understood, due

to the high dimension ($p > n$) and the sparsity of $\hat{\beta}$. Consequently, the bootstrap has been used to approximate this distribution for inference. Although the de-biased estimators follow a nice asymptotic normal distribution when $n \to \infty$, they can be far from normally distributed when n is finite. Indeed, recent papers have proposed to bootstrap the de-biased lasso as a better alternative [5, 28]. Then, here comes the second challenge: How to efficiently simulate from the bootstrap distribution or an estimated sampling distribution of $\hat{b}$? Without an explicit characterization of the finite-sample distribution of the group lasso (or lasso), it appears that one can only rely on bootstrap, which can be computationally inefficient, or even impractical, for some calculations, such as approximating tail probabilities in significance tests and conditional sampling given a selected model in post-selection inference.

1.2. Contributions of this work

To meet the aforementioned challenges in group inference, we develop the method of estimator augmentation for the block lasso. Partition the predictors into J disjoint groups $\mathcal{G}_j \subset \mathbb{N}_p$ for $j = 1, \ldots, J$. For $\beta = (\beta_1, \ldots, \beta_p)$, let $\beta_{(j)} = (\beta_k)_{k \in \mathcal{G}_j}$ for $j \in \mathbb{N}_J$. Given $\alpha \in [1, \infty]$, the block lasso is defined via minimizing a penalized loss function $L(\beta; \alpha)$:

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ L(\beta; \alpha) := \frac{1}{2n} \|y - X\beta\|^2 + \lambda \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_\alpha \right\}, \tag{1.2}$$

where $\|\cdot\|$ denotes the Euclidean norm. The weight $w_j > 0$ usually depends on the group size $p_j = |\mathcal{G}_j|$. The regularizer is the block-$(1, \alpha)$ norm (when $w_j = 1$) of β, hence the name block lasso. Note that the lasso and the group lasso can be regarded as the special cases of $\alpha = 1$ and $\alpha = 2$, respectively. In the context of group inference, we can always choose a partition so that $G = \mathcal{G}_j$ for some j, which translates our task into inference about $\beta_{0(j)}$ using some function of $\hat{\beta}$. Instead of the distribution of $\hat{\beta}$, we work with the joint distribution of a so-called augmented estimator $(\hat{\beta}, S)$, where $S = S(y, X) \in \mathbb{R}^p$ is a vector. Under a particular choice of S, we are able to obtain a closed-form density for the *exact* distribution of the augmented estimator for any finite n and p and for all $\alpha \in [1, \infty]$. Given the density, one may use Monte Carlo methods, such as importance sampling, to draw from the joint distribution and simultaneously obtain samples of $\hat{\beta}$ and any function of $\hat{\beta}$, such as the estimator $\hat{b}$ used in an inferential method. This method serves as a powerful and efficient alternative to parametric bootstrap for $\hat{b}$, and can be applied in any group inference approach that utilizes some function of the block lasso. Estimator augmentation is especially useful in determining the significance in a hypothesis test and approximating $[\hat{b} \mid \hat{\beta} \in B]$ for some event B. In both scenarios, we need to sample from a rare event, which is known to be difficult and sometimes impossible for the bootstrap. We will demonstrate such applications with two

group inference approaches, one using the group lasso and the other a de-biased group lasso.

Estimator augmentation [29] was first developed for the lasso, which does not respect any group structure. Generalizing the method to the block lasso for all block norms ($\alpha \in (1, \infty]$) turns out to be very challenging technically. The sample space of the augmented estimator $(\hat{\beta}, S)$, which can be represented by a collection of manifolds with nonzero curvature, becomes more complicated. The joint distribution is thus defined over a curved space, a significant distinction from the augmented lasso estimator. Along the development, we also identify a set of sufficient conditions for solution uniqueness for the block lasso, which are weaker and more transparent than known results. In fact, the method of estimator augmentation applies to a large class of regularized estimators. To illustrate this view and promote its practical use, we further apply our method to find the joint density of an augmented scaled block lasso $(\hat{\beta}, S, \hat{\sigma})$, which has a coherent estimator $\hat{\sigma}^2$ for the error variance. When we were finalizing the first version of this paper, Tian Harris et al. posted a preprint [19], in which they generalize the technique of estimator augmentation to derive densities for selective sampling in a randomized convex learning program. This exemplifies that estimator augmentation may have much wider applications than what has been considered in our paper.

In addition to the above theoretical contributions, the significance of this work is also seen from its application in group inference, especially when the group size p_j is large. Mitra and Zhang [12] prove that de-biasing a scaled group lasso can achieve an efficiency gain in group inference by a factor of $\sqrt{p_j}$ over a de-biased lasso. Zhou and Min [30] show that bootstrap inference with the group lasso can reach an optimal rate of $n^{-1/2}$ if $\log J = O(p_j)$, which never holds for the lasso ($p_j = 1$) in the high-dimensional setting $p \gg n$. These results demonstrate the benefit of group sparsity in making inference about a group of parameters. Our development of estimator augmentation for the block lasso enables efficient simulation from the sampling distributions of the group lasso and the de-biased group lasso, which is an essential component in practical applications of these inferential approaches.

Notation used in this paper is defined as follows. Let $A \subset \mathbb{N}_p$ be an index set. For a vector $v = (v_j)_{1:p}$, we define $v_A = (v_j)_{j \in A}$. For a matrix $M = (M_{ij})_{n \times p}$, write its columns as M_j, $j = 1, \ldots, p$, and define $M_A = (M_j)_{j \in A}$ as a matrix of size $n \times |A|$ consisting of columns in A. Similarly, we define $M_{BA} = (M_{ij})_{i \in B, j \in A}$ and $M_{B\bullet} = (M_{ij})_{i \in B}$ for $B \subset \mathbb{N}_n$. Given the group structure $\mathcal{G}$, let $\mathcal{G}_A = \cup_{j \in A} \mathcal{G}_j \subset \mathbb{N}_p$ for $A \subset \mathbb{N}_J$. Define $v_{(A)} = v_{\mathcal{G}_A}$ with the special case $v_{(j)} = v_{\mathcal{G}_j}$ and let $G(v) = \{j \in \mathbb{N}_J : v_{(j)} \neq 0\}$ be the active groups of v. For an $n \times p$ matrix M, $M_{(A)} = M_{\mathcal{G}_A}$, and for a $p \times p$ matrix M, $M_{(AB)} = M_{\mathcal{G}_A \mathcal{G}_B}$, where $A, B \subset \mathbb{N}_J$. Denote by M^+ the Moore-Penrose pseudo-inverse of a matrix M so $M^+ = (M^\mathsf{T} M)^+ M^\mathsf{T}$ when M is not a square matrix. We use $\mathrm{row}(M)$ and $\mathrm{null}(M)$ to denote the row space and the null space of M, respectively. Let $\mathrm{diag}(M, M')$ be the block-diagonal matrix with M and M' as the diagonal blocks. Denote by $\phi_n(\bullet; c)$ the density of $\mathcal{N}_n(0, c\mathbf{I}_n)$ for $c > 0$. Let

$\mathbb{S}_\alpha^{m-1} = \{v \in \mathbb{R}^m : \|v\|_\alpha = 1\}$ be the unit ℓ_α-sphere in $\mathbb{R}^m$. We may suppress $(m-1)$ and simply write $\mathbb{S}_\alpha$ when the dimension does not need to be specified explicitly.

Throughout the paper, let α^* be conjugate to α in the sense that $\frac{1}{\alpha} + \frac{1}{\alpha^*} = 1$. We will assume that $\alpha \in (1, \infty)$ unless noted otherwise in next three sections, and leave to Section 5.1 the special case $\alpha = \infty$ whose technical details are slightly more complicated. Although not the focus of this paper, the results for the lasso can be obtained as another special case ($\alpha = 1$) after some simple modifications of the results for $\alpha \in (1, \infty)$. The block norm reduces to the usual ℓ_1 norm when $\alpha = 1$, effectively ignoring the block structure, and thus in this case we always assume $p_j = 1$ for all $j \in \mathbb{N}_J$ without loss of generality.

2. The basic idea

In this section, we give an overview of the idea of estimator augmentation. We start with the Karush-Kuhn-Tucker (KKT) conditions for the minimization problem (1.2). Under uniqueness of the block lasso, we will establish a bijection between y and the augmented estimator and derive the joint density of its sampling distribution. We note that solution uniqueness for the block lasso is an interesting topic in its own right, and the sufficient conditions in this work are much more transparent than those in the existing literature.

2.1. The KKT conditions

Denote by $\mathrm{sgn}(\cdot)$ the sign function with the convention that $\mathrm{sgn}(0) = 0$. For a scalar function $f : \mathbb{R} \to \mathbb{R}$ and a vector $v = (v_i) \in \mathbb{R}^m$, we define

$$f(v) := (f(v_1), \ldots, f(v_m)). \tag{2.1}$$

Definition 1. For $\alpha \in (1, \infty)$, let $\rho = \alpha^*/\alpha \in (0, \infty)$ and define $\eta : [-1, 1] \to [-1, 1]$ by

$$\eta(x) = \eta(x; \rho) = \mathrm{sgn}(x)|x|^\rho.$$

Denote its inverse function by $\eta^{-1}(x) = \mathrm{sgn}(x)|x|^{1/\rho}$.

Some basic properties about η are given in Lemma A.1 in Appendix A. In particular, $\eta(v)$ for $v \in \mathbb{S}_{\alpha^*}$, interpreted in the sense of (2.1), is a bijection from $\mathbb{S}_{\alpha^*}$ onto $\mathbb{S}_\alpha$. This fact is used in (2.2) below.

Let $S = (S_1, \ldots, S_p) \in \mathbb{R}^p$ such that $S_{(j)} \in \mathbb{R}^{p_j}$ is a subgradient of $\|\beta_{(j)}\|_\alpha$ evaluated at the solution $\hat{\beta}_{(j)}$ of (1.2). According to Lemma A.2 on the subdifferential of $\|\cdot\|_\alpha$, we have

$$\begin{cases} S_{(j)} = \eta^{-1}(\hat{\beta}_{(j)}/\|\hat{\beta}_{(j)}\|_\alpha) \in \mathbb{S}_{\alpha^*}^{p_j - 1} & \text{if } \hat{\beta}_{(j)} \neq 0, \\ \|S_{(j)}\|_{\alpha^*} \leq 1 & \text{if } \hat{\beta}_{(j)} = 0. \end{cases} \tag{2.2}$$

For the case $\alpha = \alpha^* = 2$ (group lasso), $\eta(v) = \eta^{-1}(v) = v$ and the above subgradient reduces to the familiar result in [26]. Put $W = \mathrm{diag}(w_1 \mathbf{I}_{p_1}, \ldots, w_J \mathbf{I}_{p_J})$,

which is a $p \times p$ matrix. The KKT conditions of (1.2), which are both sufficient and necessary, are

$$\frac{1}{n} X^{\mathsf{T}} X \hat{\beta} + \lambda W S = \frac{1}{n} X^{\mathsf{T}} y \tag{2.3}$$

for a vector S satisfying (2.2).

Definition 2. Let S be defined by (2.2) and (2.3). We will call $(\hat{\beta}, S) \in \mathbb{R}^{2p}$ an augmented solution to the block lasso problem (1.2). When we study the sampling distribution of $\hat{\beta}$, the random vector $(\hat{\beta}, S)$ will be called an *augmented estimator*.

If $(\hat{\beta}, S)$ is unique for each y, then (2.3) defines a bijective mapping from the space of $(\hat{\beta}, S)$ onto the space of y, which is the inverse of the minimization program (1.2) that maps y to $(\hat{\beta}, S)$. From the density of y or ε, it is hopeful to derive the joint density of the augmented estimator $(\hat{\beta}, S)$ via this bijective mapping. Then one may apply Monte Carlo methods, such as Markov chain Monte Carlo (MCMC) and importance sampling, to draw from the joint distribution of the augmented estimator. As a marginal distribution, the sampling distribution of $\hat{\beta}$ can be readily approximated by Monte Carlo samples, as well as any function of $(\hat{\beta}, S)$. This is the basic idea of estimator augmentation. Although the idea seems intuitive, there are a few technical difficulties in the implementation:

1. To establish the uniqueness of $(\hat{\beta}, S)$ under fairly general situations.
2. To characterize the sample space for $(\hat{\beta}, S)$, which appears to be a $2p$-vector but in fact lives in the union of a finite number of n-dimensional manifolds. This makes the aforementioned bijection conceivable since $\varepsilon \in \mathbb{R}^n$.
3. To calculate the Jacobian of the mapping and obtain the target density via a change of variable.

We will establish the solution uniqueness in the remainder of this section, and take care of the other two major steps in Section 3. Although the basic idea follows from that in [29], there are substantial new technical issues in each of the three steps, which will be discussed in the sequel.

2.2. Uniqueness

We briefly present here the most relevant results about solution uniqueness for the block lasso, while leaving many useful intermediate results and proofs to Appendix A. Despite that the KKT conditions only require the existence of a subgradient, it turns out that S is always unique:

Lemma 2.1. *For any y, X, $\lambda > 0$, and $\alpha \in [1, \infty]$, every $\hat{\beta}$ (1.2) gives the same fitted value $X\hat{\beta}$ and the same subgradient S.*

This lemma covers the full domain of α, including the boundary cases $\alpha = 1$ (lasso) and $\alpha = \infty$. Hereafter, we call S *the* subgradient vector due to its uniqueness. Next, we establish the uniqueness of $\hat{\beta}$ and thus the uniqueness

of the augmented solution $(\hat{\beta}, S)$. The lasso solution is unique if the columns of X are in general position [21], which says that the affine span of the set $\{s_j X_j : s_j \in \{-1, 1\}, j \in K \subset \mathbb{N}_p\}$ for every $|K| \leq n \wedge p$ does not contain any $\pm X_i$ for $i \notin K$. We generalize this definition to establish the uniqueness of the block lasso.

Definition 3. We say that the columns of a matrix $M \in \mathbb{R}^{n \times p}$ is in blockwise general position with respect to $(\mathcal{G}, \alpha)$ if for all $s \in \operatorname{row}(M)$, the vectors $\{M_{(j)}\eta(s_{(j)}) : j \in \mathcal{E}\}$ are in general position, where $\mathcal{E} = \{j \in \mathbb{N}_J : \|s_{(j)}\|_{\alpha^*} = 1\}$.

Let $U = \frac{1}{n} X^{\mathsf{T}} \varepsilon \in \mathbb{R}^p$, and denote the Gram matrix by $\Psi = \frac{1}{n} X^{\mathsf{T}} X$ hereafter. The KKT conditions (2.3) can be written as

$$\Psi\hat{\beta} + \lambda W S - \Psi\beta_0 = U. \tag{2.4}$$

Since $U \in \operatorname{row}(X)$ and $\Psi(\hat{\beta} - \beta_0) \in \operatorname{row}(X)$, we have

$$WS \in \operatorname{row}(X) \Leftrightarrow S \in \operatorname{row}(XW^{-1}) := \mathcal{V} \subset \mathbb{R}^p. \tag{2.5}$$

The following assumptions are sufficient for the main results of this work.

Assumption 1. Every $(n \wedge p)$ columns of X are linearly independent.

Assumption 2. The columns of XW^{-1} are in blockwise general position with respect to $(\mathcal{G}, \alpha)$.

The two assumptions are quite weak. Assumption 1 simply states that X does not satisfy any additional linear constraint other than those that must be satisfied by any $n \times p$ matrix. If the entries of X are drawn from a continuous distribution, then Assumption 1 holds with probability one. To help understand the intuition behind Assumption 2, choose $W = \mathbf{I}_p$ to simplify the exposition. Then this assumption is imposed on the vectors $X_{(j)}v_{(j)}$, where $v_{(j)} = \eta(s_{(j)}) \in \mathbb{S}_\alpha$ (Lemma A.1) and $s \in \mathcal{V}$. Under Assumption 1 with $n \leq p$, $\dim(\mathcal{V}) = n$ and $v = \eta(s) \in \mathbb{R}^p$ has only n free coordinates. Thus, we essentially require linear combinations of disjoint subsets of any n columns of X be in general position, which is a mild condition in practice. For the special case of the lasso with $p_j = 1$, Assumption 2 reduces to that the columns of X are in general position.

Theorem 2.2. *Suppose that Assumption 2 holds. Then for any $\lambda > 0$ and $y \in \mathbb{R}^n$, the solution $\hat{\beta}$ to the block lasso problem (1.2) with $\alpha \in [1, \infty)$ is unique and $|G(\hat{\beta})| \leq n \wedge J$.*

Since solution uniqueness is a topic of independent interest, we make a brief comparison to some existing results. Theorem 2.2 unifies a few important special cases, including the lasso ($\alpha = 1$) and the group lasso ($\alpha = 2$). For $\alpha = 1$, this theorem is comparable to the result in [21], while the existing results about the uniqueness of the group lasso involve conditions that are much less transparent than the one stated here. As an example, Theorem 3 in [17] states that, under Assumption 1, the group lasso solution $\hat{\beta}$ (with $\alpha = 2$) is unique if (i) $|\mathcal{G}_A| \leq n$, where $A = G(\hat{\beta})$ is the active groups, and (ii) $A = \{j \in \mathbb{N}_J : \|S_{(j)}\| = 1\}$.

Unlike Assumption 2 which is imposed on X explicitly, conditions (i) and (ii) are implicit in nature and can be verified only after a particular solution is calculated. According to Theorem 2.2, it is possible to have a unique solution when $|\mathcal{G}_A| > n$ as long as $|A| \leq n$, i.e., there are no more than n active groups but the total number of active coefficients of $\hat{\beta}$ could be greater than n. Such cases are not covered by the result in [17]. As will become clear in next section, the set of $\hat{\beta}$ satisfying (i) and (ii) is a proper subset of the full space of unique solutions and thus, in general, will have a probability mass strictly less than one.

