Electronic Journal of Statistics
Vol. 16 (2022) 1–52
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2020

Random weighting in LASSO regression*

Tun Lee Ng and Michael A. Newton

Department of Statistics,
1300 University Ave, Madison WI 53706
e-mail: tunlee@stat.wisc.edu; newton@stat.wisc.edu

Abstract: We establish statistical properties of random-weighting methods in LASSO regression under different regularization parameters λ_n and suitable regularity conditions. The random-weighting methods in view concern repeated optimization of a randomized objective function, motivated by the need for computationally efficient uncertainty quantification in contemporary estimation settings. In the context of LASSO regression, we repeatedly assign analyst-drawn random weights to terms in the objective function, and optimize to obtain a sample of random-weighting estimators. We show that existing approaches have conditional model selection consistency and conditional asymptotic normality at different growth rates of λ_n as $n \to \infty$. We propose an extension to the available random-weighting methods and establish that the resulting samples attain conditional sparse normality and conditional consistency in a growing-dimension setting. We illustrate the proposed methodology using synthetic and benchmark data sets, and we discuss the relationship of the results to approximate nonparametric Bayesian analysis and to perturbation bootstrap methods.

MSC2020 subject classifications: 62F12, 62F40, 62F15.
Keywords and phrases: Random weights, weighted likelihood bootstrap, weighted Bayesian bootstrap, LASSO, bootstrap, perturbation bootstrap, consistency, model selection consistency.

Received February 2021.

Contents

1 Introduction 2
2 Problem setup 4
3 Main results 7
3.1 One-step procedure 7
3.2 Two-step procedure 12
3.3 Remarks 15
4 Numerical experiments 16
4.1 Simulation 16
4.2 Benchmark data example 22
5 Discussion 24

*TLN and MAN were supported in part by the University of Wisconsin Institute for the Foundations of Data Science through grants from the US National Science Foundation (1740707, 2023239).

A Appendix A 26
Acknowledgments 50
References 50

1. Introduction

Consider the well-studied linear regression model with fixed design

$$\boldsymbol{Y} = \beta_\mu \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ is the response vector, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, $X \in \mathbb{R}^{n \times p_n}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is the vector of independent and identically distributed (i.i.d.) random errors with mean 0 and variance σ_ϵ^2. Without loss of generality, we assume that the columns of X are centered, and take $\widehat{\beta}_\mu = \bar{Y}$, in which case we can replace $\boldsymbol{Y}$ in (1.1) with $\boldsymbol{Y} - \bar{Y}\mathbf{1}_n$, and concentrate on inference for $\boldsymbol{\beta}$. Again, without loss of generality, we also assume $\bar{Y} = 0$. Let $\boldsymbol{\beta_0} \in \mathbb{R}^{p_n}$ be the true model coefficients with q non-zero components, where $q \leq \min(p_n, n)$. Note that $\boldsymbol{Y}, X$ and $\boldsymbol{\epsilon}$ are all indexed by sample size n, but we omit the subscript whenever this does not cause confusion.

Recall, the LASSO estimator is given by

$$\widehat{\boldsymbol{\beta}}_n^{\text{LAS}} := \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j|, \tag{1.2}$$

for a scalar penalty λ_n (Tibshirani, 1996), where $\boldsymbol{x}_i'$ is the i^{th} row of X. The LASSO is a canonical example in the broad class of penalized inference procedures; for the purpose of uncertainty quantification in such models, Newton, Polson and Xu (2021) developed the random-weighting approach as a straightforward technique to leverage advances in optimization. They reported good performance in high-dimensional regression, trend-filtering and deep learning applications. In particular, their random-weighting version of (1.2) is

$$\widehat{\boldsymbol{\beta}}_n^{w} := \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} W_{0,j} |\beta_j| \right\}, \tag{1.3}$$

where the analyst first chooses a distribution F_W with $P(W > 0) = 1$ and $\mathbb{E}(W^4) < \infty$, and constructs $W_i \overset{iid}{\sim} F_W$ for all $i = 1, 2, \cdots, n$. The precise treatment of penalty-associated weights $\boldsymbol{W}_0 = (W_{0,1}, \cdots, W_{0,p_n})$ induces several random-weighting variations, the simplest of which has

$$W_{0,j} = 1 \ \forall \ j, \tag{1.4}$$

or the penalty terms all share a common random weight

$$W_{0,j} = W_0 \ \forall \ j, \ \text{where } (W_0, W_i) \overset{iid}{\sim} F_W \ \forall \ i, \tag{1.5}$$

and the most elaborate of which has all entries

$$(W_{0,j}, W_i) \overset{iid}{\sim} F_W \ \forall\ i, j. \tag{1.6}$$

Regardless of our treatment of the weights, (1.3) yields independent and identically distributed draws from the conditional distribution of $\widehat{\boldsymbol{\beta}}_n^w$ given data when we repeatedly realize weight vectors *in silico* by one of the random-weighting mechanisms. A computational benefit for uncertainty quantification is that random weighting is readily parallelized. Though useful inference tools already exist for LASSO regression (e.g., Friedman, Hastie and Tibshirani, 2010), we focus on this well-studied model in order to extend random-weighting theory and also to guide work for more complex settings where random weighting may be readily applied (Newton, Polson and Xu, 2021). In the present study we investigate the asymptotic properties of (1.3), with attention on properties of the conditional distribution given data. By allowing different rates of growth of the regularization parameter λ_n, and under suitable regularity conditions, we prove that the random-weighting method has the following properties:

- conditional model selection consistency (for both growing p_n and fixed p)
- conditional consistency (for fixed $p_n = p$)
- conditional asymptotic normality (for fixed $p_n = p$)

for all three weighting schemes (1.4), (1.5) and (1.6). We find there is no common λ_n that would allow random-weighting samples to have conditional sparse normality (i.e., simultaneously to enjoy conditional model selection consistency and to achieve conditional asymptotic normality on the true support of $\boldsymbol{\beta}$) even under fixed $p_n = p$ setting. Consequently, we propose an extension to the random-weighting framework (1.3) by adopting a two-step procedure in the optimization step as laid out in Algorithm 2. We prove that a common regularization rate λ_n allows two-step random-weighting samples to achieve conditional sparse normality and conditional consistency properties under growing p_n setting.

After setting regularity conditions and notation in Section 2, we report our main distributional results for random weighting in Section 3. Asymptotic techniques from Knight and Fu (2000), Zhao and Yu (2006) and Liu and Yu (2013) guide our calculations. Extensive simulations and application to a benchmark data set illustrate how two-step random weighting under schemes (1.4), (1.5) and (1.6) compares with both Bayesian and bootstrap methods for uncertainty quantification (Section 4). In Section 5 we comment on our findings in relation to the perturbation bootstrap (e.g., Das and Lahiri, 2019) and also to recent nonparametric Bayesian work that has renewed interest in the operating characteristics of random-weighting (Lyddon, Walker and Holmes, 2018; Lyddon, Holmes and Walker, 2019; Fong, Lyddon and Holmes, 2019). Detailed proofs are presented in Appendix A.

2. Problem setup

We assume throughout that the unknown number of truly relevant predictors, q, is fixed, that

$$\mathbb{E}(\epsilon_i^4) < \infty \ \forall \ i, \tag{2.1}$$

and all p_n predictors are bounded, i.e. $\exists\ M_1 > 0$ such that

$$|x_{ij}| \leq M_1 \quad \forall \quad i = 1, \ldots, n \ ; \ j = 1, \ldots, p_n, \tag{2.2}$$

where x_{ij} refers to the $(i,j)^{th}$ element of X.

Without loss of generality, we partition $\boldsymbol{\beta}_0$ into

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \boldsymbol{\beta}_{0(1)} \\ \boldsymbol{\beta}_{0(2)} \end{bmatrix},$$

where $\boldsymbol{\beta}_{0(1)}$ refers to the $q \times 1$ vector of non-zero true regression parameters, and $\boldsymbol{\beta}_{0(2)}$ is a $(p_n - q) \times 1$ zero vector. Similarly, we partition the columns of the design matrix X into

$$X = \begin{bmatrix} X_{(1)} & X_{(2)} \end{bmatrix}$$

which corresponds to $\boldsymbol{\beta}_{0(1)}$ and $\boldsymbol{\beta}_{0(2)}$ respectively.

We consider both fixed-dimensional ($p_n = p$) and growing-dimensional (p_n increases with n) settings. In the growing dimensional setting, we assume that for some $M_2 > 0$,

$$\boldsymbol{\alpha}' \left[\frac{X_{(1)}' X_{(1)}}{n} \right] \boldsymbol{\alpha} \geq M_2 \quad \forall \quad \|\boldsymbol{\alpha}\|_2 = 1. \tag{2.3}$$

Note that assumptions (2.2) and (2.3), coupled with the fact that q is fixed, ensure that $\frac{1}{n} X_{(1)}' X_{(1)}$ is invertible $\forall\ n$, a fact that we rely on in this paper.

Meanwhile, for fixed-dimensional ($p_n = p$) setting, we assume that $\text{rank}(X) = p$ and there exists a non-singular matrix C such that

$$\frac{1}{n} X'X = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \to C \quad \text{as } n \to \infty, \tag{2.4}$$

where $\boldsymbol{x}_i$ is the i^{th} row of the design matrix X.

Comments on assumptions: The fixed-q assumption is commonly found in Bayesian linear-model literature, such as Johnson and Rossell (2012), and Narisetty and He (2014). Since we intend to compare the random-weighting approach with posterior inference, we make the fixed-q assumption to align with existing Bayesian theory. The finite-moment assumption (2.1) of $\boldsymbol{\epsilon}$ is commonly found in literature (e.g., Camponovo, 2015; Das and Lahiri, 2019) is weaker than the normality assumption commonly specified under a Bayesian approach (e.g., Park and Casella, 2008; Johnson and Rossell, 2012; Narisetty and He, 2014).

Assumption (2.2) can also be found in some seminal papers, such as Zhao and Yu (2006) and Chatterjee and Lahiri (2011a), and in fact, can be (trivially) achieved by standardizing the covariates. Assumption (2.3) is equivalent to providing a lower bound to the minimum eigenvalue of $\frac{1}{n}X'_{(1)}X_{(1)}$. This eigenvalue assumption is very common in both frequentist and Bayesian literature, such as Zhao and Yu (2006) and Narisetty and He (2014). Finally, assumption (2.4) is common in the LASSO literature under fixed p setting, which can be traced back to Knight and Fu (2000) and Zhao and Yu (2006). This assumption basically explains the relationship between the predictors under a fixed design model, and can be interpreted as the direct counterpart to the variance-covariance matrix of X under a random design model. For the case of growing p_n, assumption (2.4) is no longer appropriate since the dimension of $\frac{1}{n}X'X$ grows.

Probability Space: There are two sources of variation in the random-weighting setup (1.3), namely the error terms $\boldsymbol{\epsilon}$ and the user-defined weights $\boldsymbol{W}$. In this paper, we consider a common probability space with the common probability measure $P = P_D \times P_W$, where P_D is the probability measure of the observed data $Y_1, Y_2, \cdots$, and P_W is the probability measure of the triangular array of random weights (e.g., Mason and Newton, 1992). The use of product measure reflects the independence of user-defined $\boldsymbol{W}$ and data-associated $\boldsymbol{\epsilon}$. We focus on the conditional probabilities given data, that is, given the sigma-field $\mathcal{F}_n$ generated by $\boldsymbol{\epsilon}$:

$$\mathcal{F}_n := \sigma(Y_1, \ldots, Y_n) = \sigma(\epsilon_1, \ldots, \epsilon_n).$$

The study of convergence of these conditional probabilities $P(\cdot\,|\mathcal{F}_n)$ under a weighted bootstrap framework is not new; see, for example, Mason and Newton (1992) and Lyddon, Holmes and Walker (2019). We now outline some definitions and notations in this respect.