3. Estimator augmentation

We will go through the main steps in detail to derive the joint density of the augmented estimator $(\hat{\beta}, S)$, which is useful for understanding this method. Section 3.1 characterizes the sample space of $(\hat{\beta}, S)$, Section 3.2 defines explicitly the bijective mapping from the KKT conditions, and Section 3.3 derives the joint density. A few concrete examples will follow in Section 3.4 to illustrate the method. The joint density of $(\hat{\beta}, S)$ depends on the true parameter β_0 and the error distribution. We will discuss in Section 4 how to apply estimator augmentation in high-dimensional inference. By default, we assume $p \geq n$. The results for $p < n$ will be obtained as special cases.

3.1. Sample space

Denote by $\hat{\gamma} = (\hat{\gamma}_j) \in \mathbb{R}^J$ the vector of the norms of $\hat{\beta}_{(j)}$, i.e. $\hat{\gamma}_j = \|\hat{\beta}_{(j)}\|_\alpha$. It follows from (2.2) that $\hat{\beta}_{(j)} = \hat{\gamma}_j \eta(S_{(j)})$ for all $j \in \mathbb{N}_J$. Thus, the augmented estimator $(\hat{\beta}, S)$ can be represented by the triplet $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$, where $\mathcal{A} = G(\hat{\beta})$ is a random subset of $\mathbb{N}_J$ when considering the sampling distribution. Given $\mathcal{A} = A$ for a fixed subset $A \subset \mathbb{N}_J$, it is seen from (2.2) and (2.5) that the sample space for S is

$$\mathcal{M}_A = \left\{ s \in \mathcal{V} : \|s_{(j)}\|_{\alpha^*} = 1 \ \forall j \in A \text{ and } \|s_{(j)}\|_{\alpha^*} \leq 1 \ \forall j \notin A \right\}. \tag{3.1}$$

Since $s_{(j)} \in \mathbb{S}_{\alpha^*}^{p_j - 1}$ for $j \in A$ and $\dim(\mathcal{V}) = n$ under Assumption 1, $\mathcal{M}_A$ is an $(n - |A|)$-manifold in $\mathbb{R}^p$ if $|A| \leq n$: It is the product of unit ℓ_{α^*}-spheres and balls intersecting with the linear subspace $\mathcal{V}$. Correspondingly, the space for $(\hat{\gamma}_A, S)$ given A is $\Omega_A = (\mathbb{R}^+)^{|A|} \times \mathcal{M}_A$, which is an n-manifold. Taking union over subsets of size $\leq n$, we obtain the sample space for the augmented estimator $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$:

$$\Omega = \bigcup_{|A| \leq n} \Omega_A \times \{A\}. \tag{3.2}$$

Remark 1. We do not have to consider $\{|\mathcal{A}| > n\}$, since this never happens under the assumptions of Theorem 2.2. Hereafter, we always regard the essential range of $\mathcal{A}$ as

$$\mathscr{A} := \{A \subset \mathbb{N}_J : |A| \leq n\}. \tag{3.3}$$

In summary, the sample space of the augmented estimator, represented by the triplet $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$, is the union of a finite number of n-manifolds. Thus, it is possible to find a bijective mapping from this space to $\mathbb{R}^n$, the space for ε; see the mapping $\tilde{H}$ to be defined in (3.10). For the lasso, Ω_A degenerates to a union of n-dimensional polyhedra with zero curvature.

Remark 2. Parameterizing the augmented estimator in terms of $\hat{\gamma}$ and S is a critical choice for our derivations. In this way, all the equality constraints are imposed on S as in (3.1), leading to familiar geometry for the spaces of $\hat{\gamma}$ and S, which is helpful for understanding distributions over these spaces. It is also a nature choice, since the subgradient S is alway unique (Lemma 2.1) and non-uniqueness comes solely from $\hat{\gamma}$ (Lemma A.4).

3.2. A bijective mapping

Putting $\hat{\beta}_{(j)} = \hat{\gamma}_j \eta(S_{(j)})$ for $j \in \mathcal{A}$, Equation (2.4) becomes

$$\frac{1}{n} X^{\mathsf{T}} \varepsilon = \sum_{j \in \mathcal{A}} \hat{\gamma}_j \Psi_{(j)} \eta(S_{(j)}) + \lambda W S - \Psi \beta_0 := H(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A}; \beta_0, \lambda), \tag{3.4}$$

which defines a mapping $H : \Omega \to \mathrm{row}(X)$ for any $\beta_0 \in \mathbb{R}^p$ and $\lambda > 0$. For notational brevity, we often suppress its dependence on (β_0, λ) and write the mapping as $H(\bullet)$. In what follows, we show that H is bijective, which is a consequence of the uniqueness of $(\hat{\beta}, S)$, or equivalently of $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$.

Lemma 3.1. *Suppose that $\alpha \in [1, \infty)$ and Assumption 2 holds. Then for any $\beta_0 \in \mathbb{R}^p$ and $\lambda > 0$, H is a bijection that maps Ω onto* $\mathrm{row}(X)$.

This lemma applies to $\alpha = 1$, in which case we define $\eta(x) = xI(|x| = 1)$ and $\eta^{-1}(x) = \mathrm{sgn}(x)$ by taking the limit $\rho \to \infty$ in Definition 1.

The mapping H is established at a quite abstract level so far. It will be more convenient to work with the restriction of H to Ω_A for $A \in \mathscr{A}$, defined by

$$H_A(r_A, s) := H(r_A, s, A) \quad \text{for } (r_A, s) \in \Omega_A, \tag{3.5}$$

where $r = (r_1, \ldots, r_J) \in \mathbb{R}^J$ with $r_j = 0$ for $j \notin A$. Then H can be understood as a collection of one-to-one mappings $\{H_A : A \in \mathscr{A}\}$ indexed by subsets of $\mathbb{N}_J$. Write the block lasso solution for the response y as $\hat{\beta} = \hat{\beta}(y)$. Let

$$E_A := \left\{ v \in \mathbb{R}^n : G\left(\hat{\beta}(X\beta_0 + v)\right) = A \right\}$$

be the set of noise vectors v for which the active set of the block lasso solution $\hat{\beta}(X\beta_0 + v)$ is A. Denote the block norms and the subgradient of $\hat{\beta}(X\beta_0 + v)$ by $\hat{\gamma}(X\beta_0 + v)$ and $S(X\beta_0 + v)$, respectively. Then for $v \in E_A$, we have

$$H_A(\hat{\gamma}_A(X\beta_0 + v), S(X\beta_0 + v)) = \frac{1}{n} X^{\mathsf{T}} v.$$

The one-to-one mapping H_A allows us to obtain the density for $(\hat{\gamma}_A, S)$ from the density of the noise vector via a change of variable.

It remains to find the differential of H_A so that we can calculate the Jacobian for the change of variable. A special aspect of this mapping is that H_A is defined on a manifold and thus its differential is determined with respect to local parameterizations. As an $(n-|A|)$-manifold in $\mathbb{R}^p$, a neighborhood of $s \in \mathcal{M}_A$ can be parameterized by s_F, where $F \subset \mathbb{N}_p$ may depend on (s, A) and $|F| = n-|A|$. Correspondingly, the n-manifold Ω_A will be parameterized by $(r_A, s_F) \in \mathbb{R}^n$ in a neighborhood of (r_A, s). Under this parameterization, Lemma 3.2, proven in Appendix B.1, gives an expression of dH_A in terms of a few matrices defined below. Let $\eta' : [-1, 1] \to [0, \infty]$ be the derivative of η so that $\eta'(x) = \rho|x|^{\rho-1}$. Define

$$r \circ \Psi := [r_1\Psi_{(1)}| \dots |r_J\Psi_{(J)}] \in \mathbb{R}^{p\times p}, \tag{3.6}$$

$$\Psi \circ \eta := [\Psi_{(1)}\eta(s_{(1)})| \dots |\Psi_{(J)}\eta(s_{(J)})] \in \mathbb{R}^{p\times J}, \tag{3.7}$$

and $D = D(s, A) \in \mathbb{R}^{p\times p}$ to be a diagonal matrix whose diagonal elements $D_{kk} = \eta'(s_k)$ for $k \in \mathcal{G}_A$ and $D_{kk} = 0$ otherwise.

Lemma 3.2. *Fix $p \geq n$, $\beta_0 \in \mathbb{R}^p$, $\lambda > 0$ and $A \in \mathscr{A}$. Suppose that $\alpha \in (1, \infty)$ and Assumption 1 holds. Then for any interior point $(r_A, s) \in \Omega_A$, there is a full rank matrix $T = T(s, A)$ of size $p \times (n-|A|)$ such that $ds = T(s, A)ds_F$ and*

$$dH_A = [(\Psi \circ \eta)_A \mid \{(r \circ \Psi)D + \lambda W\}T(s, A)]\, d\theta := M(r_A, s, A; \lambda)d\theta, \tag{3.8}$$

where $\theta = (r_A, s_F) \in \mathbb{R}^n$ and $r_j = 0$ for $j \notin A$.

Remark 3. This lemma applies to every interior point of Ω_A, irrespective of whether or not the corresponding solution is unique. The size of the matrix $M = M(r_A, s, A; \lambda)$ is $p \times n$. Assumption 1 is only needed to fix the dimension of the manifold $\mathcal{M}_A$. With some modifications of the proof, the result can be generalized to the situation in which Assumption 1 fails to hold. The parameterization s_F for $\mathcal{M}_A$ is defined locally for a neighborhood of s. For each $j \in A$, the unit sphere $\mathbb{S}_{\alpha^*}^{p_j-1}$, except a set of measure zero, can be covered by two parameterizations, one for each open semi-sphere.

Remark 4. For the special case $\alpha = \alpha^* = 2$ (group lasso), we have $\rho = 1$, $\eta(x) = x$ and $\eta'(x) = 1$ for $x \in [-1, 1]$. The matrix M (3.8) has a simpler form:

$$M(r_A, s, A; \lambda) = [(\Psi \circ s)_A \mid (r \circ \Psi + \lambda W)T(s, A)]\,. \tag{3.9}$$

The only reason that we excluded the case $\alpha = 1$ in the above lemma is because η' is not well-defined. We will cover this case, which reduces to the lasso, in Example 3.

Geometrically, the columns of T consist of a set of tangent vectors of the manifold $\mathcal{M}_A$ while those of M consist of tangent vectors of the mapping H_A. These tangent vectors determine the ratio between the volume element in the image space $\mathrm{row}(X)$ and that in the domain Ω_A, and thus the Jacobian of the mapping.

3.3. Joint density

Now we make an explicit link from the augmented estimator $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ to the noise vector ε. Under Assumption 1, $\text{null}(X^{\mathsf{T}}) = \{0\}$ and thus by (3.4)

$$\varepsilon/\sqrt{n} = \sqrt{n}(X^{\mathsf{T}})^{+} H(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A}; \beta_0, \lambda) := \tilde{H}(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A}; \beta_0, \lambda). \tag{3.10}$$

We note that $\tilde{H} \in \mathbb{R}^n$ is the coordinates of H with respect to the basis $(X^{\mathsf{T}}/\sqrt{n})$ of $\text{row}(X)$. By Lemma 3.1, $\tilde{H}$ is a bijection that maps Ω onto $\mathbb{R}^n$. Define $\tilde{H}_A$ similarly as for H_A in (3.5). It then follows from Lemma 3.2 that the Jacobian of $\tilde{H}_A$ is

$$J_A(r_A, s; \lambda) = \det\left[\sqrt{n}(X^{\mathsf{T}})^{+} M(r_A, s, A; \lambda)\right], \tag{3.11}$$

of which the matrix on the right side is of size $n \times n$.

Theorem 3.3. *Fix $p \geq n$, $\beta_0 \in \mathbb{R}^p$ and $\lambda > 0$. Suppose that Assumptions 1 and 2 hold. Then the distribution of the augmented estimator $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ for $\alpha \in (1, \infty)$ is given by the differential form*

$$d\mu_A := \mathbb{P}(dr_A, ds, \{A\}) = g_n(\tilde{H}_A(r_A, s; \beta_0, \lambda))|J_A(r_A, s; \lambda)|d\theta \tag{3.12}$$

for $(r_A, s, A) \in \Omega$, where $\theta = (r_A, s_F) \in \mathbb{R}^n$ and g_n is the density of $(\varepsilon/\sqrt{n})$.

See Appendix B.2 for a proof, from which we see that (3.12) is valid as long as the block lasso program (1.2) has a unique solution for almost all $y \in \mathbb{R}^n$. For each $A \in \mathscr{A}$, the n-form $d\mu_A$ defines a measure on Ω_A in the following sense. Let $k = n - |A|$ and

$$f_A(r_A, s) = g_n(\tilde{H}_A(r_A, s))|J_A(r_A, s)|. \tag{3.13}$$

Suppose that $\Gamma \subset (\mathbb{R}^{+})^{|A|}$ and $\Phi = \{\Phi(u) : u \in \Delta\} \subset \mathcal{M}_A$ is a k-surface in $\mathbb{R}^p$ with parameter domain $\Delta \subset \mathbb{R}^k$. Then by (3.12) we have

$$\mathbb{P}(\hat{\gamma}_A \in \Gamma, S \in \Phi, \mathcal{A} = A) = \int_{\Gamma \times \Phi} d\mu_A = \int_{\Gamma} \int_{\Delta} f_A(r_A, \Phi(u)) \left|\frac{\partial s_F}{\partial u}\right| du\, dr_A, \tag{3.14}$$

where the Jacobian $\partial s_F/\partial u = 1$ if Φ is parameterized by s_F. Note that for a particular k-surface, parameterizations other than by s_F may be more convenient. As shown in (3.14), the distribution of $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ is defined by a collection of measures, $\{\mu_A : A \in \mathscr{A}\}$, due to the discrete nature of $\mathcal{A}$, and f_A is the density of μ_A parameterized by $\theta = (r_A, s_F)$. It is possible that $f_A = \infty$ on a set of measure zero in Ω_A. An important special case of the above integral is

$$\mathbb{P}(\mathcal{A} = A) = \mu_A(\Omega_A) = \int_{\Omega_A} d\mu_A.$$

Lastly, summing over $\mathscr{A}$ in the above equation leads to

$$\sum_{A \in \mathscr{A}} \int_{\Omega_A} d\mu_A = \sum_{|A| \leq n} \mathbb{P}(\mathcal{A} = A) = 1.$$

Remark 5. Evaluation of the joint density in (3.12) for any $(r_A, s, A) \in \Omega$ can be done by a simple procedure:

1. Find a local parameterization s_F for s and the associated matrix T;
2. Calculate the Jacobian J_A (3.11), evaluate the mapping $\tilde{H}_A(r_A, s)$ (3.10), and plug them into (3.13) to obtain $f_A(r_A, s)$.

See Examples 1 and 2 for concrete illustrations.

As a consequence of Theorem 3.3, the density f_A under i.i.d. Gaussian errors can be found in the following corollary. See Appendix B.3 for a proof.

Corollary 3.4. *Suppose that the assumptions of Theorem 3.3 hold. If* $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$, *then for* $(r_A, s) \in \Omega_A$ *and* $A \in \mathscr{A}$,

$$f_A(r_A, s) = \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|X(b + \lambda\Psi^+ W s - \beta_0)\|^2\right] |J_A(r_A, s)|, \tag{3.15}$$

where $b \in \mathbb{R}^p$ *is such that* $b_{(j)} = r_j \eta(s_{(j)})$ *for all* $j \in A$ *and* $b_{(j)} = 0$ *otherwise.*

As we have seen, the sample space Ω (3.2) for the augmented estimator is complex due to the many constraints involved in $\mathcal{M}_A$ (3.1) and the mix of continuous and discrete components. It is quite surprising that one can find an exact joint density for the augmented estimator given β_0 and the noise distribution, which is usually simple under an i.i.d. assumption. The density gives a complete and explicit characterization of the sampling distribution according to (3.14). In light of the non-linear and sparse nature of $\hat{\beta}$ and the high dimension of the problem, the joint density itself is a significant theoretical result that improves our understanding of the block lasso estimator. Applications of this result in group inference will be discussed in Section 4.

Remark 6. We summarize the main differences between the joint density of the augmented block lasso in Theorem 3.3 and that of the augmented lasso in [29]. First, the sample space Ω_A is an n-manifold with nonzero curvature for $\alpha > 1$, and consequently the density is specified in (3.12) by a differential form of order n. In contrast, the space Ω_A has no curvature for the augmented lasso estimator, whose density can be defined with respect to the Lebesgue measure. Second, the Jacobian (3.11) depends on r_A, s and A for the block lasso, while it only depends on A for the lasso. See Example 3 for the technical reason and a geometric interpretation for these differences. Both aspects result in new challenging computational issues in this work for the development of Monte Carlo algorithms, which are discussed in Section 4.2.

For the sake of completeness, we also give the density of $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ when $p < n$, which can be obtained by simple modifications of a few steps in the proof of the result for $p \geq n$. See Appendix B.4 for details. Assume that $\operatorname{rank}(X) = p < n$, which is sufficient for both Assumptions 1 and 2 to hold. Then $\operatorname{row}(X)$ and $\mathcal{V}$ (2.5) are identical to $\mathbb{R}^p$, which implies every $s \in \mathcal{M}_A$ (3.1) can be locally

parameterized by s_F with $|F| = p - |A|$. In this case, H_A maps Ω_A into $\mathbb{R}^p$ and $M(r_A, s, A; \lambda) \in \mathbb{R}^{p \times p}$.

Corollary 3.5. *Fix $n > p$, $\beta_0 \in \mathbb{R}^p$ and $\lambda > 0$. If $\operatorname{rank}(X) = p$, the distribution of the augmented estimator $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ for $\alpha \in (1, \infty)$ is given by the p-form*

$$d\mu_A = g_{n,X}(H_A(r_A, s; \beta_0, \lambda))|\det M(r_A, s, A; \lambda)|d\theta \tag{3.16}$$

for $(r_A, s, A) \in \Omega$, where $\theta = (r_A, s_F) \in \mathbb{R}^p$ and $g_{n,X}$ is the density of $U = n^{-1}X^{\mathsf{T}}\varepsilon$.

An interesting observation is that we need the density $g_{n,X}$ of $U = n^{-1}X^{\mathsf{T}}\varepsilon$ when $p < n$, which is more difficult to determine than the density of ε needed in the high-dimensional case (3.12). The underlying reason for this can be found from the sufficient statistic $t = X^{\mathsf{T}}y$ (2.3). When $p < n$, the dimension of t is smaller than the sample size n and thus this statistic achieves the goal of data reduction. Consequently, one needs the distribution of t or U to determine the sampling distribution of $\hat{\beta}$. However, when $p \geq n$, y is the coordinates of the statistic t using the rows of X as the basis and thus the two are equivalent up to a change of basis, in which case the distribution of y or ε is all we need.

3.4. Examples

We illustrate the distribution of the augmented estimator with a few examples. Example 1 is a simple concrete example that demonstrates various key concepts, including the sample space, the density, and probability calculations. The second example shows that, under an orthonormal design, the joint distribution given by Theorem 3.3 coincides with the result from block soft-thresholding. The last example considers the lasso. Technical details involved in these examples are deferred to Appendix C.

Example 1. Consider a simple but nontrivial example with $p = 3$, $n = 2$, and $J = 2$. The two groups $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{3\}$, and pick $\alpha = 2$. Suppose that

$$\frac{1}{\sqrt{n}}X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \qquad \beta_0 = 0, \qquad \varepsilon \sim \mathcal{N}_2(0, \sigma^2\mathbf{I}_2), \qquad W = \mathbf{I}_3.$$

Put $r = (r_1, r_2)$ and $s = (s_1, s_2, s_3)$.

We first determine the space $\mathcal{M}_A$ (3.1). Incorporating the constraint that

$$s \in \mathcal{V} = \operatorname{row}(X) \Leftrightarrow s_1 + s_2 - s_3 = 0, \tag{3.17}$$

the manifold $\mathcal{M}_A$ can be expressed as

$$\mathcal{M}_A = \{(s_1, s_2, s_1 + s_2) : (s_1, s_2) \in \mathbb{D}_A\}, \tag{3.18}$$

where $\mathbb{D}_A \subset \mathbb{R}^2$ is the range for $s_{(1)} = (s_1, s_2)$. Let $\mathbb{B}^m$ be the unit ℓ_2-ball in $\mathbb{R}^m$. For $A = \varnothing$, the definition of $\mathcal{M}_A$ shows that $\mathbb{D}_\varnothing = \mathbb{B}^2 \cap \{|s_1 + s_2| \leq 1\}$, whose boundary $\partial\mathbb{D}_\varnothing$ consists of two arcs and two line segments connecting at

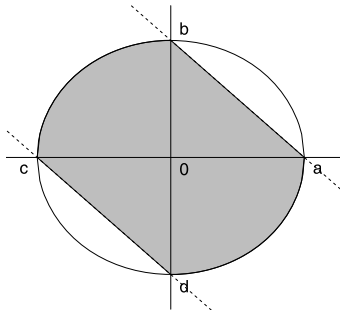

FIG 1. *Sample space of $S_{(1)}$ shown by the shaded area.*

four points: $a = (1, 0)$, $b = (0, 1)$, $c = (-1, 0)$, and $d = (0, -1)$. See Figure 1 for illustration. Use $\partial(q_1, q_2)$ to denote the boundary of $\mathbb{D}_\varnothing$ from q_1 to q_2 along the positive orientation. It is then immediate that

$$\mathbb{D}_A = \begin{cases} \partial(b,c) \cup \partial(d,a) & \text{for } A = \{1\}, \\ \partial(a,b) \cup \partial(c,d) & \text{for } A = \{2\}, \\ \{a,b,c,d\} & \text{for } A = \{1,2\}. \end{cases} \tag{3.19}$$

Plugging $\mathbb{D}_A$ back into (3.18), we see that $\mathcal{M}_\varnothing$ is a surface, $\mathcal{M}_{\{1\}}$ and $\mathcal{M}_{\{2\}}$ are curves, and $\mathcal{M}_{\{1,2\}}$ degenerates to four points in $\mathbb{R}^3$.

We find f_A (3.13) and calculate $\mathbb{P}(\mathcal{A} = A)$ for $A = \{1\}$ here. The two arcs in $\mathbb{D}_{\{1\}}$ can be parameterized by s_1 and $\Omega_{\{1\}}$ correspondingly by $\theta = (r_1, s_1)$ with two domains, $\mathbb{R}^+ \times (-1, 0)$ and $\mathbb{R}^+ \times (0, 1)$. After a few steps of algebra, we arrive at the density

$$f_{\{1\}}(r_1, s_1, s_2) = \frac{1}{\pi\sigma^2} \exp\left[-\frac{(r_1 + \lambda)^2}{\sigma^2}\right] \frac{r_1 + \lambda}{|s_2|}. \tag{3.20}$$

Integrating $f_{\{1\}} dr_1 ds_1$ over $\mathbb{R}^+ \times \mathbb{D}_{\{1\}}$ gives $\mathbb{P}(\mathcal{A} = \{1\}) = \frac{1}{2} e^{-\lambda^2/\sigma^2}$. In Appendix C.1, we provide the results for $A = \varnothing, \{2\}, \{1, 2\}$, and verify that $\mathbb{P}(\mathcal{A} = A)$ indeed sums up to one.

Example 2 (Orthogonal design)**.** Suppose $p = n = mJ$, $\Psi = \mathbf{I}_p$, and put $W = \sqrt{m}\mathbf{I}_p$. In this example, all the groups are of the same size m. It is known that under this setting, the group lasso ($\alpha = 2$) is obtained by block soft-thresholding the least-squares estimator $\tilde{\beta} = \frac{1}{n} X^\mathsf{T} y$:

$$\hat{\beta}_{(j)} = \tilde{\beta}_{(j)} \left[1 - \lambda\sqrt{m}/\|\tilde{\beta}_{(j)}\|\right]_+, \quad j = 1, \ldots, J. \tag{3.21}$$

Assume $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. Then $\tilde{\beta}_{(j)} \sim \mathcal{N}_m(\beta_{0(j)}, (\sigma^2/n)\mathbf{I}_m)$, $j \in \mathbb{N}_J$, are mutually independent. The distribution of $(\hat{\gamma}_\mathcal{A}, S, \mathcal{A})$ for $\alpha = 2$, derived in Appendix C.2, is given by

$$d\mu_A = \prod_{j \in \mathbb{N}_J} f_j(r_j, s_{(j)}) d\theta_{(j)}, \tag{3.22}$$

in which

$$f_j(r_j, s_{(j)}) = \left(\frac{2\pi\sigma^2}{n}\right)^{-\frac{m}{2}} \exp\left\{-\frac{n}{2\sigma^2}\|(r_j + \lambda\sqrt{m})s_{(j)} - \beta_{0(j)}\|^2\right\} \left|\det M_{(jj)}\right|,$$

$d\theta_{(j)} = dr_j ds_{F(j)}$ if $j \in A$ and $d\theta_{(j)} = ds_{(j)}$ (with $r_j = 0$) otherwise. Here, $F(j)$ is a chosen set of $(m-1)$ free coordinates of $s_{(j)}$, and $|\det M_{(jj)}|$ has a closed-form expression (C.10). In what follows, we exemplify that (3.22) is consistent with block soft-thresholding (3.21).

Since $d\mu_A$ factorizes into a product of J terms, different groups $(\hat{\gamma}_j, S_{(j)})$ are mutually independent. The density $f_j(r_j, s_{(j)})$ determines the distribution of $(\hat{\gamma}_j, S_{(j)})$. If $j \notin A$, letting $r_j = 0$ we have

$$f_j ds_{(j)} = (2\pi\sigma^2/n)^{-\frac{m}{2}} \exp\left[-\frac{n}{2\sigma^2}\|\lambda\sqrt{m}s_{(j)} - \beta_{0(j)}\|^2\right] (\lambda\sqrt{m})^m ds_{(j)}. \quad (3.23)$$

It then follows that

$$\begin{aligned}\mathbb{P}(\hat{\beta}_{(j)} = 0) &= \int_{\mathbb{B}^m} f_j ds_{(j)} = \int_{\|z\| \le \lambda\sqrt{m}} \phi_m(z; \beta_{0(j)}, \sigma^2 \mathbf{I}_m/n) dz \\ &= \mathbb{P}(\|\tilde{\beta}_{(j)}\| \le \lambda\sqrt{m}),\end{aligned}$$

where the last equality comes from the distribution of $\tilde{\beta}_{(j)}$. This result is clearly consistent with the soft-thresholding rule (3.21). Next we calculate $\mathbb{P}(\hat{\gamma}_j > t)$ for $j \in A$. To simplify our derivation, assume further that $\beta_{0(j)} = 0$. Integrating $f_j d\theta_{(j)}$ over the sphere $s_{(j)} \in \mathbb{S}^{m-1}$, the marginal density of $\hat{\gamma}_j$ is

$$f_j(r_j) = \frac{(n/\sigma^2)^{\frac{m}{2}}}{2^{\frac{m}{2}-1} \cdot \Gamma(m/2)} (r_j + \lambda\sqrt{m})^{m-1} \exp\left[-\frac{n}{2\sigma^2}(r_j + \lambda\sqrt{m})^2\right] \quad (3.24)$$

for $r_j > 0$. It then follows that, for $t \ge 0$,

$$\mathbb{P}(\|\hat{\beta}_{(j)}\| > t) = \int_t^\infty f_j(r_j) dr_j = \mathbb{P}\left\{\|\tilde{\beta}_{(j)}\| > t + \lambda\sqrt{m}\right\}, \quad (3.25)$$

which again coincides with the result from soft-thresholding. See Appendix C.2 for the derivation of (3.24) and (3.25).

Example 3 (lasso)**.** When $\alpha = 1$ in (1.2), the block lasso reduces to the lasso with no group structure. Thus, the result for $\alpha = 1$ can be deduced by letting $p_j = 1$ for all j and $\alpha = 2$ (or any $\alpha > 1$) in Theorem 3.3. In this case, for $j \in \mathcal{A}$ the subgradient $S_j = \operatorname{sgn}(\hat{\beta}_j) \in \{1, -1\}$ is a function of $\hat{\beta}_j$. This leads to two special properties of the matrix $T = T(s, A)$ defined in Lemma 3.2, which do not hold in the general case $p_j \ge 2$: (i) $T = T(A)$ depends only on A, (ii) the submatrix $T_{A\bullet}$ is a zero matrix; see Appendix C.3. Bearing these facts in mind, one can apply Theorem 3.3 to find the joint distribution of the augmented lasso, given by the density

$$f_A(r_A, s) dr_A ds_F = g_n(\tilde{H}_A(r_A, s)) \left|\det\left\{\sqrt{n}(X^\mathsf{T})^+ [\Psi_A \mid \lambda W_B T_{B\bullet}]\right\}\right| dr_A ds_F \quad (3.26)$$

for $(r_A, s) \in \Omega_A$, where $B = \mathbb{N}_p \setminus A$ and $F \subset B$. Owing to property (i), the Jacobian and the set F here do not depend on s, which is fundamentally different from the block lasso. A geometrical interpretation for (i) is that the space $\mathcal{M}_A$ (3.1) for the lasso is a union of polyhedra and the set of tangent vectors that forms the columns of T is invariant at each $s \in \mathcal{M}_A$, while in the block lasso case $\mathcal{M}_A$ is curved with a different tangent space at different points. This gives one of the aspects in which this work represents a highly nontrivial generalization to the result for the lasso.

As adopted in [29], the augmented lasso estimator can also be represented by $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{B}}, \mathcal{A})$, where $\mathcal{B} = \mathbb{N}_p \setminus \mathcal{A}$ is the set of zero components of $\hat{\beta}$. With the change of variable, $\hat{\beta}_j = \hat{\gamma}_j S_j$ for $j \in \mathcal{A}$, one can easily obtain the density under this alternative parameterization from (3.26), which is identical to the joint density in Theorem 2 of [29] with the choice of $(X^{\mathsf{T}}/\sqrt{n})$ as a basis for $\operatorname{row}(X)$. See Appendix C.3 for the technical details.

4. Applications in statistical inference

In this section, we develop Monte Carlo methods to make inference about β_0 by utilizing the joint density of the augmented block lasso estimator. Recall that we want to test the hypothesis $H_{0,G} : \beta_{0G} = 0$ or to construct confidence regions for β_{0G}. Without loss of generality, assume $G = \mathcal{G}_j$ for some j so that our goal is to infer $\beta_{0(j)}$. Denote the null hypotheses by $H_{0,j} : \beta_{0(j)} = 0$ for $j \in \mathbb{N}_J$.

4.1. Parametric bootstrap

Consider inference with an estimator in the form of $\hat{b} = \hat{b}(\hat{\beta}, S) \in \mathbb{R}^p$, a mapping of the augmented estimator $(\hat{\beta}, S)$. One such approach that has drawn recent attention is the de-biased lasso and its generalization to the de-biased group lasso. Given a $p \times p$ matrix $\hat{\Theta} = \hat{\Theta}(X)$, a form of the de-biased estimator may be expressed as

$$\hat{b} = \hat{\beta} + \hat{\Theta} X^{\mathsf{T}}(y - X\hat{\beta})/n = \hat{\beta} + \lambda \hat{\Theta} W S, \tag{4.1}$$

where $(\hat{\beta}, S)$ is either the augmented lasso or the augmented group lasso. Different de-biased estimators have been constructed with different $\hat{\Theta}$, which is often some version of a relaxed inverse of the Gram matrix Ψ. It is usually impossible to obtain the exact distribution of $(\hat{b} - \beta_0)$ for a finite sample. Thus, bootstrap methods have been developed [5, 28] with improved performance compared to asymptotic approximation for the de-biased methods.

Assuming the error distribution is $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ with a known σ^2 for now, a parametric bootstrap for the augmented estimator $(\hat{\beta}, S)$ contains two steps:

Algorithm 1 $(PB(\tilde{\beta}, \sigma^2, \lambda))$**.** Given $\sigma^2 > 0$, $\lambda > 0$ and a point estimate $\tilde{\beta} \in \mathbb{R}^p$,

(1) draw $\varepsilon^* \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ and set $y^* = X\tilde{\beta} + \varepsilon^*$;
(2) solve (1.2) with y^* in place of y to obtain $\hat{\beta}^*$ and calculate S^* via (2.3).

Let $\hat{b}^* = \hat{b}(\hat{\beta}^*, S^*)$. Choosing a function $h_j : \mathbb{R}^{p_j} \to [0, \infty)$, we estimate its $(1-\delta)$-quantile $h_{j,(1-\delta)}$ from a large bootstrap sample such that

$$\mathbb{P}\left\{ h_j(\hat{b}^*_{(j)} - \tilde{\beta}_{(j)}) > h_{j,(1-\delta)} \middle| \tilde{\beta} \right\} = \delta.$$

Then, a $(1-\delta)$ confidence region for $\beta_{0(j)}$ can be constructed in the form of

$$R_j(\delta) = \left\{ \theta \in \mathbb{R}^{p_j} : h_j(\hat{b}_{(j)} - \theta) \le h_{j,(1-\delta)} \right\}. \tag{4.2}$$

By duality the p-value for testing $H_{0,j}$ is approximated by the tail probability

$$\mathbb{P}\left\{ h_j(\hat{b}^*_{(j)} - \tilde{\beta}_{(j)}) \ge h_j(\hat{b}_{(j)}) \middle| \tilde{\beta} \right\}. \tag{4.3}$$

Common choices of h_j include, for example, various norms and $h_j(\theta) = \|X_{(j)}\theta\|$. Although out of the scope of this paper, the asymptotic validity of (4.2) and (4.3) comes from the fact that $(\hat{b}_{(j)} - \beta_{0(j)})$ is an asymptotic pivot with a careful choice of $\hat{\Theta}$ [12, 22].

An interesting and key observation is that the joint density of $[\hat{\beta}^*, S^* \mid \tilde{\beta}]$ is explicitly given by (3.12) in Theorem 3.3, with $\tilde{\beta}$ in place of β_0, through its equivalent representation. Denote this density (3.13) by $f_A(r_A, s; \tilde{\beta}, \sigma^2, \lambda)$ to emphasize its dependence on $(\tilde{\beta}, \sigma^2, \lambda)$. In principle, we can use Monte Carlo methods, such as importance sampling and MCMC, to draw $(\hat{\beta}^*, S^*)$ and obtain a sample of $\hat{b}^* = \hat{b}(\hat{\beta}^*, S^*)$, which serve as alternatives to the above bootstrap sampling. Monte Carlo methods may bring computational efficiency and flexibility compared to parametric bootstrap. In the following, we will demonstrate the efficiency of importance sampling in calculating tail probabilities as in (4.3), which is a prominent difficulty for the bootstrap. Monte Carlo methods for other applications, including those with an estimated error distribution, are discussed in Section 4.5.

4.2. Importance sampling

The following simple fact about the parameterization of $\mathcal{M}_A$ (3.1), proved in Appendix B.5, is useful for designing proposal distributions in importance sampling.