Conditional Convergence Notations: Let random variables (or vectors) $U, V_1, V_2, \ldots$ be defined on $(\Omega, \mathcal{A})$. We say V_n converges in conditional probability *a.s.* P_D to U if for every $\delta > 0$,

$$P(\|V_n - U\| > \delta|\mathcal{F}_n) \to 0 \quad a.s.\ P_D$$

as $n \to \infty$. The notation *a.s.* P_D is read as *almost surely under* P_D, and means *for almost every infinite sequence of data* $Y_1, Y_2, \cdots$. For brevity, this convergence is denoted

$$V_n \xrightarrow{\text{c.p.}} U \quad a.s.\ P_D.$$

Similarly, we say V_n converges in conditional distribution *a.s.* P_D to U if for any Borel set $A \subset \mathbb{R}$,

$$P(V_n \in A|\mathcal{F}_n) \to P(U \in A) \quad a.s.\ P_D$$

as $n \to \infty$. For brevity, this convergence is denoted

$$V_n \xrightarrow{\text{c.d.}} U \quad a.s.\ P_D.$$

In addition, for random variables (or vectors) $V_1, V_2, \ldots$ and random variables $U_1, U_2, \ldots$, we say

$$V_n = O_p(U_n) \quad a.s.\ P_D$$

if and only if, for any $\delta > 0$, there is a constant $C_\delta > 0$ such that $a.s.\, P_D$,

$$\sup_n P\left(\|V_n\| \geq C_\delta |U_n| \ \middle| \mathcal{F}_n \right) < \delta;$$

whereas

$$V_n = o_p(U_n) \quad a.s.\ P_D$$

if and only if

$$\frac{V_n}{U_n} \xrightarrow{\text{c.p.}} 0 \quad a.s.\ P_D.$$

Other Notation: Following the usual convention, denote $\Phi\{.\}$ as the cumulative distribution function of the standard normal distribution. For two random variables U and V, the expression $U \perp V$ is read as "U is independent of V". Denote $\|\cdot\|_2$ and $\|\cdot\|_F$ as the l_2 norm and Frobenius norm respectively. Let $\mathbf{1}_k$ and I_k be $k \times 1$ vector of ones and $k \times k$ identity matrix respectively for some integer $k \geq 2$. Besides that, for any two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ of the same dimension, we denote $\boldsymbol{u} \circ \boldsymbol{v}$ as the Hadamard (entry-wise) product of the two vectors. In addition, define

$$\begin{bmatrix} C_{n(11)} & C_{n(12)} \\ C_{n(21)} & C_{n(22)} \end{bmatrix} := \frac{1}{n} X'X = \frac{1}{n} \begin{bmatrix} X'_{(1)}X_{(1)} & X'_{(1)}X_{(2)} \\ X'_{(2)}X_{(1)} & X'_{(2)}X_{(2)} \end{bmatrix}.$$

Notice that an immediate consequence of Assumption (2.4) is that

$$C_{n(ij)} \to C_{ij} \ \forall\ i, j = 1, 2,$$

where C_{11} is invertible. Furthermore, denote μ_W and σ_W^2 as the mean and variance of the random weight distribution F_W. Let $D_n = diag(W_1, \ldots, W_n)$, and define

$$\begin{bmatrix} C^w_{n(11)} & C^w_{n(12)} \\ C^w_{n(21)} & C^w_{n(22)} \end{bmatrix} := \frac{1}{n} X' D_n X = \frac{1}{n} \begin{bmatrix} X'_{(1)}D_nX_{(1)} & X'_{(1)}D_nX_{(2)} \\ X'_{(2)}D_nX_{(1)} & X'_{(2)}D_nX_{(2)} \end{bmatrix}.$$

Notice that D_n does not contain any penalty weights $W_{0,j}$. For weighting scheme (1.6), the penalty weights $\boldsymbol{W}_0 = (W_{0,1}, \cdots, W_{0,p_n})$ could also be partitioned into

$$\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{W}_{0(1)} \\ \boldsymbol{W}_{0(2)} \end{bmatrix},$$

which corresponds to the partition of $\boldsymbol{\beta}_0$. For ease of notation, define

$$\boldsymbol{Z}^w_{n(1)} = \frac{1}{\sqrt{n}} X'_{(1)} D_n \boldsymbol{\epsilon},$$

$$\boldsymbol{Z}^w_{n(2)} = \frac{1}{\sqrt{n}} X'_{(2)} D_n \boldsymbol{\epsilon},$$
$$\boldsymbol{Z}^w_{n(3)} = C_{n(21)} C^{-1}_{n(11)} \boldsymbol{Z}^w_{n(1)} - \boldsymbol{Z}^w_{n(2)},$$
$$\widetilde{C}^w_n = C^w_{n(21)} \left(C^w_{n(11)} \right)^{-1} - C_{n(21)} C^{-1}_{n(11)}.$$

Finally, the function $\mathrm{sgn}(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero. An estimator $\widehat{\boldsymbol{\beta}}$ is said to be equal in sign to the true parameter $\boldsymbol{\beta}_0$, if

$$\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}_0),$$

and is denoted as

$$\widehat{\boldsymbol{\beta}} \overset{s}{=} \boldsymbol{\beta}_0.$$

3. Main results

3.1. One-step procedure

We investigate the asymptotic properties of random-weighting draws (1.3) obtained from Algorithm 1, which coincides with the weighted Bayesian bootstrap method considered by Newton, Polson and Xu (2021). For convenience, we shall call this the "one-step procedure" to distinguish it from the extended framework that we shall discuss in Section 3.2.

Algorithm 1: Random-Weighting in LASSO regression

Input :
- data: $D = (\boldsymbol{y}, X)$
- regularization parameter: λ_n
- number of draws: B
- choice of random weight distribution: F_W
- choice of weighting schemes: (1.4), (1.5) or (1.6)

Output : B parameter samples $\{\widehat{\boldsymbol{\beta}}^{w,b}_n\}_{b=1}^B$

for $b = 1$ *to* B **do**
  Draw i.i.d. random weights from F_W and substitute them into (1.3) ;
  Store $\widehat{\boldsymbol{\beta}}^{w,b}_n$ obtained by optimizing (1.3) ;
end

First, we establish the property of conditional model selection given data. In particular, we are interested in the conditional probability of the random-weighting samples matching the signs of $\boldsymbol{\beta}_0$. Notably, sign consistency is stronger than variable selection consistency, which requires only matching of zeros. Nevertheless, we agree with Zhao and Yu (2006)'s argument of considering sign consistency – it allows us to avoid situations where models have matching zeroes but reversed signs, which hardly qualify as correct models. We begin with a result that establishes the lower bound for this conditional probability.

Proposition 3.1. *Suppose $p_n \leq n$ and $\text{rank}(X) = p_n$. Assume (2.1), (2.2) and (2.3). Furthermore, assume the* ***strong irrepresentable condition*** *(Zhao and Yu, 2006): there exists a positive constant vector $\boldsymbol{\eta}$ such that*

$$\left| C_{n(21)} \left(C_{n(11)} \right)^{-1} sgn\left(\boldsymbol{\beta}_{0(1)} \right) \right| \leq \mathbf{1}_{p_n - q} - \boldsymbol{\eta}, \tag{3.1}$$

where $0 < \eta_j \leq 1 \ \forall \ j = 1, \ldots, p_n - q$, and the inequality holds element-wise. Then, for all $n \geq p_n$,

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \stackrel{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n \right) \geq P\left(A_n^w \cap B_n^w \middle| \mathcal{F}_n \right),$$

where

(a) for weighting scheme (1.4),

$$A_n^w \equiv \left\{ \left| \left(C_{n(11)}^w \right)^{-1} \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n}{2\sqrt{n}} sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right) \right| \leq \sqrt{n} \left| \boldsymbol{\beta}_{0(1)} \right| \text{ element-wise} \right\}$$

$$B_n^w \equiv \left\{ \left| \widetilde{C}_n^w \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n}{2\sqrt{n}} sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right) + \boldsymbol{Z}_{n(3)}^w \right| \leq \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{\eta} \text{ element-wise} \right\};$$

(b) for weighting scheme (1.5),

$$A_n^w \equiv \left\{ \left| \left(C_{n(11)}^w \right)^{-1} \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n W_0}{2\sqrt{n}} sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right) \right| \leq \sqrt{n} \left| \boldsymbol{\beta}_{0(1)} \right| \text{ element-wise} \right\}$$

$$B_n^w \equiv \left\{ \left| \widetilde{C}_n^w \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n W_0}{2\sqrt{n}} sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right) + \boldsymbol{Z}_{n(3)}^w \right| \leq \frac{\lambda_n W_0}{2\sqrt{n}} \boldsymbol{\eta} \text{ element-wise} \right\};$$

(c) for weighting scheme (1.6),

$$A_n^w \equiv \left\{ \left| \left(C_{n(11)}^w \right)^{-1} \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{W}_{0(1)} \circ sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right) \right| \right.$$
$$\left. \leq \sqrt{n} \left| \boldsymbol{\beta}_{0(1)} \right| \text{ element-wise} \right\}$$

$$B_n^w \equiv \left\{ \left| \widetilde{C}_n^w \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{W}_{0(1)} \circ sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right) + \boldsymbol{Z}_{n(3)}^w \right| \right.$$
$$\left. \leq \frac{\lambda_n}{2\sqrt{n}} \left(\boldsymbol{W}_{0(2)} - \left| C_{n(21)} \left(C_{n(11)} \right)^{-1} \boldsymbol{W}_{0(1)} \circ sgn\left[\boldsymbol{\beta}_{0(1)} \right] \right| \right) \text{ element-wise} \right\}.$$

The $\text{rank}(X) = p_n \leq n$ assumption in Proposition 3.1 ensures that the random-weighting setup (1.3) has a unique solution (Osborne, Presnell and Turlach, 2000). For a random-design setting, the $\text{rank}(X) = p_n \leq n$ assumption can be replaced with the assumption that X is drawn from a joint continuous distribution (Tibshirani, 2013).

The strong irrepresentable condition (3.1) can be seen as a constraint on the relationship between active covariates and inactive covariates, that is, the total amount of an irrelevant covariate "represented" by a relevant covariate must be

strictly less than one. Similar to Zhao and Yu (2006)'s argument, A_n^w refers to recovery of the signs of coefficients for $\boldsymbol{\beta}_{0(1)}$, and B_n^w further implies obtaining $\widehat{\boldsymbol{\beta}}_{n(2)}^w = \mathbf{0}$ given A_n^w. The regularization parameter λ_n continues to play the role of trade-off between A_n^w and B_n^w: higher λ_n leads to larger B_n^w but smaller A_n^w, which forces the random-weighting method to drop more covariates, and vice versa. Meanwhile, larger $\boldsymbol{\eta}$ in (3.1), which could be interpreted as lower "correlation" between active covariates and inactive covariates, increases B_n^w but does not affect A_n^w, thus allowing the random-weighting method to better select the true model. Zhao and Yu (2006) also gave a few sufficient conditions that ensure the following designs of X satisfy condition (3.1):

- constant positive correlation,
- bounded correlation,
- power-decay correlation,
- orthogonal design, and
- block-wise design.

Again, we would like to highlight the fact that conditional on $\mathcal{F}_n$, the randomness of A_n^w and B_n^w derives from the random weights instead of $\boldsymbol{\epsilon}$. Besides that, notice how the presence of different penalty weights in weighting scheme (1.6) affects the strong irrepresentable condition (3.1) in B_n^w. We will see how these different weighting schemes affect the constraints on p_n and λ_n in order to achieve conditional model selection consistency.

Theorem 3.1. ***(Conditional Model Selection Consistency)*** *Assume assumptions in Proposition 3.1.*

(a) Under weighting schemes (1.4) and (1.5), if there exists $\frac{1}{2} < c_1 < c_2 < 1.5 - c_1$ *and* $0 \leq c_3 < \min\{2(c_2 - c_1), 2c_1 - 1\}$ *for which* $\lambda_n = \mathcal{O}(n^{c_2})$ *and* $p_n = \mathcal{O}(n^{c_3})$*, then as* $n \to \infty$*,*

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D.$$

(b) Under weighting scheme (1.6), if $(W_i, W_{0,j}) \overset{iid}{\sim} \mathrm{Exp}(\theta_{\mathrm{w}})$ *for some* $\theta_w > 0$*, and if* $\boldsymbol{\eta} = \mathbf{1}_{p_n - q}$*, and if there exists* $\frac{1}{2} < c_1 < c_2 < 1.5 - c_1$ *and* $0 \leq c_3 < \min\{\frac{2}{3}(c_2 - c_1), 2c_1 - 1\}$ *for which* $\lambda_n = \mathcal{O}(n^{c_2})$ *and* $p_n = \mathcal{O}(n^{c_3})$*, then as* $n \to \infty$*,*

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \overset{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D.$$

Theorem 3.1 could be interpreted as the "concentration" of the conditional distribution of signs of $\widehat{\boldsymbol{\beta}}_n^w$ around the neighborhood of the true signs of $\boldsymbol{\beta}$ as $n \to \infty$. Comparing the three weighting schemes, we can see that assigning random weights on the penalty term further impedes how fast p_n could increase with n while achieving conditional model selection consistency, especially when the penalty terms do not share a common random weight in weighting scheme (1.6). This adversely affects/violates the strong irrepresentable assumption (3.1), unless under a stringent condition where $\boldsymbol{\eta} = \mathbf{1}$. One sufficient condition for $\boldsymbol{\eta} = \mathbf{1}$

would be zero correlation between any relevant predictor and any irrelevant predictor, i.e. $C_{n(21)} = \mathbf{0}$ for all n.