Lemma 4.1. *Let $\alpha \in (1, \infty)$. For each $A \in \mathscr{A}$, the manifold $\mathcal{M}_A$, except for a set of measure zero, can be parameterized by s_F such that the index set $F = F(A)$ only depends on A.*

A consequence of Lemma 4.1 is that we may use the same volume element $d\theta = dr_A ds_F$ almost everywhere in the subspace Ω_A, which eases our development of a Monte Carlo algorithm. Suppose that $q_A(r_A, s)$ is the density of a distribution over Ω with respect to $d\theta$ such that $\sum_A \int_{\Omega_A} q_A(r_A, s) d\theta = 1$. As long as the support of q_A is Ω_A for all $A \in \mathscr{A}$, it can be used as a proposal distribution in importance sampling. With a little abuse of notation, put

$\theta = (r_A, s) \in \Omega_A$ so that (θ, A) represents a point in the sample space Ω at which the volume element is $d\theta$. Suppose we want to estimate the expectation of a function $h(\hat{\beta}, S) = h(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ with respect to f_A, using $(\hat{\beta}, S)$ and $(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ interchangeably. Importance sampling can be readily implemented given the densities f_A and q_A. Draw (A_t, θ_t) from the proposal $q_A(\theta)$ for $t = 1, \ldots, N$ and calculate importance weights $w_t = f_{A_t}(\theta_t)/q_{A_t}(\theta_t)$. Then by the law of large numbers, the weighted sample mean

$$\hat{h} = \frac{\sum_{t=1}^{N} w_t h(\theta_t, A_t)}{\sum_{t=1}^{N} w_t} \xrightarrow{a.s.} \mathbb{E}[h(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})]$$

provides the desired estimate. To estimate the probability in (4.3), h is taken to be the indicator function of the event of interest. When the true $\beta_{0(j)} \neq 0$, the p-value (4.3) can be tiny, and bootstrap (Algorithm 1) may fail to provide a meaningful estimate of the significance level. In such cases, it is much more efficient to use importance sampling with a proposal distribution that has a higher chance to reach the tail of the bootstrap distribution $f_A(r_A, s; \tilde{\beta}, \sigma^2, \lambda)$.

We design two types of proposal distributions. The first type of proposals draw $(\hat{\beta}^*, S^*)$ by the bootstrap algorithm $PB(\beta^\dagger, M\sigma^2, \lambda^\dagger)$ with a proper choice of $(\beta^\dagger, M, \lambda^\dagger)$, where $M > 0$ is a constant. The proposal distribution has density $f_A(r_A, s; \beta^\dagger, M\sigma^2, \lambda^\dagger)$, again by Theorem 3.3. By increasing the error variance with $M > 1$, choosing $\beta^\dagger \neq \tilde{\beta}$, and possibly with a different $\lambda^\dagger$, we can propose samples in the region of interest in (4.3) which has a small probability with respect to the target distribution. The Jacobian term $J_A(r_A, s; \lambda)$ (3.11) is the time-consuming part in evaluating the densities for calculating importance weights. If we choose $\lambda^\dagger = \lambda$, however, this term will cancel out and the importance weight is simply the ratio of two normal densities, whose calculation is almost costless. Our empirical study shows that this choice gives comparable estimation accuracy and thus we always let $\lambda^\dagger = \lambda$. Denote by $IS(\beta^\dagger, M)$ the importance sampling with the first type of proposals. Our second design uses a mixture of two proposal distributions with different $\beta^\dagger$ and M, which has more flexibility in shifting samples to multiple regions of interest. Again the Jacobian term cancels out in the importance weight (4.4). Our importance sampling with a mixture proposal is detailed in the following algorithm. For brevity, write

$$\tilde{H}(\hat{\beta}, S; \beta_0) = \sqrt{n}(X^{\mathsf{T}})^{+}(\Psi\hat{\beta} + \lambda W S - \Psi\beta_0),$$

which is identical to the $\tilde{H}$ in (3.10).

Algorithm 2 $(IS(a_1, \beta_1^\dagger, M_1; a_2, \beta_2^\dagger, M_2))$**.** Given $a_1 + a_2 = 1$, $\beta_1^\dagger, \beta_2^\dagger \in \mathbb{R}^p$ and $M_1, M_2 > 0$,

(1) draw Z from $\{1, 2\}$ with probabilities $\{a_1, a_2\}$, and given Z, draw $(\hat{\beta}^*, S^*)$ from $PB(\beta_Z^\dagger, M_Z\sigma^2, \lambda)$;
(2) calculate importance weight

$$w^* = \frac{\phi_n\left(\tilde{H}(\hat{\beta}^*, S^*; \tilde{\beta}); \sigma^2/n\right)}{\sum_{k=1}^{2} a_k\, \phi_n\left(\tilde{H}(\hat{\beta}^*, S^*; \beta_k^\dagger); M_k\sigma^2/n\right)}. \tag{4.4}$$

Remark 7. The first algorithm $IS(\beta^\dagger, M)$ can be regarded as a special case of Algorithm 2 with $a_1 = 1$, $\beta_1^\dagger = \beta^\dagger$ and $M_1 = M$. One can easily generalize Algorithm 2 to a mixture proposal with $K \geq 3$ component distributions. For other error distributions, we simply replace ϕ_n in (4.4) by g_n, the density of $\varepsilon/\sqrt{n}$.

In our numerical results, the efficiency of an importance sampling estimate is measured by its coefficient of variation (cv) across multiple independent runs and compared with direct bootstrap outlined in Algorithm 1.

4.3. Group lasso

We begin with a simpler application to test the complete null hypothesis H_0 : $\beta_0 = 0$ using the statistic $T = h(\hat{\beta}) = \sum_j \|\hat{\beta}_{(j)}\|$, where $\hat{\beta}$ is the group lasso for a particular λ. In this case, our target density $f_A(r_A, s; \beta_0 = 0, \sigma^2, \lambda)$ determines the exact distribution of T under H_0.

We set the group size $p_j = 10$ for all groups and fixed $\sigma^2 = 1$. Each row of X was drawn from $\mathcal{N}_p(0, \Sigma)$, where the diagonal elements of Σ are all 1. The off-diagonal elements $\Sigma_{ij} = \rho_1$ if i, j are in the same group and $\Sigma_{ij} = \rho_2$ otherwise. We simulated 30 datasets with parameters (n, p, ρ_1, ρ_2) reported in Table 1. Put $v = (1, 1, 1, 1, -1, -1, -1, -1, 0, 0)$. For the first 10 datasets, we chose $\beta_0 = 0$ so that H_0 is true. For the other 20 datasets, the first two groups of β_0 were active, with $\beta_{0(1)} = \beta_{0(2)} = v/2$ for datasets 11 to 20 and $\beta_{0(1)} = \beta_{0(2)} = v$ for datasets 21 to 30. For each dataset, λ was chosen to be the smallest value such that the group lasso solution had two active groups. The range of λ and that of the statistic T across the simulated datasets are reported in Table 1 as well.

TABLE 1
Simulated datasets for testing complete null hypothesis

Dataset	(n, p)	(ρ_1, ρ_2)	λ	T
1-10	$(30, 100)$	$(0, 0)$	$(0.396, 0.796)$	$(0.017, 0.337)$
11-20	$(30, 100)$	$(0, 0)$	$(0.554, 1.613)$	$(0.460, 1.964)$
21-30	$(30, 100)$	$(0.5, 0)$	$(0.956, 2.650)$	$(0.045, 2.186)$

We applied the algorithm $IS(0, 5)$ to generate $N = 100,000$ samples. Denote the samples by $\hat{\beta}_t^*$, with importance weight w_t, for $t = 1, \ldots, N$. The p-value for the observed statistic T was then estimated by

$$\hat{q}^{(IS)} = \frac{\sum_{t=1}^N w_t I(h(\hat{\beta}_t^*) \geq T)}{\sum_{t=1}^N w_t}. \tag{4.5}$$

This procedure was repeated 20 times independently for each dataset to calculate the mean $\bar{q}$ and the standard deviation of $\hat{q}^{(IS)}$, from which we calculated $\mathrm{cv}(\hat{q}^{(IS)})$. If we had used the bootstrap algorithm $PB(0, \sigma^2, \lambda)$ for the same N to estimate the p-value, denoted by $\hat{q}^{(PB)}$, its cv would have been close to $\sqrt{(1-\bar{q})/(N\bar{q})}$. Figure 2 plots $\log_{10}(\bar{q})$, $\mathrm{cv}(\hat{q}^{(IS)})$ and $\log_{10}\{\mathrm{cv}(\hat{q}^{(PB)})/\mathrm{cv}(\hat{q}^{(IS)})\}$ for the 30 datasets. We observe from the ratios of cv's in panel (c) that, for

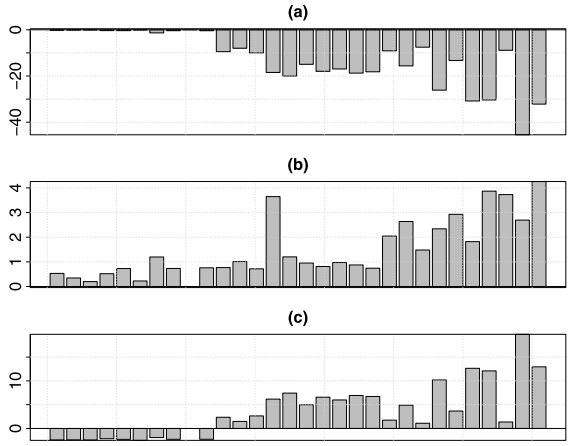

FIG 2. *Estimation of p-values for testing H_0 with the group lasso. (a)* $\log_{10} \bar{q}$, *(b)* $\mathrm{cv}(\hat{q}^{(IS)})$ *and (c)* $\log_{10}\{\mathrm{cv}(\hat{q}^{(PB)})/\mathrm{cv}(\hat{q}^{(IS)})\}$. *The result for a dataset is reported by a vertical bar in each plot.*

datasets 11 to 30, the importance sampling estimates are much more accurate, while the estimated p-values, as shown in panel (a), are very small. For many of these 20 datasets, the improvement of importance sampling over bootstrap can be five or more orders of magnitude. The p-values are insignificant for the first 10 datasets, in which the null hypothesis is true. In a majority of these cases, the importance sampling estimates are slightly less accurate than the bootstrap estimates, which is fully expected.

4.4. A de-biased approach

The second application concerns a de-biased group lasso in the form of (4.1). Since our method applies to any choice of $\hat{\Theta}$, to simplify the discussion we set $\hat{\Theta} = \Sigma^{-1}$ instead of using a particular estimate, where Σ is the population covariance of X. The test statistic is chosen as $h_j(\hat{b}_{(j)}) = \|X_{(j)}\hat{b}_{(j)}\| := T_j$ in (4.3).

We simulated 20 datasets independently under the same settings as those for datasets 11 to 30 in Table 1. The tuning parameter λ was chosen by the same method as in Section 4.3 to calculate the group lasso $\hat{\beta}$ and the de-biased estimate $\hat{b}$ (4.1) for each dataset. Figure 3 plots these two estimates for one dataset, in which $\beta_{0(1)} = \beta_{0(2)} = v$ and $\beta_{0(j)} = 0$ for $j > 2$. We see that the de-biased group lasso $\hat{b}$ is not sparse, $\hat{b}_{(j)} \neq 0$ for all j, and its first two groups are closer to the active groups of β_0 than the group lasso. This largely removed the shrinkage in the active coefficients of the group lasso solution and substantially reduced its bias. Our goal here is to test $H_{0,1} : \beta_{0(1)} = 0$ by estimating the probability (4.3) for $T_1 = \|X_{(1)}\hat{b}_{(1)}\|$ with a plug-in point estimate $\tilde{\beta}$. The observed value of the test statistic T_1 ranges from 4.4 to 21.2 across the

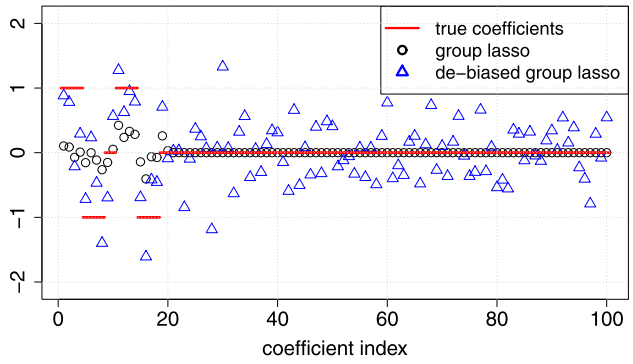

FIG 3. *The group lasso and de-biased group lasso solutions for one dataset with* $p = 100$, *where the size of each group is 10.*

20 datasets. Due to the asymptotic normality of $\hat{b}_{(j)}$, the bootstrap distribution $[\hat{b}^*_{(j)} - \tilde{\beta}_{(j)} \mid \tilde{\beta}]$ is not sensitive to the choice of $\tilde{\beta}$ as long as it is sparse. Thus, we choose $\tilde{\beta} = \hat{\beta}$, the group lasso estimate. See [5] for related discussions.

We designed the following mixture proposal for Algorithm 2 to approximate the p-value (4.3) by importance sampling:

$$a_1 = a_2 = 1/2; M_1 = 2, M_2 = 4; \beta^\dagger_1 = \hat{\beta}, \beta^\dagger_{2(1)} = \hat{\beta}_{(1)}/2, \beta^\dagger_{2(-1)} = \hat{\beta}_{(-1)}.$$

Note that $\beta^\dagger_{2(1)}$ is the middle point between $\hat{\beta}_{(1)}$ and 0, serving as a bridge between the target distribution and the null hypothesis $H_{0,1}$. To achieve a wider coverage of the sample space, the error variances of both component distributions were chosen to be greater than σ^2. We applied Algorithm 2 to generate $N = 100,000$ weighted samples $(\hat{\beta}^*_t, S^*_t)$, with weights w_t, for each dataset. Similar to (4.5), the p-value for T_1 was estimated as

$$\hat{q}^{(IS)} = \frac{\sum_{t=1}^N w_t I(\|X_{(1)}(\hat{b}^*_{t(1)} - \hat{\beta}_{(1)})\| \geq T_1)}{\sum_{t=1}^N w_t}, \tag{4.6}$$

where $\hat{b}^*_t = \hat{b}(\hat{\beta}^*_t, S^*_t)$ as in (4.1). We replicated this procedure 20 times independently to calculate the cv of $\hat{q}^{(IS)}$ as we did in the previous example. The same comparisons were conducted and the results are reported in Figure 4. Strong majority of the p-values were estimated to be significant, since $\beta_{0(1)} \neq 0$ for all 20 datasets. The cv's of the importance sampling estimates are seen to be quite small, which is especially satisfactory for those tiny tail probabilities on the order of 10^{-10} or smaller. As shown in Figure 4(c), our importance sampling estimation is more efficient than parametric bootstrap for at least 13 out of the 20 datasets, many showing orders of magnitude of improvement. For most of the other datasets, the importance sampling results are very comparable to the results from bootstrap.

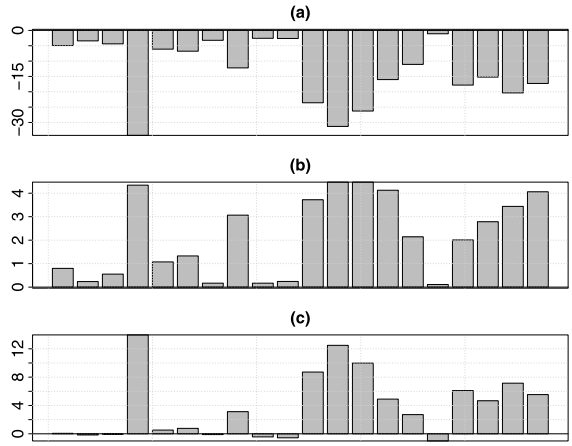

FIG 4. *Estimation of p-values for testing $H_{0,1}$ with a de-biased group lasso. Plots are in the same format as those in Figure 2.*

Compared to the parametric bootstrap in Algorithm 1, the only additional step in our importance sampling algorithms is to evaluate importance weights, such as (4.4), of which the computing time is negligible relative to computing group lasso solutions. As a result, the total running time of the importance sampling is almost identical to that of the bootstrap sampling. The above two applications thus exemplify the huge gain in estimation accuracy by importance sampling via estimator augmentation at almost identical computing cost. It is worth mentioning that accurate estimation of small p-values is crucial for ranking the importance of predictors and controlling false discoveries in large-scale screening.

4.5. Other applications

Given the joint density $f_A(r_A, s; \tilde{\beta}, \sigma^2, \lambda)$, one may design MCMC algorithms to draw samples $(\hat{\beta}^*, S^*)$ from this distribution, which is identical to the distribution of a bootstrap sample generated by $PB(\tilde{\beta}, \sigma^2, \lambda)$ in Algorithm 1. The advantage of an MCMC algorithm is that it does not need to solve a convex optimization program in any of its steps. But evaluating the Jacobian term in f_A could be time-consuming. Another potential application is conditional sampling from $[\hat{\beta}^*, S^* \mid \hat{\beta}^* \in B]$, which will be useful in post-selection inference. For example, conditioning on the model selected by $\hat{\beta}$, i.e. $G(\hat{\beta}^*) = G(\hat{\beta})$, we may wish to sample from an estimator $\hat{b}^*$ with a nice asymptotic distribution for inference. For this problem, bootstrap may be impractical since the conditioning event is often a rare event. However, from the joint density one can easily obtain the conditional density $\propto f_G(r_G, s)$, where $G = G(\hat{\beta})$, and implement an MCMC algorithm to draw from this conditional distribution. In the case of

the lasso, Zhou [29] implemented an Metropolis-Hastings sampler for such conditional sampling. The more general case for a block lasso will be considered in the future.