We also point out that the conditional model selection consistency property under a fixed dimensional ($p_n = p$) setting could be easily obtained by taking $c_3 = 0$ in Theorem 3.1.

The next two results concern with the properties of conditional consistency and conditional asymptotic normality of the random-weighting samples under a fixed-dimension ($p_n = p$) setting.

Theorem 3.2. *Suppose $p_n = p$ is fixed. Assume (2.1), (2.2) and (2.4).*

(a) ***(Conditional Consistency)*** *If $\dfrac{\lambda_n}{n} \to 0$, then for all three weighting schemes (1.4), (1.5) and (1.6),*

$$\widehat{\boldsymbol{\beta}}_n^w \xrightarrow{c.p.} \boldsymbol{\beta}_0 \quad a.s.\ P_D.$$

(b) If $\dfrac{\lambda_n}{n} \to \lambda_0 \in (0, \infty)$, then

$$\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right) \xrightarrow{c.d.} \arg\min_{\boldsymbol{u}} g(\boldsymbol{u}) \quad a.s.\ P_D,$$

where

$$g(\boldsymbol{u}) = \mu_W \boldsymbol{u}' C \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} W_j |\beta_{0,j} + u_j|$$

and

(i) $W_j = 1$ for all j under weighting scheme (1.4),

(ii) $W_j = W_0$ for all j and $W_0 \sim F_W$ under weighting scheme (1.5),

(iii) $W_j \overset{iid}{\sim} F_W$ under weighting scheme (1.6).

In other words, the conditional distribution of $\widehat{\boldsymbol{\beta}}_n^w$ concentrates in the neighborhood of $\arg\min_{\boldsymbol{u}} g(\boldsymbol{u})$ as the sample size increases. In fact, for part (b)(i) of Theorem 3.2, conditional convergence in probability takes place since $g(\boldsymbol{u})$ is not a random function (i.e., does not involve any non-degenerate random variables).

Theorem 3.3. ***(Asymptotic Conditional Distribution)*** *Suppose $p_n = p$ is fixed. Assume (2.1), (2.2) and (2.4). Let $\widehat{\boldsymbol{\beta}}_n^{SC}$ be a strongly consistent estimator of $\boldsymbol{\beta}$ in the linear model (1.1) such that for $\boldsymbol{e}_n = \boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}_n^{SC}$,*

$$\frac{1}{\sqrt{n}} X' \boldsymbol{e}_n \to \mathbf{0} \quad a.s.\ P_D. \tag{3.2}$$

If $q = p$ and $\dfrac{\lambda_n}{\sqrt{n}} \to \lambda_0 \in [0, \infty)$, then

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{SC}\right) \xrightarrow{c.d.} \arg\min_{\boldsymbol{u}} V(\boldsymbol{u}) \quad a.s.\ P_D,$$

where

$$V(\boldsymbol{u}) = -2\boldsymbol{u}'\Psi + \mu_W \boldsymbol{u}' C \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} W_j \left[u_j \, sgn(\beta_{0,j})\right],$$

for $\Psi \sim N\left(\mathbf{0}, \sigma_W^2 \sigma_\epsilon^2 C\right)$, *and*

(i) $W_j = 1$ *for all* j *under weighting scheme (1.4),*
(ii) $W_j = W_0$ *for all* j, $W_0 \sim F_W$ *and* $W_0 \perp \Psi$ *under weighting scheme (1.5),*
(iii) $W_j \overset{iid}{\sim} F_W$ *and* $W_j \perp \Psi$ *for all* j *under weighting scheme (1.6).*

In particular, if $\lambda_0 = 0$, *then for all three weighting schemes (1.4), (1.5) and (1.6),*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{SC}\right) \xrightarrow{c.d.} N\left(\mathbf{0}\,,\ \frac{\sigma_W^2 \sigma_\epsilon^2}{\mu_W^2} C^{-1}\right) \quad a.s.\ P_D.$$

The OLS estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}$ and the standard LASSO estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{LAS}}(\lambda_n^*)$ with $\lambda_n^* = o(\sqrt{n})$ are two qualified candidates for $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}$ to satisfy the conditions in Theorem 3.3. (Note that λ_n^* does not necessarily have to be the same as the λ_n that we use for the random-weighting approach.) Firstly, due to Assumption (2.4), $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}$ is strongly consistent (Lai, Robbins and Wei, 1978), and

$$X'\boldsymbol{e}_n^{\mathrm{OLS}} = \left(X'Y - X'X(X'X)^{-1}X'Y\right) = \mathbf{0}.$$

Meanwhile, since $\mathbb{E}(|\epsilon_i|) < \infty$ for all i and $\lambda_n^* = o(\sqrt{n})$, $\widehat{\boldsymbol{\beta}}_n^{\mathrm{LAS}}(\lambda_n^*)$ is strongly consistent (Chatterjee and Lahiri, 2011a), and the KKT conditions ensure that

$$\frac{1}{\sqrt{n}} \left\| X'\boldsymbol{e}_n^{\mathrm{LAS}} \right\|_2 = \frac{1}{\sqrt{n}} \left\| X'\left(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_n^{\mathrm{LAS}}\right) \right\|_2 \le \frac{\lambda_n^* \sqrt{p}}{\sqrt{n}} \to 0 \quad a.s.\ P_D.$$

We also point out that centering on the true regression parameter

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right).$$

results in additional terms that depend on the sample path of realized data $\{y_1, y_2, \cdots\}$. Consequently, convergence in conditional distribution almost surely under P_D (just like the result in Theorem 3.3) could not be achieved. We refer readers to Remark A.1 in the Appendix for more details.

On the other hand, a more sophisticated argument is needed to establish the asymptotic conditional distribution for the case of $0 < q < p$. First, note that for $j \in \{j : \beta_{0,j} = 0\}$, $\sqrt{n}\widehat{\beta}_{n,j}^{SC}$ has an asymptotic normal distribution (denoted Z_j) under P_D. By the Skorokhod representation theorem, there exists random variables $U_{n,j}$ and U_j such that $U_{n,j} \overset{d}{=} \sqrt{n}\widehat{\beta}_{n,j}^{SC}$, $U_j \overset{d}{=} Z_j$, and $U_{n,j} \to U_j$ $a.s.\ P_D$. Then, for $(\lambda_n/\sqrt{n}) \to \lambda_0 \in [0, \infty)$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}\right) \xrightarrow{\mathrm{c.d.}} \arg\min_{\boldsymbol{u}} V^*(\boldsymbol{u}) \quad a.s.\ P_D, \tag{3.3}$$

where

$$V^*(\boldsymbol{u}) = -2\boldsymbol{u}'\Psi + \mu_W \boldsymbol{u}' C \boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} W_j \left[u_j \operatorname{sgn}(\beta_{0,j}) \mathbb{1}_{\{\beta_{0,j}\neq 0\}} + (|U_j + u_j| - |U_j|)\, \mathbb{1}_{\{\beta_{0,j}=0\}} \right],$$

for Ψ and $\{W_j\}_{1\leq j\leq p}$ defined in Theorem 3.3.

The results presented above fulfill our first objective to study and extend the asymptotic properties of the "one-step" random-weighting procedure that was considered by Newton, Polson and Xu (2021). However, we also recognize that the current "one-step" random-weighting setup (1.3) in Algorithm 1 does not produce random-weighting samples that have conditional sparse normality property. From Theorems 3.1 and 3.3, it is evident that even under a fixed dimensional ($p_n = p$) setting, the random weighting samples achieve conditional model selection consistency when $\lambda_n = \mathcal{O}(n^c)$ for some $\frac{1}{2} < c < 1$, whereas conditional asymptotic normality happens when $\lambda_n = o(\sqrt{n})$.

Unsurprisingly, this finding about (lack of) conditional sparse normality approximation coincides with many existing Bayesian and frequentist results. For instance, in the Bayesian framework, Theorem 7 of Castillo, Schmidt-Hieber and van der Vaart (2015) proved that the Bayesian LASSO approach (Park and Casella, 2008) could not achieve asymptotic sparse normality for any one given λ_n due to the conflicting demands of sparsity-inducement and normality approximation on the regularization parameter λ_n. In the frequentist setting, Liu and Yu (2013) pointed out that there does not exist one λ_n that allows a standard LASSO estimator (1.2) to simultaneously achieve model selection and asymptotic normality. Consequently, many variations of "two-step" LASSO estimators (e.g., Zou (2006)'s ALasso), and their corresponding bootstrap procedures (e.g., Das, Gregory and Lahiri (2019)'s perturbation bootstrap of ALasso) were introduced to overcome this shortcoming.

3.2. *Two-step procedure*

To overcome the regularization problem, we propose an extension to random weighting in LASSO regression. We retain the random-weighting framework of repeatedly assigning random-weights and optimizing the objective function (1.3), except we propose optimization in two-steps: In step one, we optimize

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} W_{0,j} |\beta_j| \right\} \tag{3.4}$$

to select variables. Let $\widehat{S}_n^w \subseteq \{1, \cdots, p_n\}$ be the set of variables being selected in (3.4), and let $(\widehat{S}_n^w)^c$ be the set of discarded variables. In addition, denote $X_{\widehat{S}_n^w}$ as the $n \times |\widehat{S}_n^w|$ submatrix of X whose columns correspond to the selected

variables in (3.4). Then, in step two, we obtain our random-weighting samples by solving

$$\widehat{\boldsymbol{\beta}}_n^w := \begin{bmatrix} \widehat{\boldsymbol{\beta}}^w_{n,\widehat{S}_n^w} \\ \\ \widehat{\boldsymbol{\beta}}^w_{n,(\widehat{S}_n^w)^c} \end{bmatrix} := \begin{bmatrix} \left(X'_{\widehat{S}_n^w} D_n X_{\widehat{S}_n^w}\right)^{-1} X'_{\widehat{S}_n^w} D_n Y \\ \\ \mathbf{0} \end{bmatrix}, \tag{3.5}$$

where the partition of $\widehat{\boldsymbol{\beta}}_n^w$ corresponds to $\widehat{S}_n^w$ and $\left(\widehat{S}_n^w\right)^c$.

Algorithm 2: Random-Weighting in LASSO+LS regression

Input :

- data: $D = (\boldsymbol{y}, X)$
- regularization parameter: λ_n
- number of draws: B
- choice of random weight distribution: F_W
- choice of weighting schemes: (1.4), (1.5) or (1.6)

Output :

- B sets of selected variables $\{\widehat{S}_n^{w,b}\}_{b=1}^B$
- B parameter samples $\{\widehat{\boldsymbol{\beta}}_n^{w,b}\}_{b=1}^B$

for $b = 1$ *to* B **do**

Draw i.i.d. random weights from F_W and substitute them into (1.3) ;
Optimize (3.4) to obtain $\widehat{S}_n^{w,b}$;
Based on the selected set of variables $\widehat{S}_n^{w,b}$, obtain $\widehat{\boldsymbol{\beta}}_n^{w,b}$ by solving (3.5) ;

end

For convenience, we shall refer to this proposed extension as a "two-step procedure", which is laid out in detail in Algorithm 2. This extension can be seen as the random-weighting version of Liu and Yu (2013)'s LASSO+LS procedure, i.e., a LASSO step (1.2) for variable selection followed by a least-square estimation for the selected variables. (Belloni and Chernozhukov (2013) had also studied the finite-sample and asymptotic properties of the post-LASSO OLS estimator.) We shall denote this unweighted two-step LASSO+LS estimator as $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$, and let $\widehat{S}_n$ be the set of variables selected (in the first step) by this estimator. Notice that $\widehat{S}_n$ and $\widehat{S}_n^w$ may be different due to the presence of random-weights in the selection step of (3.4). The superscript w of $\widehat{S}_n^w$ helps to remind readers that the set of selected variables in (3.4) could change with different sets of assigned random weights.

In this subsection, we adopt the same assumptions as we did in Theorem 3.1, including the fact that $p_n \leq n$ and X is full rank for all n. Thus $X_{\widehat{S}_n^w}$ is full rank and consequently,

$$X'_{\widehat{S}_n^w} D_n X_{\widehat{S}_n^w}$$

is also full rank and is invertible for all n.