Under a Gaussian error assumption, it is a common practice to plug an estimated variance $\hat{\sigma}^2$ in the bootstrap $PB(\tilde{\beta}, \hat{\sigma}^2, \lambda)$. As long as $\hat{\sigma}^2$ is consistent with a certain rate, inference will be valid asymptotically [5, 30]. Therefore, we can use our importance sampling algorithms with $f_A(r_A, s; \tilde{\beta}, \hat{\sigma}^2, \lambda)$ as the target density. Note that the density f_A (3.13) depends on the error distribution only through the density g_n of $\varepsilon/\sqrt{n}$. Under a general i.i.d. error assumption, estimating g_n reduces to estimating the density of an univariate distribution, which can be done quite accurately even when n is moderate by either a parametric or a nonparametric method. Given an estimate $\hat{g}_n$, our target density is readily obtained with g_n replaced by $\hat{g}_n$. An appealing alternative is to de-bias a scaled block lasso, which estimates σ^2 in a coherent way, for inference as in [12]. Estimator augmentation can be applied to derive the joint density of an augmented scaled block lasso, including its variance estimator $\hat{\sigma}^2$, as outlined in Section 5.2. Given the density, one can follow the same importance sampling algorithms for tail probability approximation.

5. Generalizations

We generalize estimator augmentation to the block lasso with $\alpha = \infty$ and to a scaled block lasso. In both cases, the subgradient has more structure.

5.1. Block-$(1, \infty)$ norm

In this subsection, we consider the case $\alpha = \infty$ ($\alpha^* = 1$). The difference between this case and the case $\alpha < \infty$ comes from the subgradient vector S. Let $\mathcal{B}_j = \operatorname{argmax}_{k \in \mathcal{G}_j} |\hat{\beta}_k| \subset \mathcal{G}_j$, which may contain multiple elements when a tie occurs, and $\mathcal{B}_j^c = \mathcal{G}_j \setminus \mathcal{B}_j$ for $j \in \mathbb{N}_J$. It follows from Lemma A.3 that (i) for $\hat{\beta}_{(j)} \neq 0$, $\|S_{(j)}\|_1 = 1$ and

$$S_k = \begin{cases} t_k \operatorname{sgn}(\hat{\beta}_k) & k \in \mathcal{B}_j \\ 0 & k \in \mathcal{B}_j^c \end{cases}, \tag{5.1}$$

where $\sum_{\mathcal{B}_j} t_k = 1$ and $t_k \geq 0$; (ii) $\|S_{(j)}\|_1 \leq 1$ for $\hat{\beta}_{(j)} = 0$.

Compared to (2.2), the discreteness of $\{S_k = 0\}$ for some k as in (5.1) distinguishes the $(1, \infty)$ norm from other cases of $\alpha < \infty$. Accordingly, the augmented estimator $(\hat{\beta}, S)$ will have more structure. Recall that the active blocks of $\hat{\beta}$ are denoted by $\mathcal{A} = G(\hat{\beta})$. For $j \in \mathcal{A}$, define

$$\mathcal{K}_j = \{k \in \mathcal{G}_j : S_k \neq 0\} \quad \text{and} \quad \mathcal{K}_j^c = \mathcal{G}_j \setminus \mathcal{K}_j. \tag{5.2}$$

Put $\mathcal{K} = \cup\{\mathcal{K}_j : j \in \mathcal{A}\}$ and $\mathcal{K}^c = \cup\{\mathcal{K}_j^c : j \in \mathcal{A}\}$. It follows from (5.1) that $\hat{\beta}_{\mathcal{K}_j} = \hat{\gamma}_j \operatorname{sgn}(S_{\mathcal{K}_j})$ for $j \in \mathcal{A}$, where $\hat{\gamma}_j = \|\hat{\beta}_{(j)}\|_\infty$. We can then represent $(\hat{\beta}, S)$

by

$$(\hat{\gamma}_{\mathcal{A}}, \hat{\beta}_{\mathcal{K}^c}, S, \mathcal{A}, \mathcal{K}) \quad \text{with } S_{\mathcal{K}^c} = 0, \tag{5.3}$$

subject to the constraints that $\|\hat{\beta}_{\mathcal{K}_j^c}\|_\infty \le \hat{\gamma}_j$ for $j \in \mathcal{A}$. For $\mathcal{A} = A \in \mathscr{A}$ (3.3), Proposition A.6 implies that assuming solution uniqueness the range of $\mathcal{K}$ is

$$\mathscr{K}(A) = \{K \subset \mathcal{G}_A : K \cap \mathcal{G}_j \neq \varnothing \ \forall j \in A \text{ and } |\mathcal{G}_A \setminus K| \le n - |A|\}.$$

Let $K_j = K \cap \mathcal{G}_j$ and $K_j^c = K^c \cap \mathcal{G}_j$ for $j \in A$, where $K^c = \mathcal{G}_A \setminus K$. The sample space for S given $\mathcal{A} = A$ and $\mathcal{K} = K$ is

$$\mathcal{M}_{A,K} = \{s \in \mathcal{M}_A : s_k \neq 0 \ \forall k \in K, s_{K^c} = 0\},$$

where $\mathcal{M}_A$ is as in (3.1) with $\alpha^* = 1$. The sample space for $(\hat{\gamma}_j, \hat{\beta}_{K_j^c})$ is the cone

$$\mathcal{C}_j = \left\{(r, v) \in \mathbb{R}^+ \times \mathbb{R}^{|K_j^c|} : \|v\|_\infty \le r\right\}.$$

Then the sample space for $(\hat{\gamma}_A, \hat{\beta}_{K^c}, S)$ is the product $\Omega_{A,K} = (\prod_{j\in A} \mathcal{C}_j) \times \mathcal{M}_{A,K}$. Taking union over the range of the sets $A \in \mathscr{A}$ and $K \in \mathscr{K}(A)$ determines the space Ω for the augmented estimator (5.3). Compared to the case $\alpha < \infty$, the subgradient S has lost $|\mathcal{K}^c|$ free dimensions due to the constraints that $S_k = 0$ for all $k \in \mathcal{K}^c$. Consequently, for every interior point $s \in \mathcal{M}_{A,K}$, there is a neighborhood that may be parameterized by s_F with $|F| = n - |A| - |K^c| := q$. Note that $ds_k = 0$ for each $k \in K^c$. Let $I = \mathbb{N}_p \setminus K^c$. Similar to Lemma 3.2, we can then find a matrix $T \in \mathbb{R}^{(p-|K^c|)\times q}$ such that $ds_I = T ds_F$.

For notational brevity we will use $(\hat{\beta}, S)$ and its equivalent representation (5.3) interchangeably. Write the mappings H (3.4) and $\tilde{H}$ (3.10) as $H(b, s)$ and $\tilde{H}(b, s)$, respectively, where (b, s) denotes the value of $(\hat{\beta}, S)$. For $(r_A, b_{K^c}, s) \in \Omega_{A,K}$, let $\tilde{H}_{A,K}(r_A, b_{K^c}, s) = \tilde{H}(b, s)$ with (r_A, b_{K^c}, s, A, K) being the equivalent representation of (b, s). Define two matrices

$$\Psi \circ \operatorname{sgn}(s) = [\Psi_{(1)} \operatorname{sgn}(s_{(1)}) | \dots | \Psi_{(J)} \operatorname{sgn}(s_{(J)})] \in \mathbb{R}^{p\times J},$$
$$M(s, A, K; \lambda) = [\{\Psi \circ \operatorname{sgn}(s)\}_A \mid \Psi_{K^c} \mid \lambda W_I T] \in \mathbb{R}^{p\times n},$$

and a related Jacobian

$$J_{A,K}(s; \lambda) = \det\left[\sqrt{n}(X^\mathsf{T})^+ M(s, A, K; \lambda)\right]. \tag{5.4}$$

Parallel to Theorem 3.3, we have the following explicit density for the augmented estimator under block-$(1, \infty)$ sparsity, which is proved in Appendix B.6.

Theorem 5.1. *Fix $p \ge n$, $\beta_0 \in \mathbb{R}^p$ and $\lambda > 0$. Suppose Assumption 1 holds and that the program (1.2) for $\alpha = \infty$ has a unique solution for almost all $y \in \mathbb{R}^n$. Let g_n be the density of $(\varepsilon/\sqrt{n})$. Then the distribution of the augmented estimator $(\hat{\beta}, S)$ is given by the n-form*

$$\begin{aligned} d\mu_{A,K} &:= \mathbb{P}(dr_A, db_{K^c}, ds, \{A, K\}) \\ &= g_n(\tilde{H}_{A,K}(r_A, b_{K^c}, s; \beta_0, \lambda))|J_{A,K}(s; \lambda)| d\theta := f_{A,K}(r_A, b_{K^c}, s) d\theta \end{aligned} \tag{5.5}$$

for $(r_A, b_{K^c}, s, A, K) \in \Omega$, where $\theta = (r_A, b_{K^c}, s_F) \in \mathbb{R}^n$.

The sufficient condition for solution uniqueness in this case is discussed in Appendix A.4. The density $f_{A,K}$ is defined in terms of the parameterization θ. Suppose that $\Gamma \subset \mathbb{R}^{|A|+|K^c|}$ is a subset of the product cone $\prod_{j\in A}\mathcal{C}_j$ and $\Phi = \{\Phi(s_F) : s_F \in \Delta\} \subset \mathcal{M}_{A,K}$ is a q-surface in $\mathbb{R}^p$ with parameter domain $\Delta \subset \mathbb{R}^q$. Then we have

$$\mathbb{P}\left\{(\hat{\gamma}_A, \hat{\beta}_{K^c}) \in \Gamma, S \in \Phi, \mathcal{A} = A, \mathcal{K} = K\right\} = \int_{\Gamma\times\Delta} f_{A,K}(r_A, b_{K^c}, \Phi(s_F))d\theta,$$

which interprets the differential form (5.5). The density here differs from that in (3.12) only in the Jacobian term. Clearly, the same importance sampling method (Algorithm 2) can be used here due to the cancellation of the Jacobian.

5.2. A scaled block lasso

As another generalization, we consider estimator augmentation for a scaled block lasso, which is scale invariant and provides an estimate of σ^2 simultaneously. Mitra and Zhang [12] have developed inference methods via de-biasing a scaled group lasso, which is related to the square-root group lasso [2]. Following their formulation, we define a scaled block lasso

$$(\hat{\beta}, \hat{\sigma}) \in \underset{\beta\in\mathbb{R}^p, \sigma>0}{\operatorname{argmin}} \left\{ L(\beta, \sigma) := \frac{1}{2n\sigma}\|y - X\beta\|^2 + \frac{\sigma}{2} + \lambda\sum_{j=1}^{J} w_j\|\beta_{(j)}\|_\alpha \right\}. \quad (5.6)$$

Since the loss function in (5.6) is convex, $(\hat{\beta}, \hat{\sigma})$ is given by the KKT conditions

$$\frac{1}{n}X^\mathsf{T}(y - X\hat{\beta}) = \lambda\hat{\sigma}WS, \quad (5.7)$$

$$\hat{\sigma} = \|y - X\hat{\beta}\|/\sqrt{n}. \quad (5.8)$$

Under Assumption 1, (5.7) is equivalent to $y - X\hat{\beta} = \lambda n\hat{\sigma}(X^\mathsf{T})^+WS$, plugging which into (5.8), we arrive at

$$\lambda\sqrt{n}\|(X^\mathsf{T})^+WS\| = 1. \quad (5.9)$$

It is easy to see that (5.7) and (5.9) imply (5.8), and thus are sufficient and necessary for $(\hat{\beta}, \hat{\sigma})$ to be a scaled block lasso solution. This shows that the subgradient S here satisfies an additional equality constraint. Therefore, for $\mathcal{A} = G(\hat{\beta}) = A$, its sample space

$$\widetilde{\mathcal{M}}_A = \{s \in \mathcal{M}_A : \lambda\sqrt{n}\|(X^\mathsf{T})^+Ws\| = 1\} \quad (5.10)$$

is an $(n-|A|-1)$-manifold, where $\mathcal{M}_A$ is defined in (3.1). Note that $\sqrt{n}(X^\mathsf{T})^+z$ is the coordinates of $z \in \operatorname{row}(X)$ with respect to the basis $(X^\mathsf{T}/\sqrt{n})$. Thus, the vector λWS for a scaled block lasso is normalized with respect to this basis. Let

$(\hat{\gamma}_{\mathcal{A}}, S, \mathcal{A})$ be the equivalent representation of $(\hat{\beta}, S)$. Given $\mathcal{A} = A$, the space for $(\hat{\gamma}_A, S, \hat{\sigma})$ is

$$\widetilde{\Omega}_A = (\mathbb{R}^+)^{|A|} \times \widetilde{\mathcal{M}}_A \times \mathbb{R}^+,$$

which is still a manifold of dimension n.

Suppose the program (5.6) has a unique minimizer for almost all $y \in \mathbb{R}^n$. Substituting y by $X\beta_0 + \varepsilon$, (5.7) defines a bijective mapping $F : (r_A, s, \hat{\sigma}, A) \mapsto v = \varepsilon/\sqrt{n}$, for $(r_A, s, \hat{\sigma}) \in \widetilde{\Omega}_A$. Recall that (5.7) with $S \in \widetilde{\mathcal{M}}_{\mathcal{A}}$ is equivalent to the KKT conditions, and thus this mapping is sufficient for determining the distribution of $(\hat{\beta}, S, \hat{\sigma})$. The restriction of F to a fixed A defines a one-to-one mapping $F_A : \widetilde{\Omega}_A \to \mathbb{R}^n$, given by

$$F_A(r_A, s, \hat{\sigma}; \beta_0, \lambda) := \tilde{H}(r_A, s, A; \beta_0, \lambda\hat{\sigma}),$$

where $\tilde{H}$ is defined in (3.10). As in Lemma 3.2, we parameterize a neighborhood of $s \in \widetilde{\mathcal{M}}_A$ by s_F, with $|F| = n - |A| - 1$, so that $ds = T(s, A)ds_F$, where $T = T(s, A)$ is a $p \times |F|$ matrix. Following a similar derivation as in the proof of Lemma 3.2 in Appendix B.1, we find the Jacobian of F_A,

$$J_A(r_A, s, \hat{\sigma}; \lambda) = \det\left[\sqrt{n}(X^{\mathsf{T}})^+ M(r_A, s, \hat{\sigma}, A; \lambda)\right], \tag{5.11}$$

with respect to the parameterization $\theta = (r_A, s_F, \hat{\sigma}) \in \mathbb{R}^n$, where the matrix

$$M(r_A, s, \hat{\sigma}, A; \lambda) = [(\Psi \circ \eta)_A \mid \{(r \circ \Psi)D + \lambda\hat{\sigma}W\}T \mid \lambda W s] \in \mathbb{R}^{p \times n}.$$

The joint distribution of $(\hat{\beta}, S, \hat{\sigma})$ is then given by the n-form,

$$f_A(r_A, s, \hat{\sigma})d\theta = g_n(F_A(r_A, s, \hat{\sigma}; \beta_0, \lambda))|J_A(r_A, s, \hat{\sigma}; \lambda)|d\theta, \tag{5.12}$$

where g_n is the density of $(\varepsilon/\sqrt{n})$. This result applies to all $\alpha \in [1, \infty)$, under the convention that $p_j = 1$ for all j when $\alpha = 1$, in which case (5.6) reduces to a scaled lasso [1, 18].

An assumption for the above result is the uniqueness of $(\hat{\beta}, \hat{\sigma})$. Given $\sigma > 0$, denote by $\hat{\beta}(\sigma\lambda)$ the restricted minimizer of (5.6), which is the same as the block lasso (1.2) with tuning parameter $\sigma\lambda$. Therefore, under Assumptions 1 and 2, $\hat{\beta}(\sigma\lambda)$ is unique for any $\sigma > 0$. As established in Lemma 2 of [12], the profile loss function $L(\hat{\beta}(\sigma\lambda), \sigma)$ is convex and continuously differentiable in σ. Thus, $\hat{\sigma}$ is given by any solution to the equation

$$\frac{dL(\hat{\beta}(\sigma\lambda), \sigma)}{d\sigma} = \frac{1}{2} - \frac{\|y - X\hat{\beta}(\sigma\lambda)\|^2}{2n\sigma^2} = 0. \tag{5.13}$$

If this equation has a unique solution in $(0, \infty)$, then $(\hat{\beta}, \hat{\sigma})$ is unique. We summarize this result into the following theorem.

Theorem 5.2. *Fix $p \geq n$, $\beta_0 \in \mathbb{R}^p$, $\lambda > 0$, and $\alpha \in [1, \infty)$. Suppose that Assumptions 1 and 2 hold, and that Equation (5.13) has a unique solution in*

$(0,\infty)$ *for almost all* $y \in \mathbb{R}^n$. *Then the distribution of the augmented scaled block lasso* $(\hat{\beta}, S, \hat{\sigma})$ *defined by* (5.6) *is given by the* n*-form,*

$$\mathbb{P}(dr_A, ds, d\hat{\sigma}, \{A\}) = f_A(r_A, s, \hat{\sigma})d\theta$$

as in (5.12), *for* $(r_A, s, \hat{\sigma}) \in \widetilde{\Omega}_A$ *and* $|A| \le n-1$.

We believe (5.13) indeed has a unique solution for almost all $y \in \mathbb{R}^n$ under fairly weak but perhaps technical assumptions. See Appendix A.5 for a detailed discussion. To avoid this technical issue in practice, one may include another additive term $a\sigma^2$ in the loss function in (5.6), where a is a small positive constant. The uniqueness of $\hat{\sigma}$ is then an immediate consequence of the strong convexity of the modified profile loss, $L(\hat{\beta}(\sigma\lambda), \sigma) + a\sigma^2$.