For ease of presentation, we introduce a bit of additional notation. Let S_0 be the true set of relevant variables. To be consistent with our previous notation, we remind readers that $S_0 = \{1, \cdots, q\}$ without loss of generality, and $X_{S_0} = X_{(1)}$. We also partition $\widehat{\boldsymbol{\beta}}_n^w$ and $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$ into

$$\widehat{\boldsymbol{\beta}}_n^w = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{n(1)}^w \\ \\ \widehat{\boldsymbol{\beta}}_{n(2)}^w \end{bmatrix} \qquad \text{and} \qquad \widehat{\boldsymbol{\beta}}_n^{LAS+LS} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{n(1)}^{LAS+LS} \\ \\ \widehat{\boldsymbol{\beta}}_{n(2)}^{LAS+LS} \end{bmatrix}$$

respectively, which correspond to the partition of $\boldsymbol{\beta}_0 = \begin{bmatrix} \boldsymbol{\beta}_{0(1)} & \boldsymbol{\beta}_{0(2)} \end{bmatrix}'$. We observe that if $\widehat{S}_n^w = S_0$, then

$$\widehat{\boldsymbol{\beta}}_{n,\widehat{S}_n^w}^w = \widehat{\boldsymbol{\beta}}_{n(1)}^w \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_{n,(\widehat{S}_n^w)^c}^w = \widehat{\boldsymbol{\beta}}_{n(2)}^w = \boldsymbol{\beta}_{0(2)} = \mathbf{0}.$$

Similarly, if $\widehat{S}_n = S_0$, then

$$\widehat{\boldsymbol{\beta}}_{n,\widehat{S}_n}^{LAS+LS} = \widehat{\boldsymbol{\beta}}_{n(1)}^{LAS+LS} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_{n,(\widehat{S}_n)^c}^{LAS+LS} = \widehat{\boldsymbol{\beta}}_{n(2)}^{LAS+LS} = \boldsymbol{\beta}_{0(2)} = \mathbf{0}.$$

We are now ready to establish the conditional sparse normality property of the two-step random-weighting samples (3.5) under growing p_n setting with appropriate regularity conditions.

Theorem 3.4. ***(Conditional Sparse Normality)*** *Adopt all regularity assumptions as stated in Theorem 3.1 (including assumptions about the different rates of λ_n and p_n for weighting schemes (1.4), (1.5) and (1.6)). Furthermore, assume $\mu_W = 1$ and $C_{n(11)} \to C_{11}$ for some nonsingular matrix C_{11}. Let $\widehat{\boldsymbol{\beta}}_n^w$ be the two-step random-weighting samples defined in (3.5), and let $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$ be the unweighted two-step LASSO+LS estimator (i.e. a LASSO variable selection step (1.2) followed by least-squares estimation for the selected variables). Then,*

$$P\left(\widehat{S}_n^w = S_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D,$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{n(1)}^w - \widehat{\boldsymbol{\beta}}_{n(1)}^{LAS+LS}\right) \xrightarrow{c.d.} N_q\left(\mathbf{0}\ ,\ \sigma_W^2 \sigma_\epsilon^2 C_{11}^{-1}\right) \quad a.s.\ P_D.$$

Theorem 3.4 highlights the improvement brought about by the extended random-weighting framework as compared to the original "one-step" procedure considered by Newton, Polson and Xu (2021). With a common regularization parameter λ_n (and all regularity conditions that apply), the two-step random-weighting samples attain conditional model selection consistency and achieve conditional asymptotic normality (by centering at the unweighted two-step LASSO+LS estimator) on the true support S_0 under growing p_n setting. We note that the rate of convergence for selection above is relatively slow (see Lemma A.11 in the Appendix). In any case, the theorem assures that appropriate confidence intervals constructed from the two-step random-weighting samples have coverage that converges to the nominal value.

We conclude this section by establishing that the random-weighting samples from the two-step procedure also achieve the conditional consistency property under growing p_n setting. This could be viewed as an improvement to the result that we have in Theorem 3.2(a) which applies to fixed dimensional setting only.

Theorem 3.5. ***(Conditional Consistency)*** *Adopt all regularity assumptions as stated in Theorem 3.1 (including assumptions about the different rates of λ_n and p_n for weighting schemes (1.4), (1.5) and (1.6)). Let $\widehat{\boldsymbol{\beta}}_n^w$ be the two-step random-weighting samples defined in (3.5). Then*

$$\left\|\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right\|_2 \xrightarrow{c.p.} 0 \quad a.s.\ P_D.$$

Theorem 3.5 indicates a concentration of the conditional distribution of $\widehat{\boldsymbol{\beta}}_n^w$ near $\boldsymbol{\beta}_0$ with increasing sample size given almost any data set.

3.3. Remarks

The two-step random-weighting procedure is a valid bootstrap procedure for Liu and Yu (2013)'s LASSO+LS estimator $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$ under growing p_n setting. Using very similar regularity assumptions, Liu and Yu (2013) showed that their LASSO+LS method gives consistent model selection under P_D, and

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{n(1)}^{LAS+LS} - \boldsymbol{\beta}_{0(1)}\right)$$

converges to $N\left(\mathbf{0}\,,\,\sigma_\epsilon^2 C_{11}^{-1}\right)$ under P_D. Hence, based on Theorem 3.4, by fulfilling the appropriate regularity assumptions and drawing random weights from F_W with unitary mean and variance ($\mu_W = \sigma_W^2 = 1$), the conditional distribution of the two-step random-weighting samples $\widehat{\boldsymbol{\beta}}_n^w$ converges to the same distributional limit of the LASSO+LS estimator under P_D. This enables the two-step random-weighting procedure to produce bootstrap samples that provide valid distributional approximation to the LASSO+LS estimator. It also assures that the coverage of confidence intervals constructed from the random weighting samples (e.g. by percentile method) will converge to the nominal coverage.

We point out that by capitalizing on the sub-Gaussian nature of $\boldsymbol{\epsilon}$, Liu and Yu (2013)'s proposed residual bootstrap procedure for their LASSO+LS estimator works under high-dimensional setting where p_n grows nearly exponential with sample size n. On the other hand, in this paper, we only require finite fourth moment assumptions for both error term $\boldsymbol{\epsilon}$ and random weights $\boldsymbol{W}$, and our random-weighting procedure only allows p_n to grow at a polynomial rate of $o(\sqrt{n})$.

Similarly, under fixed dimensional ($p_n = p$) setting where $\boldsymbol{\beta}_0$ is not sparse (i.e. $q = p$), our one-step random-weighting approach in Algorithm 1 could also be a valid bootstrap procedure for the standard LASSO estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{LAS}}(\lambda_n)$. Specifically, Knight and Fu (2000) proved that for $(\lambda_n/\sqrt{n}) \to \lambda_0 \in [0, \infty)$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{LAS}}(\lambda_n) - \boldsymbol{\beta}_0\right)$$

converges to the same distributional limit stated in Theorem 3.3 under P_D. However, for the case where $q < p$, the one-step random-weighting procedure no longer provides valid distributional approximation to $\widehat{\boldsymbol{\beta}}_n^{\mathrm{LAS}}(\lambda_n)$, as evident from the Skorokhod argument. This mimics the asymptotic conditional distribution of the LASSO parametric residual bootstrap (Knight and Fu, 2000).