6. Concluding remarks

By augmenting the sample space to that of $(\hat{\beta}, S)$, we have derived a closed-form density for the sampling distribution of the augmented block lasso estimator. Given the density, we have demonstrated the use of importance sampling in group inference, which can be orders of magnitude more efficient than the corresponding parametric bootstrap. For high-dimensional data, sparsity seems an essential assumption for inference, and consequently, an inference method is often built upon a non-regular penalized estimator. It is unlikely to work out an exact pivot in this setting, and thus, simulation-based approaches have been widely used. Our work of estimator augmentation opens the door to a large class of Monte Carlo methods for such simulations, which in our view is the main intellectual contribution. Due to the complexity of the sample space of an augmented estimator, development of efficient Monte Carlo algorithms is a highly demanding job and an interesting future direction.

Appendix A: Uniqueness of the block lasso

A.1. Auxiliary lemmas

Lemma A.1. *If* $\alpha \in (1,\infty)$, *then* η *is a bijection that maps* $\mathbb{S}_{\alpha^*}$ *onto* $\mathbb{S}_\alpha$ *and* $\langle \eta(v), v\rangle = 1$ *for any* $v \in \mathbb{S}_{\alpha^*}$.

Proof. For any $v = (v_i) \in \mathbb{S}_{\alpha^*}$,

$$\|\eta(v)\|_\alpha^\alpha = \|(v_i^\rho)\|_\alpha^\alpha = \sum_i |v_i|^{\alpha^*} = 1.$$

Similarly, we can show that $\eta^{-1}(u) \in \mathbb{S}_{\alpha^*}$ for any $u \in \mathbb{S}_\alpha$. By definition, $\rho + 1 = \alpha^*$. Then, straightforward calculation leads to

$$\langle \eta(v), v\rangle = \langle \operatorname{sgn}(v)|v|^\rho, \operatorname{sgn}(v)|v|\rangle = \sum_i |v_i|^{\rho+1} = \sum_i |v_i|^{\alpha^*} = 1.$$

Here, $|\cdot|$ and $\operatorname{sgn}(\cdot)$ are applied on v in the sense of (2.1). □

Lemma A.2. *Let $h(v) = \|v\|_\alpha$ for $\alpha \in (1,\infty)$ and $v \in \mathbb{R}^m$. If $v \neq 0$, then*

$$\nabla h(v) = \eta^{-1}(\tilde{v}) \in \mathbb{S}_{\alpha^*}^{m-1} \tag{A.1}$$

with $\tilde{v} = v/\|v\|_\alpha \in \mathbb{S}_\alpha^{m-1}$. If $v = 0$, the subdifferential of h

$$\partial h(0) = \{u \in \mathbb{R}^m : \|u\|_{\alpha^*} \le 1\}. \tag{A.2}$$

Proof. For $v \neq 0$,

$$\frac{\partial h}{\partial v_i} = \frac{\operatorname{sgn}(v_i)|v_i|^{\alpha-1}}{\|v\|_\alpha^{\alpha-1}} = \operatorname{sgn}(\tilde{v}_i)|\tilde{v}_i|^{1/\rho} = \eta^{-1}(\tilde{v}_i),$$

using the simple fact that $\alpha - 1 = 1/\rho$. Since by definition $\tilde{v} \in \mathbb{S}_\alpha$, Lemma A.1 implies that $\eta^{-1}(\tilde{v}) \in \mathbb{S}_{\alpha^*}$. This proves (A.1). By Hölder's inequality,

$$\langle u, v\rangle \le \|u\|_{\alpha^*}\|v\|_\alpha \le h(v), \quad \forall v \in \mathbb{R}^m,$$

if and only if $\|u\|_{\alpha^*} \le 1$, which implies (A.2). □

Lemma A.3 (Lemma 1 in [14])**.** *Let $h(v) = \|v\|_\infty$ for $v \in \mathbb{R}^m$ and $K = \operatorname{argmax}_i |v_i| \subset \mathbb{N}_m$. For $v \neq 0$, $u \in \partial h(v)$ if and only if*

$$u_i = \begin{cases} t_i \operatorname{sgn}(v_i) & i \in K \\ 0 & \text{otherwise} \end{cases}$$

for some $(t_i)_{i\in K}$ so that $\sum_i t_i = 1$ and $t_i \ge 0$. For $v = 0$,

$$\partial h(0) = \{u \in \mathbb{R}^m : \|u\|_1 \le 1\}. \tag{A.3}$$

A.2. Characterization of solutions

Proof of Lemma 2.1. Suppose that $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are two minimizers of $L(\beta;\alpha)$ such that $X\hat{\beta}^{(1)} \neq X\hat{\beta}^{(2)}$. The convexity of L implies that

$$L(\hat{\beta}^{(1)};\alpha) = L(\hat{\beta}^{(2)};\alpha) = L^*.$$

Since $\|x\|^2$ is strictly convex in x, for any $c \in (0,1)$,

$$\|y - X[c\hat{\beta}^{(1)} + (1-c)\hat{\beta}^{(2)}]\|^2 < c\|y - X\hat{\beta}^{(1)}\|^2 + (1-c)\|y - X\hat{\beta}^{(2)}\|^2$$

by the hypothesis that $X\hat{\beta}^{(1)} \neq X\hat{\beta}^{(2)}$. Therefore,

$$L(c\hat{\beta}^{(1)} + (1-c)\hat{\beta}^{(2)};\alpha) < cL(\hat{\beta}^{(1)};\alpha) + (1-c)L(\hat{\beta}^{(2)};\alpha) = L^*,$$

which is contradictory to the assumption that the minimum of L is L^*. The uniqueness of S is an immediate consequence of the uniqueness of $X\hat{\beta}$ and that

$$S = (n\lambda W)^{-1}X^{\mathsf{T}}(y - X\hat{\beta}) \tag{A.4}$$

by the KKT conditions (2.3). □

We will first establish sufficient conditions for solution uniqueness for $\alpha < \infty$, while deferring the case $\alpha = \infty$ to Appendix A.4. We start with more explicit expressions for $X\hat{\beta}$ and $\hat{\beta}$. Write the KKT conditions for each block in (2.3),

$$\frac{1}{n} X_{(j)}^{\mathsf{T}} X\hat{\beta} + \lambda w_j S_{(j)} = \frac{1}{n} X_{(j)}^{\mathsf{T}} y, \quad j = 1, \ldots, J. \tag{A.5}$$

Define

$$\mathcal{E} = \left\{ j \in \mathbb{N}_J : \frac{1}{w_j n} \left\| X_{(j)}^{\mathsf{T}} (y - X\hat{\beta}) \right\|_{\alpha^*} = \lambda \|S_{(j)}\|_{\alpha^*} = \lambda \right\}. \tag{A.6}$$

By (A.5) and (2.2), $\hat{\beta}_{(-\mathcal{E})} = 0$. Now the $\mathcal{E}$ block of (A.5) with $\hat{\beta}_{(-\mathcal{E})} = 0$ reads

$$\frac{1}{n} X_{(\mathcal{E})}^{\mathsf{T}} (y - X_{(\mathcal{E})} \hat{\beta}_{(\mathcal{E})}) = \lambda W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})}, \tag{A.7}$$

which shows that $W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})} \in \operatorname{row}(X_{(\mathcal{E})})$. Thus, we have

$$W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})} = X_{(\mathcal{E})}^{\mathsf{T}} (X_{(\mathcal{E})}^{\mathsf{T}})^{+} W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})}, \tag{A.8}$$

since the right side is the projection of $W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})}$ onto $\operatorname{row}(X_{(\mathcal{E})})$. Plugging the above identity into (A.7), we arrive at

$$X_{(\mathcal{E})}^{\mathsf{T}} X_{(\mathcal{E})} \hat{\beta}_{(\mathcal{E})} = X_{(\mathcal{E})}^{\mathsf{T}} \left[y - n\lambda (X_{(\mathcal{E})}^{\mathsf{T}})^{+} W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})} \right]. \tag{A.9}$$

A solution to the above equation is

$$\begin{aligned} \hat{\beta}_{(\mathcal{E})} &= (X_{(\mathcal{E})}^{\mathsf{T}} X_{(\mathcal{E})})^{+} X_{(\mathcal{E})}^{\mathsf{T}} \left[y - n\lambda (X_{(\mathcal{E})}^{\mathsf{T}})^{+} W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})} \right] \\ &= (X_{(\mathcal{E})})^{+} \left[y - n\lambda (X_{(\mathcal{E})}^{\mathsf{T}})^{+} W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})} \right]. \end{aligned}$$

Then by the uniqueness of the fit $X\hat{\beta}$ (Lemma 2.1), for all solutions $\hat{\beta}$ we have

$$X\hat{\beta} = X_{(\mathcal{E})} \hat{\beta}_{(\mathcal{E})} = X_{(\mathcal{E})} (X_{(\mathcal{E})})^{+} \left[y - n\lambda (X_{(\mathcal{E})}^{\mathsf{T}})^{+} W_{(\mathcal{E}\mathcal{E})} S_{(\mathcal{E})} \right] := \hat{y}. \tag{A.10}$$

To make the relation between $\hat{\beta}_{(\mathcal{E})}$ and $S_{(\mathcal{E})}$ more explicit, write

$$\hat{\beta}_{(j)} = \|\hat{\beta}_{(j)}\|_{\alpha} \eta(S_{(j)}) = \hat{\gamma}_j \eta(S_{(j)}) \quad \text{for } j \in \mathcal{E}, \tag{A.11}$$

which follows from (2.2). For $B \subset \mathbb{N}_p$, let $\mathbb{R}^B$ be $|B|$-dimensional Euclidean space with coordinates index by B so that a vector $v \in \mathbb{R}^B$ has components v_j, $j \in B$. Similarly, $\mathbb{R}^{m \times B}$ denotes the space of matrices with columns indexed by B. Put

$$Z_j = X_{(j)} \eta(S_{(j)}) \in \mathbb{R}^n, \quad Z = (Z_j)_{j \in \mathcal{E}} \in \mathbb{R}^{n \times \mathcal{E}}. \tag{A.12}$$

Then (A.10) can be rewritten

$$\sum_{j \in \mathcal{E}} \hat{\gamma}_j X_{(j)} \eta(S_{(j)}) = Z\hat{\gamma}_{\mathcal{E}} = \hat{y}.$$

Now we have the following characterization of the block lasso solutions:

Lemma A.4. *If $\hat{\beta}$ is a block lasso solution (1.2) for $\alpha \in (1, \infty)$, then*

$$Z\hat{\gamma}_{\mathcal{E}} = \hat{y} \quad \text{and} \quad \hat{\beta}_{(-\mathcal{E})} = 0. \tag{A.13}$$

Moreover, $\mathcal{E}$, $\hat{y}$ and Z are unique for any y, X and $\lambda > 0$.

The uniqueness of $(\hat{y}, \mathcal{E}, Z)$ is an immediate consequence of Lemma 2.1. So non-uniqueness can only come from $\hat{\gamma}_{\mathcal{E}}$ when the linear system $Zx = \hat{y}$ has multiple solutions for x, which happens only if $\mathrm{null}(Z) \neq \{0\}$. Therefore, every block lasso solution $\hat{\beta}$ satisfies:

$$\hat{\beta}_{(-\mathcal{E})} = 0 \quad \text{and} \quad \hat{\gamma}_{\mathcal{E}} = Z^{+}\hat{y} + \gamma, \tag{A.14}$$

provided that

$$\gamma \in \mathrm{null}(Z) \subset \mathbb{R}^{\mathcal{E}} \quad \text{and} \quad (Z^{+}\hat{y} + \gamma)_j \geq 0 \ \text{ for } j \in \mathcal{E}. \tag{A.15}$$

A.3. Proof of sufficiency

If $\mathrm{null}(Z) = \{0\}$, then $\hat{\beta}$ is uniquely given by (A.14) with $\gamma = 0$. In this case $\hat{\gamma}_j = (Z^{+}\hat{y})_j$ is necessarily nonnegative as there always exists a solution to the block lasso problem. Furthermore, $|\mathcal{E}| \leq n$ and thus this solution has at most $(n \wedge J)$ nonzero blocks. This leads to our first sufficient condition for the uniqueness of $\hat{\beta}$.

Proposition A.5. *Suppose $\lambda > 0$ and $\mathrm{null}(Z) = \{0\}$. Then the block lasso solution $\hat{\beta}$ for $\alpha \in (1, \infty)$ is uniquely given by*

$$\hat{\beta}_{(-\mathcal{E})} = 0, \qquad \hat{\gamma}_{\mathcal{E}} = (Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}\hat{y}, \quad \text{and} \quad \hat{\beta}_{(j)} = \hat{\gamma}_j \eta(S_{(j)}) \text{ for } j \in \mathcal{E}. \tag{A.16}$$

Furthermore, $|G(\hat{\beta})| \leq n \wedge J$.

In the following, we prove Theorem 2.2 for $\alpha \in (1, \infty)$. Note that the case $\alpha = 1$ is equivalent to the case $\alpha = 2$ with $p_j = 1$ for all j. Thus, this part covers the range of $\alpha \in [1, \infty)$ as in Theorem 2.2. The result for $\alpha = \infty$ will be established in next subsection.

Proof of Theorem 2.2. Suppose that $\mathrm{null}(Z) \neq \{0\}$. Then for some $i \in \mathcal{E}$, there is a set $A \subset \mathcal{E} \setminus \{i\}$ and $|A| \leq n$ such that

$$Z_i/w_i = \sum_{j \in A} c_j (Z_j/w_j),$$

where we may assume that Z_j, $j \in A$, are linearly independent and $c_j \neq 0$ without loss of generality. Let $r = y - X\hat{\beta}$ denote the block lasso residual. By (A.5), for every $j \in \mathcal{E}$,

$$\langle Z_j, r \rangle = \langle X_{(j)}\eta(S_{(j)}), r \rangle = n\lambda w_j \langle \eta(S_{(j)}), S_{(j)} \rangle = n\lambda w_j,$$

where the last equality follows from Lemma A.1 since $S_{(j)} \in \mathbb{S}_{\alpha^*}$. Therefore, for $\lambda > 0$ we have

$$1 = \sum_{j \in A} c_j.$$

Note that $Z_j = X_{(j)}\eta(S_{(j)})$, $S \in \operatorname{row}(XW^{-1})$ and $\|S_{(j)}\|_{\alpha^*} = 1$ for $j \in \mathcal{E}$. The above equality is thus contradictory to the assumption that the columns of XW^{-1} are in blockwise general position (Assumption 2). □

A.4. The case of $\alpha = \infty$

Recall the KKT conditions in (A.5). Let $\alpha^* = 1$ in (A.6) to define $\mathcal{E}$. By definition $\hat{\beta}_{(j)} = 0$ for $j \notin \mathcal{E}$. For $j \in \mathcal{E}$ define $\mathcal{K}_j$ and $\mathcal{K}_j^c$ as in (5.2). Note that both $\mathcal{E}$ and $\mathcal{K}_j$ are unique due to the uniqueness of S and $\hat{y} = X\hat{\beta}$ for any y, X and $\lambda > 0$ (Lemma 2.1). It follows from (5.1) and (5.2) that $\hat{\beta}_{\mathcal{K}_j} = \hat{\gamma}_j \operatorname{sgn}(S_{\mathcal{K}_j})$ for each $j \in \mathcal{E}$. Then the fitted value $\hat{y}$ can be expressed as

$$\hat{y} = X_{(\mathcal{E})}\hat{\beta}_{(\mathcal{E})} = \sum_{j \in \mathcal{E}} \left\{ \hat{\gamma}_j X_{\mathcal{K}_j} \operatorname{sgn}(S_{\mathcal{K}_j}) + X_{\mathcal{K}_j^c}\hat{\beta}_{\mathcal{K}_j^c} \right\} = Z\hat{\zeta},$$

where we define

$$Z = \left[X_{\mathcal{K}_j} \operatorname{sgn}(S_{\mathcal{K}_j}) \mid X_{\mathcal{K}_j^c} \right]_{j \in \mathcal{E}} \quad \text{and} \quad \hat{\zeta} = (\hat{\gamma}_j, \hat{\beta}_{\mathcal{K}_j^c})_{j \in \mathcal{E}}. \tag{A.17}$$

If $\operatorname{null}(Z) = \{0\}$, then $\hat{\zeta}$ and hence $\hat{\beta}$ will be unique and Z has at most n columns. Now we generalize Proposition A.5 to the block-$(1, \infty)$ norm regularization.

Proposition A.6. *Suppose $\lambda > 0$ and $\operatorname{null}(Z) = \{0\}$. Then the solution $\hat{\beta}$ to the block lasso problem (1.2) with $\alpha = \infty$ is uniquely given by*

$$\hat{\beta}_{(-\mathcal{E})} = 0, \qquad \hat{\zeta} = (Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}\hat{y}, \quad \text{and} \quad \hat{\beta}_{\mathcal{K}_j} = \hat{\gamma}_j \operatorname{sgn}(S_{\mathcal{K}_j}) \text{ for } j \in \mathcal{E}.$$

Furthermore, $|G(\hat{\beta})| \le |\mathcal{E}| \le n \wedge J$ and $|\mathcal{E}| + \sum_{j \in \mathcal{E}} |\mathcal{K}_j^c| \le n \wedge p$.

A.5. Solution uniqueness of Equation (5.13)

Without loss of generality, let $\lambda = 1$. Due to the convexity of $L(\hat{\beta}(\sigma), \sigma)$ in σ, the solution set of (5.13) can always be written as an interval $[\sigma_1, \sigma_2]$, which reduces to a single point when $\sigma_1 = \sigma_2$. Denote by $[\sigma_1(y), \sigma_2(y)]$ the solution set for y.

Suppose the solution to (5.13) is not unique for some $y^* \in \mathbb{R}^n$, so that $\sigma_2(y^*) > \sigma_1(y^*) > 0$. Let us assume that the mapping $y \mapsto (\sigma_1(y), \sigma_2(y)) \in \mathbb{R}^2$ is continuous at y^*. Then, choosing a sufficiently small ball $B(\delta)$ centered at y^* with radius $\delta > 0$, we can find a

$$\sigma^* \in \bigcap_{y \in B(\delta)} [\sigma_1(y), \sigma_2(y)]$$

such that $G(\hat{\beta}(\sigma^*)) = A \subset \mathbb{N}_J$ for all $y \in B(\delta)$. Recall that $\hat{\beta}(\sigma^*) = \hat{\beta}(y, \sigma^*)$ is the block lasso (1.2) with tuning parameter $\lambda = \sigma^*$. It follows from the KKT conditions (2.3) and Assumption 1 that

$$y - X\hat{\beta}(\sigma^*) = n\sigma^*(X^\mathsf{T})^+ WS,$$

which with (5.13) implies that $\sqrt{n}\|(X^\mathsf{T})^+ WS\| = 1$ for all $y \in B(\delta)$. Clearly, the set

$$\Phi = \{s \in \mathcal{M}_A : \sqrt{n}\|(X^\mathsf{T})^+ Ws\| = 1\}$$

has measure zero in $\mathcal{M}_A$ (3.1). Thus, by Theorem 3.3, $\mathbb{P}(S \in \Phi, \mathcal{A} = A) = 0$. Regarding $y = X\beta_0 + \varepsilon$ as a random vector, we have

$$\mathbb{P}(y \in B(\delta)) \le \mathbb{P}(S \in \Phi, \mathcal{A} = A) = 0.$$

This apparently will lead to a contradiction, as long as the density of y is positive over $B(\delta)$. This argument leads to the following result:

Proposition A.7. *Suppose that $\lambda > 0$, ε has a positive density on $\mathbb{R}^n$, and Assumptions 1 and 2 hold. If the mapping $y \mapsto (\sigma_1(y), \sigma_2(y))$ is continuous at y^*, then the solution to (5.13) is unique for $y = y^*$.*

The continuity of $\sigma_1(y)$ and $\sigma_2(y)$ needs further verification, which may be technical and is left as future work. If $(\sigma_1(y), \sigma_2(y))$ is continuous on $\mathbb{R}^n$, then (5.13) has a unique solution for all y.

Appendix B: Remaining proofs

B.1. Proof of Lemma 3.2

Consider the equality constraints on s that are involved in the definition of $\mathcal{M}_A$ (3.1). Let $Q = Q(X) \in \mathbb{R}^{p\times(p-n)}$ be a matrix whose columns form a basis for $\mathcal{V}^\perp = \text{null}(XW^{-1})$. By Assumption 1, $\text{rank}(Q) = p-n$. The equality constraints on s are

$$Q^\mathsf{T} s = 0, \tag{B.1}$$

$$\|s_{(j)}\|_{\alpha^*} = 1 \quad \forall j \in A, \tag{B.2}$$

where (B.1) is equivalent to $s \in \mathcal{V}$. These $(p - n + |A|)$ independent equality constraints define the interior of $\mathcal{M}_A$ as a differentiable manifold of dimension $(n - |A|)$. Consequently, every interior point $s \in \mathcal{M}_A$ has a neighborhood that can be parameterized by s_F for some $F = F(s, A) \subset \mathbb{N}_p$ with $|F| = n - |A|$. Then there exists a matrix $T = T(s, A) \in \mathbb{R}^{p\times(n-|A|)}$ of rank $n - |A|$ such that $ds = T(s, A)ds_F$ in this neighborhood.

Fixing $\mathcal{A} = A$ in (3.4) leads to the differentiation of H_A:

$$dH_A = \sum_{j\in A} \left[(dr_j)\Psi_{(j)}\eta(s_{(j)}) + r_j\Psi_{(j)}d\eta(s_{(j)})\right] + \lambda W ds.$$

Since $d\eta(s_{(j)}) = D_{(jj)}ds_{(j)}$ for $j \in A$, we arrive at

$$dH_A = (\Psi \circ \eta)_A dr_A + \{(r \circ \Psi)D + \lambda W\}ds. \tag{B.3}$$

Plugging $ds = Tds_F$ into the above and letting $\theta = (r_A, s_F) \in \mathbb{R}^n$ complete the proof.

From (B.1) and (B.2), we have the following constraints on ds:

$$Q^\mathsf{T} ds = 0, \tag{B.4}$$

$$\langle \eta(s_{(j)}), ds_{(j)} \rangle = 0 \quad \forall j \in A. \tag{B.5}$$

These linear equations can be used to find the matrix $T(s, A)$ explicilty.

B.2. Proof of Theorem 3.3

Put $v = \tilde{H}_A(r_A, s)$ for $A \in \mathscr{A}$ and $(r_A, s) \in \Omega_A$. The differential of $\tilde{H}_A$ leads to

$$dv = \sqrt{n}(X^\mathsf{T})^+ M(r_A, s, A; \lambda)d\theta, \tag{B.6}$$

where M and $\theta = (r_A, s_F)$ are as in (3.8). Let $\Phi \subset \mathcal{M}_A$ be a neighborhood of s with parameter domain Δ, i.e. $\Phi = \{\Phi(s_F) : s_F \in \Delta \subset \mathbb{R}^k\}$, where $k = n - |A|$. Suppose that $\Gamma \subset (\mathbb{R}^+)^{|A|}$ is open and contains r_A. Denote by $V \subset \mathbb{R}^n$ the image of $\Gamma \times \Phi$ under $\tilde{H}_A$. To establish the n-form in (3.12), it is sufficient to show (3.14) for $u = s_F$. The bijective nature of $\tilde{H}$ (3.10) under Assumption 2 implies that

$$\mathbb{P}(\hat{\gamma}_A \in \Gamma, S \in \Phi, \mathcal{A} = A) = \mathbb{P}(\varepsilon/\sqrt{n} \in V) = \int_V g_n(v)dv.$$

Applying a change of variable in differential forms, we arrive at

$$\begin{aligned}\int_V g_n(v)dv &= \int_{\Gamma\times\Delta} g_n\left(\tilde{H}_A(r_A, \Phi(s_F))\right)\left|\frac{\partial v}{\partial \theta}\right| d\theta \\ &= \int_{\Gamma\times\Delta} f_A(r_A, \Phi(s_F))d\theta,\end{aligned}$$

where the Jacobian is determined by (B.6) and f_A is defined in (3.13). This completes the proof.

B.3. Proof of Corollary 3.4

If $\varepsilon \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$, then $g_n(v) = \phi_n(v; \sigma^2/n)$. It follows from (3.4) and (3.10) that

$$\tilde{H}_A(r_A, s) = \sqrt{n}(X^\mathsf{T})^+(\Psi b + \lambda W s - \Psi\beta_0).$$

Since $\Psi^+ = nX^+(XX^\mathsf{T})^{-1}X$ by Assumption 1, we have

$$\Psi\Psi^+ = X^\mathsf{T}(XX^\mathsf{T})^{-1}X = P_{X^\mathsf{T}},$$

which is the projection onto row(X). Thus $Ws = \Psi\Psi^+Ws$, since $Ws \in \text{row}(X)$. Putting this together with the identity $(X^\mathsf{T})^+\Psi = X/n$, we have

$$\tilde{H}_A(r_A, s) = \frac{1}{\sqrt{n}} X(b + \lambda\Psi^+Ws - \beta_0),$$

and thus

$$g_n(\tilde{H}_A(r_A, s)) = \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|X(b + \lambda\Psi^+Ws - \beta_0)\|^2\right]. \quad \text{(B.7)}$$

Then (3.15) follows immediately.

B.4. Proof of Corollary 3.5

It is easy to see that Assumptions 1 and 2 hold trivially if $\text{rank}(X) = p < n$. Thus, by Lemma 3.1 the mapping H is bijective. In this case, $\text{row}(X) = \mathbb{R}^p$ and $\mathcal{V}^\perp = \{0\}$, which imply that the constraint (B.4) no longer exists. Therefore, $|F| = p - |A|$, $T(s, A)$ is $p \times (p - |A|)$, and $M(r_A, s, A; \lambda)$ is $p \times p$. Now, the desired result is established by the same arguments as in Appendix B.2 with $U = X^\mathsf{T}\varepsilon/n$ in place of $(\varepsilon/\sqrt{n})$ and H_A in place of $\tilde{H}_A$.

B.5. Proof of Lemma 4.1

First, the matrix Q that defines the constraint (B.1) does not depend on s. Second, the sphere $\mathbb{S}_{\alpha^*}^{p_j-1}$, $j \in A$, can be parameterized by $s_{(j)\setminus k}$ for almost every point on the sphere, where $k \in \mathcal{G}_j$ is chosen as the last component in the group. More specifically, we may parameterize the positive half of $\mathbb{S}_{\alpha^*}^{p_j-1}$ as

$$\mathbb{S}_{\alpha^*}^{p_j-1} \cap \{s_k > 0\} = \left\{\left(s_{(j)\setminus k}, \left[1 - \|s_{(j)\setminus k}\|_{\alpha^*}^{\alpha^*}\right]^{1/\alpha^*}\right) : s_{(j)\setminus k} \in \mathbb{B}_{\alpha^*}^{p_j-1}\right\},$$

and the negative half in a similar way, both using the variables indexed by $\mathcal{G}_j \setminus k$. Therefore, we can always choose $F(s, A) = F(A)$ to parameterize almost every point in $\mathcal{M}_A$.

B.6. Proof of Theorem 5.1

Put $R = Q^\mathsf{T}$ with the matrix Q as in (B.1) and let $I = K \cup \mathcal{G}_{A^c} \subset \mathbb{N}_p$. Any $s \in \mathcal{M}_{A,K}$ must satisfy the following equality constraints:

$$s_{K^c} = 0, \qquad R_I s_I = 0, \quad \text{and} \quad \|s_{K_j}\|_1 = 1 \quad \forall j \in A,$$

which in turn impose constraints on ds_I:

$$R_I ds_I = 0 \quad \text{and} \quad \langle \text{sgn}(s_{K_j}), ds_{K_j} \rangle = 0 \quad \forall j \in A.$$

Under Assumption 1, there are $p-n+|A|$ independent equality constraints on s_I in the above. Thus, ds_I has $q=n-|A|-|K^c|$ free coordinates and there is a matrix $T=T(s_I,A,K)$ so that

$$ds_I = T(s_I,A,K)ds_F, \tag{B.8}$$

where $T\in\mathbb{R}^{(p-|K^c|)\times q}$ is a rank q matrix and $F=F(s_I,A,K)\subset\mathbb{N}_p$ with $|F|=q$.

Starting from (2.4), we have

$$\begin{aligned} H(\hat\beta,S;\beta_0,\lambda) &= \Psi\hat\beta+\lambda WS-\Psi\beta_0 \\ &= \sum_{j\in\mathcal{A}}\left[\hat\gamma_j\Psi_{\mathcal{K}_j}\operatorname{sgn}(S_{\mathcal{K}_j})+\Psi_{\mathcal{K}_j^c}\hat\beta_{\mathcal{K}_j^c}\right]+\lambda W_{\mathcal{I}}S_{\mathcal{I}}-\Psi\beta_0, \end{aligned}$$

where the index set $\mathcal{I}=\mathcal{K}\cup\mathcal{G}_{\mathcal{A}^c}$. Recalling the definition of the matrix $\Psi\circ\operatorname{sgn}(s)$, we arrive at

$$H(\hat\beta,S)=[\Psi\circ\operatorname{sgn}(S)]_{\mathcal{A}}\hat\gamma_{\mathcal{A}}+\Psi_{\mathcal{K}^c}\hat\beta_{\mathcal{K}^c}+\lambda W_{\mathcal{I}}S_{\mathcal{I}}-\Psi\beta_0,$$

where we have used the fact that $\operatorname{sgn}(S_{\mathcal{K}_j^c})=0$ for $j\in\mathcal{A}$. Denote the value of $\hat\beta$ by $b\in\mathbb{R}^p$. Fixing $\mathcal{A}=A$ and $\mathcal{K}=K$, the differentiation of $H_{A,K}$ at $(r_A,b_{K^c},s)\in\Omega_{A,K}$ is

$$\begin{aligned} dH_{A,K} &= [\Psi\circ\operatorname{sgn}(s)]_A dr_A+\Psi_{K^c}db_{K^c}+\lambda W_I ds_I \\ &= [\Psi\circ\operatorname{sgn}(s)]_A dr_A+\Psi_{K^c}db_{K^c}+\lambda W_I T ds_F = M(s,A,K;\lambda)d\theta, \end{aligned}$$

by plugging in (B.8) for ds_I and putting $\theta=(r_A,b_{K^c},s_F)\in\mathbb{R}^n$. Then the Jacobian of the mapping $\tilde H_{A,K}=\sqrt{n}(X^{\mathsf T})^{+}H_{A,K}$ is given by (5.4). Similar to Lemma 3.1, H and hence $\tilde H$ are bijections. Now following the proof of Theorem 3.3 leads to the desired joint density.

Appendix C: Technical details in the examples

C.1. Results and derivations in Example 1

We present the complete results for Example 1 in this appendix. The densities are given by the following differential forms:

$$f_{\varnothing}ds_1ds_2=\frac{\lambda^2}{\pi\sigma^2}\exp\left[-\frac{\lambda^2(s_1^2+s_2^2)}{\sigma^2}\right]ds_1ds_2, \tag{C.1}$$

$$f_{\{1\}}dr_1ds_1=\frac{1}{\pi\sigma^2}\exp\left[-\frac{(r_1+\lambda)^2}{\sigma^2}\right]\frac{r_1+\lambda}{|s_2|}dr_1ds_1, \tag{C.2}$$

$$f_{\{2\}}dr_2ds_1=\frac{2\lambda}{\pi\sigma^2}\exp\left[-\frac{2r_2(r_2+\lambda)}{\sigma^2}\right]\exp\left[-\frac{\lambda^2(s_1^2+s_2^2)}{\sigma^2}\right]dr_2ds_1, \tag{C.3}$$

$$f_{\{1,2\}}dr_1dr_2=\frac{1}{\pi\sigma^2}\exp\left[-\frac{(r_1+r_2+\lambda)^2+r_2^2}{\sigma^2}\right]dr_1dr_2, \tag{C.4}$$

where $r_j > 0$ for $j \in A$ and $s_{(1)} = (s_1, s_2) \in \mathbb{D}_A$ for $A = \varnothing, \{1\}, \{2\}, \{1,2\}$. Special care must be taken when integrating over the parameter domains of these densities. For example,

$$(r_1, r_2, s_{(1)}) \in (\mathbb{R}^+)^2 \times \{a, b, c, d\} := \Theta \quad \text{for } A = \{1,2\},$$

and therefore

$$\mathbb{P}(\mathcal{A} = \{1,2\}) = \int_\Theta d\mu_{\{1,2\}} = 4 \int_0^\infty \int_0^\infty f_{\{1,2\}}(r_1, r_2) dr_1 dr_2.$$

Next, we derive these results and find $\mathbb{P}(\mathcal{A} = A)$. A few pre-calculations are in order:

$$\sqrt{n}(X^\mathsf{T})^+ = \frac{1}{3}\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{bmatrix}, \qquad \Psi = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

which lead to

$$r \circ \Psi = \begin{bmatrix} r_1 & 0 & r_2 \\ 0 & r_1 & r_2 \\ r_1 & r_1 & 2r_2 \end{bmatrix}, \qquad \Psi \circ s = \begin{bmatrix} s_1 & s_3 \\ s_2 & s_3 \\ s_3 & 2s_3 \end{bmatrix}.$$

The density function $g_n(v) = \phi_2(v; \sigma^2/2)$ and, with (3.17),

$$\tilde{H}_A(r_A, s) = \sqrt{n}(X^\mathsf{T})^+ \left[\sum_{j \in A} r_j (\Psi \circ s)_j + \lambda s \right] = (r_1 + \lambda) s_{(1)} + r_2 s_3 \mathbf{1},$$

where $\mathbf{1} = (1,1)$ is a (column) vector of ones and $r_j = 0$ if $j \notin A$. Constraint (3.17) implies that

$$ds_3 = ds_1 + ds_2. \tag{C.5}$$

We will first go through the calculations for $A = \{1\}$ and then move to the other three cases. In what follows, let $\tau = \sqrt{2}\lambda/\sigma$ and $Z = (Z_1, Z_2) \sim \mathcal{N}_2(0, \mathbf{I}_2)$.

Case 1: $A = \{1\}$, $r_1 > 0$ and $r_2 = 0$. Combining constraints (C.5) and $\|s_{(1)}\|^2 = 1$ leads to

$$ds = \begin{bmatrix} 1 \\ -s_1/s_2 \\ 1 - (s_1/s_2) \end{bmatrix} ds_1 \Rightarrow T(s, \{1\}) = \begin{bmatrix} 1 \\ -s_1/s_2 \\ 1 - (s_1/s_2) \end{bmatrix}.$$

Plugging in $r_2 = 0$ and $s_3 = s_1 + s_2$, it is then easy to verify that

$$\begin{aligned} M(r_1, s, \{1\}) &= [s \mid (r_1 + \lambda)T], \\ J_{\{1\}}(r_1, s) &= -(r_1 + \lambda)/s_2, \\ \tilde{H}_{\{1\}}(r_1, s) &= (r_1 + \lambda) s_{(1)}. \end{aligned}$$

Consequently, we obtain the density as in (C.2). The two arcs in $\mathbb{D}_{\{1\}}$ (its interior) are parameterized as

$$\partial(b,c) = \left\{ \left(s_1, (1-s_1^2)^{1/2} \right) : -1 < s_1 < 0 \right\},$$
$$\partial(d,a) = \left\{ \left(s_1, -(1-s_1^2)^{1/2} \right) : 0 < s_1 < 1 \right\}.$$

It then follows that

$$\begin{aligned}
\mathbb{P}(\mathcal{A} = \{1\}) &= \int_{-1}^{0} \int_{0}^{\infty} f_{\{1\}} \left(r_1, s_1, (1-s_1^2)^{1/2} \right) dr_1 ds_1 \\
&\quad + \int_{0}^{1} \int_{0}^{\infty} f_{\{1\}} \left(r_1, s_1, -(1-s_1^2)^{1/2} \right) dr_1 ds_1 \\
&= \frac{1}{\pi\sigma^2} \left\{ \int_0^\infty \exp\left[-\frac{(r_1+\lambda)^2}{\sigma^2} \right] (r_1+\lambda) dr_1 \right\} \left\{ 2 \int_0^1 \frac{1}{\sqrt{1-s_1^2}} ds_1 \right\} \\
&= \frac{1}{2} e^{-\lambda^2/\sigma^2} = \frac{1}{2} \mathbb{P}(\|Z\| \geq \tau). \qquad \text{(C.6)}
\end{aligned}$$

Case 2: $A = \varnothing$, $r_1 = r_2 = 0$ and $s_{(1)} \in \mathbb{D}_{\varnothing}$. By (C.5), we have

$$T(s, \varnothing) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix},$$

which in combination with $r_1 = r_2 = 0$ leads to the following intermediate results:

$$M(s, \varnothing) = \lambda T(s, \varnothing), \quad J_{\varnothing}(s) = \lambda^2, \quad \tilde{H}_{\varnothing}(s) = \lambda s_{(1)}.$$

Then the density $f_{\varnothing}$ is obtained immediately as in (C.1) and

$$\mathbb{P}(\mathcal{A} = \varnothing) = \int_{\mathbb{D}_{\varnothing}} f_{\varnothing}(s_{(1)}) ds_1 ds_2 = \mathbb{P}(Z \in \tau \mathbb{D}_{\varnothing}). \qquad \text{(C.7)}$$

Case 3: $A = \{2\}$, $r_1 = 0$ and $r_2 > 0$. The interior of $\mathbb{D}_{\{2\}}$ is parameterized as

$$\partial(a,b) = \{ (s_1, 1-s_1) : 0 < s_1 < 1 \},$$
$$\partial(c,d) = \{ (s_1, -(1+s_1)) : -1 < s_1 < 0 \}.$$

Since $s_3 = s_1 + s_2 \in \{1, -1\}$ in this case, we have $ds_3 = ds_2 + ds_1 = 0$ and thus

$$T(s, \{2\}) = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \quad \text{and} \quad M(s, \{2\}) = \begin{bmatrix} s_3 & \lambda \\ s_3 & -\lambda \\ 2s_3 & 0 \end{bmatrix},$$

using the fact that $r_1 = 0$. Now straightforward calculations give

$$J_{\{2\}}(s) = -2\lambda s_3, \quad \tilde{H}_{\{2\}}(r_2, s) = \lambda s_{(1)} + r_2 s_3 \mathbf{1}.$$

Substituting s_3 by $s_1 + s_2$ with the fact that $|s_3| = 1$ leads to the density in (C.3). Consequently,

$$\mathbb{P}(\mathcal{A} = \{2\}) = \frac{2\lambda}{\pi\sigma^2}\left\{\int_0^\infty \exp\left[-\frac{2r_2(r_2+\lambda)}{\sigma^2}\right] dr_2\right\}\left\{2\int_0^1 \exp\left[-\frac{\lambda^2(s_1^2+(1-s_1)^2)}{\sigma^2}\right] ds_1\right\},$$

utilizing the symmetry of $(s_1^2 + s_2^2)$ between $\partial(a,b)$ and $\partial(c,d)$. The second integral

$$\int_0^1 \exp\left[-\frac{\lambda^2(s_1^2+(1-s_1)^2)}{\sigma^2}\right] ds_1 = \frac{\sqrt{2\pi\sigma^2}}{2\lambda}\exp\left(-\frac{\lambda^2}{2\sigma^2}\right)\mathbb{P}(|Z_2| \le \lambda/\sigma).$$

After completing the first integral, we have

$$\begin{aligned}\mathbb{P}(\mathcal{A} = \{2\}) &= 2\cdot\mathbb{P}(Z_1 \ge \lambda/\sigma \text{ and } |Z_2| \le \lambda/\sigma)\\ &= 2\cdot\mathbb{P}(Z_1 + Z_2 \ge \tau \text{ and } |Z_1 - Z_2| \le \tau).\end{aligned} \tag{C.8}$$

Case 4: $A = \{1,2\}$, $r_1, r_2 > 0$ and $\mathbb{D}_{\{1,2\}} = \{a,b,c,d\}$. Since $|A| = n = 2$,

$$M(r,s,\{1,2\}) = \Psi \circ s \quad \text{and} \quad J_{\{1,2\}}(r,s) = s_1^2 - s_2^2.$$

It is easy to see that $|J_{\{1,2\}}| = 1$ for all $s_{(1)} \in \{a,b,c,d\}$ and

$$\tilde{H}_{\{1,2\}}(r,s) = (r_1+\lambda)s_{(1)} + r_2 s_3 \mathbf{1}.$$

Then we obtain the density $f_{\{1,2\}}$ in (C.4) immediately, which leads to

$$\begin{aligned}\mathbb{P}(\mathcal{A} = \{1,2\}) &= 4\int_0^\infty\int_0^\infty \frac{1}{\pi\sigma^2}\exp\left[-\frac{(r_1+r_2+\lambda)^2 + r_2^2}{\sigma^2}\right] dr_1 dr_2\\ &= 4\cdot\mathbb{P}(Z_1 \ge 0 \text{ and } Z_2 - Z_1 \ge \tau).\end{aligned} \tag{C.9}$$

Finally, by (C.6), (C.7), (C.8), and (C.9) one can easily verify that

$$\sum_A \mathbb{P}(\mathcal{A} = A) = \mathbb{P}(Z \in \mathbb{R}^2) = 1.$$

C.2. Derivations in Example 2

This section is divided into three parts:

Part 1: Derivation of (3.22). For $v \in \mathbb{R}^p$ and $j \in \mathbb{N}_J$, define $u = v_{\langle j\rangle} \in \mathbb{R}^p$ so that $u_k = v_k$ for $k \in \mathcal{G}_j$ and $u_k = 0$ otherwise. Let $b \in \mathbb{R}^p$ denote the value for $\hat{\beta}$, i.e. $b_{(j)} = r_j s_{(j)}$ for $j \in \mathbb{N}_J$. Straightforward algebra leads to:

$$\begin{aligned}H_A(r_A, s) &= b + \lambda\sqrt{m}s - \beta_0,\\ \Psi \circ s &= \left[s_{\langle 1\rangle}|\ldots|s_{\langle J\rangle}\right] \in \mathbb{R}^{p\times J},\\ r\circ\Psi + \lambda W &= \operatorname{diag}\left\{(r_j + \lambda\sqrt{m})\mathbf{I}_m : j \in \mathbb{N}_J\right\} \in \mathbb{R}^{p\times p}.\end{aligned}$$

Since $\mathrm{row}(X) = \mathbb{R}^p$, the constraint (B.4) disappears. For $j \in A$, choose $k(j) \in \mathcal{G}_j$ such that $s_{k(j)} \neq 0$ and put $F(j) = \mathcal{G}_j \setminus k(j)$. Then constraint (B.5) can be written as

$$ds_{k(j)} = -\frac{1}{s_{k(j)}}\langle s_{F(j)}, ds_{F(j)}\rangle \quad \text{for } j \in A.$$

Without loss of generality, assume that $k(j) = m \cdot j$ is chosen to be the last component in the group. The matrix $T = T(s, A)$ has a block-diagonal structure and its j^{th} block

$$T(j) = \begin{bmatrix} \mathbf{I}_{m-1} \\ -s_{F(j)}^{\mathsf{T}}/s_{k(j)} \end{bmatrix} \text{ for } j \in A \quad \text{and} \quad T(j) = \mathbf{I}_m \text{ for } j \notin A.$$

It follows immediately that

$$(r \circ \Psi + \lambda W)T(s, A) = \mathrm{diag}\left\{(r_j + \lambda\sqrt{m})T(j) : j \in \mathbb{N}_J\right\}.$$

Permuting the columns of M (3.8) to put $s_{\langle j\rangle}$, $j \in A$, to the right of the j^{th} block of the above matrix, M is also seen to be block-diagonal with each block $M_{(jj)}$ of size $m \times m$. For $j \in A$, the j^{th} block

$$M_{(jj)} = \left[(r_j + \lambda\sqrt{m})T(j) \mid s_{(j)}\right],$$

and for $j \notin A$, since $r_j = 0$,

$$M_{(jj)} = (\lambda\sqrt{m})\mathbf{I}_m.$$

Simple calculation with $\|s_{(j)}\|^2 = 1$ for $j \in A$ shows that

$$|\det M_{(jj)}| = \begin{cases} (r_j + \lambda\sqrt{m})^{m-1}/|s_{k(j)}| & j \in A, \\ (\lambda\sqrt{m})^m & j \notin A. \end{cases} \tag{C.10}$$

Under the hypotheses, $\sqrt{n}(X^{\mathsf{T}})^+ = X/\sqrt{n}$ is an orthogonal matrix whose determinant is ± 1. Consequently, the Jacobian (3.11) is

$$|J_A| = \prod_{j=1}^{J} |\det M_{(jj)}|. \tag{C.11}$$

Plugging (C.11) into (3.15) with $\Psi = \mathbf{I}_p$, we obtain the differential form in (3.22).

Part 2: Derivation of the marginal density of $\hat{\gamma}_j$ (3.24) for $j \in A$, assuming $\beta_{0(j)} = 0$. Let $d\mu_j = f_j(r_j, s_{(j)})d\theta_{(j)}$, which specifies the joint distribution of $\hat{\gamma}_j$ and $S_{(j)}$. We start from the integral

$$\begin{aligned} f_j(r_j)dr_j &= \int_{\mathbb{S}^{m-1}} d\mu_j \\ &= C(m)(2\pi\sigma^2/n)^{-\frac{m}{2}}(r_j + \lambda\sqrt{m})^{m-1}\exp\left[-\frac{n}{2\sigma^2}(r_j + \lambda\sqrt{m})^2\right]dr_j, \end{aligned}$$

where $C(m) > 0$ is a constant:

$$C(m) = \int_{\mathbb{S}^{m-1}} \frac{1}{|s_{k(j)}|} ds_{F(j)} = 2 \int_{\mathbb{B}^{m-1}} \left(1 - \|v\|^2\right)^{-1/2} dv.$$

With a change of variable, $v = x/\sqrt{1 + \|x\|^2}$,

$$C(m) = 2 \int_{\mathbb{R}^{m-1}} \left(1 + \|x\|^2\right)^{-m/2} dx = \frac{2 \cdot \pi^{m/2}}{\Gamma(m/2)},$$

by the normalizing constant of the multivariate t-distribution with one degree of freedom.

Part 3: Proof of (3.25). The distribution of $\tilde{\beta}_{(j)}$ implies $(n/\sigma^2)\|\tilde{\beta}_{(j)}\|^2 \overset{d}{=} \chi^2_m$ follows a χ^2-distribution with m degrees of freedom. Letting $z = (r_j + \lambda\sqrt{m})^2$, we have

$$\begin{aligned}
\int_t^\infty f_j(r_j) dr_j &= \int_{(t+\lambda\sqrt{m})^2}^\infty \frac{\left(n/\sigma^2\right)^{\frac{m}{2}}}{2^{m/2} \cdot \Gamma(m/2)} z^{m/2-1} \exp\left(-\frac{n}{2\sigma^2} z\right) dz \\
&= \mathbb{P}\left\{(\sigma^2/n)\chi^2_m > (t + \lambda\sqrt{m})^2\right\} \\
&= \mathbb{P}\left\{\|\tilde{\beta}_{(j)}\|^2 > (t + \lambda\sqrt{m})^2\right\},
\end{aligned}$$

which completes the proof.

C.3. Derivations in Example 3

We note that the constraint (B.5) reduces to $ds_j = 0$ for $j \in A$, which implies that $T_{A\bullet} = \mathbf{0}$ as in property (ii). Recall that $B = \mathbb{N}_p \setminus A$. The constraint imposed on ds_B comes from (B.4) and is thus independent of s, hence property (i). As a consequence, the set of free coordinates of s is always a subset of B, i.e. $F \subset B$, and $|F| = n - |A|$. Since $r_B = 0$ by definition, $(r \circ \Psi)_j = 0$ for all $j \in B$. It follows that $(r \circ \Psi)T = \mathbf{0}$ and thus, as defined in (3.9),

$$M(r_A, s, A) = [(\Psi \circ s)_A \mid \lambda W T] = [(\Psi \circ s)_A \mid \lambda W_B T_{B\bullet}].$$

Since $|s_j| = 1$ for $j \in A$ and $p_j = 1$,

$$\begin{aligned}
|J_A(r_A, s)| &= \left|\det\left\{\sqrt{n}(X^\mathsf{T})^+ [(\Psi \circ s)_A \mid \lambda W_B T_{B\bullet}]\right\}\right| \\
&= \left|\det\left\{\sqrt{n}(X^\mathsf{T})^+ [\Psi_A \mid \lambda W_B T_{B\bullet}]\right\}\right|.
\end{aligned}$$

Substituting this into (3.12) gives the density in (3.26).

To compare (3.26) with Theorem 2 in [29], we apply the following change of variable: Let $b_j = s_j r_j$ denote the value for $\hat{\beta}_j$ for $j \in A$. Plugging into (3.26) that $r_j = |b_j|$, $s_j = \operatorname{sgn}(b_j)$ and $dr_j = s_j db_j$ for $j \in A$, we obtain the density for $(\hat{\beta}_\mathcal{A}, S_\mathcal{B}, \mathcal{A})$ parameterized by (b_A, s_F):

$$g_n(\tilde{H}_A(|b_A|, s)) \left|\det\left\{\sqrt{n}(X^\mathsf{T})^+ [\Psi_A \mid \lambda W_B T_{B\bullet}]\right\}\right| db_A ds_F, \tag{C.12}$$

where we have again used $|s_j| = 1$ for $j \in A$ in the change of the volume elements.

References

[1] Antoniadis, A. (2010). Comments on: ℓ_1-penalization for mixture regression models. *TEST* **19**, 2, 257–258. MR2677723

[2] Bunea, F., Lederer, J., and She, Y. (2014). The group square-root lasso: theoretical properties and fast algorithms. *IEEE Trans. Inform. Theory* **60**, 2, 1313–1325. MR3164977

[3] Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41**, 3, 1232–1259. MR3113809

[4] Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and `r`-software `hdi`. *Statist. Sci.* **30**, 4, 533–558. MR3432840

[5] Dezeure, R., Bühlmann, P., and Zhang, C. (2016). High-dimensional simultaneous inference with the bootstrap. *Preprint*, arXiv:1606.03940.

[6] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–2909. MR3277152

[7] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 3, 907–927. MR3485948

[8] Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 2, 413–468. MR3210970

[9] Meinshausen, N. (2015). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77**, 5, 923–945. MR3414134

[10] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72**, 4, 417–473. MR2758523

[11] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104**, 488, 1671–1681. MR2750584

[12] Mitra, R. and Zhang, C.-H. (2016). The benefit of group sparsity in group inference with de-biased scaled group Lasso. *Electron. J. Stat.* **10**, 2, 1829–1873. MR3522662

[13] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27**, 4, 538–557. MR3025133

[14] Negahban, S. N. and Wainwright, M. J. (2011). Simultaneous support recovery in high dimensions: benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Trans. Inform. Theory* **57**, 6, 3841–3863. MR2817058

[15] Neykov, M., Ning, Y., Liu, J., and Liu, H. (2015). A unified theory of confidence regions and testing for high dimensional estimating equations. *Preprint*, arXiv:1510.08986.

[16] Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45**, 1, 158–195. MR3611489

[17] Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear model: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine Learning*.

[18] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 4, 879–898. MR2999166

[19] Tian Harris, X., Panigrahi, S., Markovic, J., Bi, N., and Taylor, J. (2016). Selective sampling after solving a convex problem. *Preprint*, arXiv:1609.05609v1.

[20] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 1, 267–288. MR1379242

[21] Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7**, 1456–1490. MR3066375

[22] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 3, 1166–1202. MR3224285

[23] van de Geer, S. and Stucky, B. (2016). χ^2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data*. Abel Symp., Vol. **11**. Springer, Cham, 279–306. MR3616273

[24] Voorman, A., Shojaie, A., and Witten, D. (2014). Inference in high dimensions with the penalized score test. *Preprint*, arXiv:1401.2678.

[25] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 5A, 2178–2201. MR2543689

[26] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 1, 49–67. MR2212574

[27] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**, 1, 217–242. MR3153940

[28] Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112**, 518, 757–768. MR3671768

[29] Zhou, Q. (2014). Monte carlo simulation for Lasso-type problems by estimator augmentation. *J. Amer. Statist. Assoc.* **109**, 508, 1495–1516. MR3293606

[30] Zhou, Q. and Min, S. (2017). Uncertainty quantification under group sparsity. *Biometrika* **104**, doi: 10.1093/biomet/asx037.