4. Numerical experiments

We perform simulation studies and data analysis using R (R Core Team, 2019); all source code is available at the Github public repository: https://github.com/wiscstatman/optimizetointegrate/tree/master/Tun.

4.1. Simulation

A simulation study of one-step random-weighting procedures (Algorithm 1) was previously reported (Newton, Polson and Xu, 2021), and so here we study performance of the two-step random-weighting procedure (Algorithm 2) for all three weighting schemes (1.4), (1.5) and (1.6) – denoted RW1, RW2 and RW3 respectively – in several experimental settings, and compare it with:

- Bayesian LASSO (Park and Casella, 2008), which can be easily implemented with R package `monomvn` (Gramacy, Moler and Turlach, 2019)
- parametric residual bootstrap (Knight and Fu, 2000), which is a very common and easily implementable bootstrap procedure in LASSO regression. We denote this method as RB thereafter.

We drew inspiration from Das and Lahiri (2019), Liu and Yu (2013) and Newton, Polson and Xu (2021) in setting up our simulation schemes. Specifically, we consider 8 simulation settings as tabulated in Table 1. In all settings, the generative state $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})'$ is defined as $\beta_{0,j} = (3/4) + (1/4)j$ for $j = 1, \cdots, q$ and $\beta_{0,j} = 0$ for $j = q+1, \cdots, p$. The predictors $\boldsymbol{x}_i$ are drawn from p-variate normal distribution with different covariance structures. $\Sigma^{(1)}$ has the following structure

$$\Sigma^{(1)}_{i,j} = \mathbb{1}_{\{i=j\}} + \mathbb{1}_{\{i\neq j\}} \times \left(0.3^{|i-j|}\mathbb{1}_{\{i\leq q\}}\mathbb{1}_{\{j\leq q\}}\right) \quad \text{for} \quad 1 \leq i, j \leq 10. \tag{4.1}$$

$\Sigma^{(3)}$ also has the same structure as (4.1), except that it has larger dimension $p = 50$. Meanwhile, $\Sigma^{(2)}$ has the following structure: for $1 \leq i, j \leq 10$,

$$\Sigma^{(2)}_{i,j} = \mathbb{1}_{\{i=j\}} + \mathbb{1}_{\{i\neq j\}} \times \left[0.4\mathbb{1}_{\{i\leq q\}}\mathbb{1}_{\{j\leq q\}} + 0.5\left(1 - \mathbb{1}_{\{i\leq q\}}\mathbb{1}_{\{j\leq q\}}\right)\right].$$

We verify that only simulation settings 5 and 6 violate the strong irrepresentable condition (3.1), whereas the other six simulation settings satisfy assumption (3.1). By simulating i.i.d. ϵ_i and $\boldsymbol{x}_i$, we generate $y_i = \boldsymbol{x}_i\boldsymbol{\beta}_0 + \epsilon_i$ for $i = 1, \cdots, n$.

Purpose of simulation setup: The even-numbered simulation settings share the same specifications as their odd-numbered counterparts except with

TABLE 1
Simulation Settings

Setting	n	p	q	ϵ_i	$\boldsymbol{x}_i \sim N_p(\mathbf{0}, \Sigma)$
1	100	10	6	$N(0,1)$	$\Sigma = \Sigma^{(1)}$
2	500	10	6	$N(0,1)$	$\Sigma = \Sigma^{(1)}$
3	100	10	6	$\chi^2_2 - 2$	$\Sigma = \Sigma^{(1)}$
4	500	10	6	$\chi^2_2 - 2$	$\Sigma = \Sigma^{(1)}$
5	100	10	6	$N(0,1)$	$\Sigma = \Sigma^{(2)}$
6	500	10	6	$N(0,1)$	$\Sigma = \Sigma^{(2)}$
7	100	50	6	$N(0,1)$	$\Sigma = \Sigma^{(3)}$
8	500	50	6	$N(0,1)$	$\Sigma = \Sigma^{(3)}$

larger sample size n (e.g. Setting 2 versus Setting 1, Setting 4 versus Setting 3, et cetera). Simulation Settings 3 and 4 are used as an example of cases where the error term $\boldsymbol{\epsilon}$ is no longer normally distributed, whereas Simulation Settings 5 and 6 are set up to illustrate the situations where the strong irrepresentable condition (3.1) is violated. Finally, we increase the dimension p of predictors by five-fold in Settings 7 and 8 to compare performances in higher-dimensional setting.

For each simulation setting, we generate $T = 500$ independent datasets. For each simulated data set, we draw $B = 1000$ posterior/bootstrap samples from the 5 aforementioned methods: Bayesian LASSO (BLASSO), two-step random-weighting with schemes (1.4), (1.5) and (1.6), and residual bootstrap. For the Bayesian LASSO procedure, we specify a 2000 burn-in period. In addition, Bayesian LASSO imposes a noninformative marginal prior on σ^2_ϵ, $\pi(\sigma^2_\epsilon) \sim 1/\sigma^2_\epsilon$, and a Jeffrey's prior on λ_n. To induce sparsity in the MCMC samples of $\boldsymbol{\beta}$, the posterior distribution is sampled by a Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm (Green, 1995), with a uniform prior specified on the number of non-zero coefficients to be included in the model. For the three random-weighting schemes, all i.i.d. random weights are drawn from a standard exponential distribution. The regularization parameter λ_n is chosen via cross-validation using Liu and Yu (2013)'s (unweighted) LASSO+LS procedure, and then the same λ_n is used to draw the 1000 random-weighting samples according to Algorithm 2. We note that the optimization step (3.4) can be easily computed using R package `glmnet` (Friedman, Hastie and Tibshirani, 2010). Meanwhile for residual bootstrap, its regularization parameter λ_n^{RB} is chosen via cross-validation using standard LASSO, and values of λ_n^{RB} are thereafter fixed for all bootstrap computations on the same dataset.

For each of the five aforementioned methods, we obtain $\{\widehat{\beta}_j^{(b,t)}\}$ that represents the j^{th} component of sampled/bootstrapped $\boldsymbol{\beta}$ in the b^{th} iteration for the

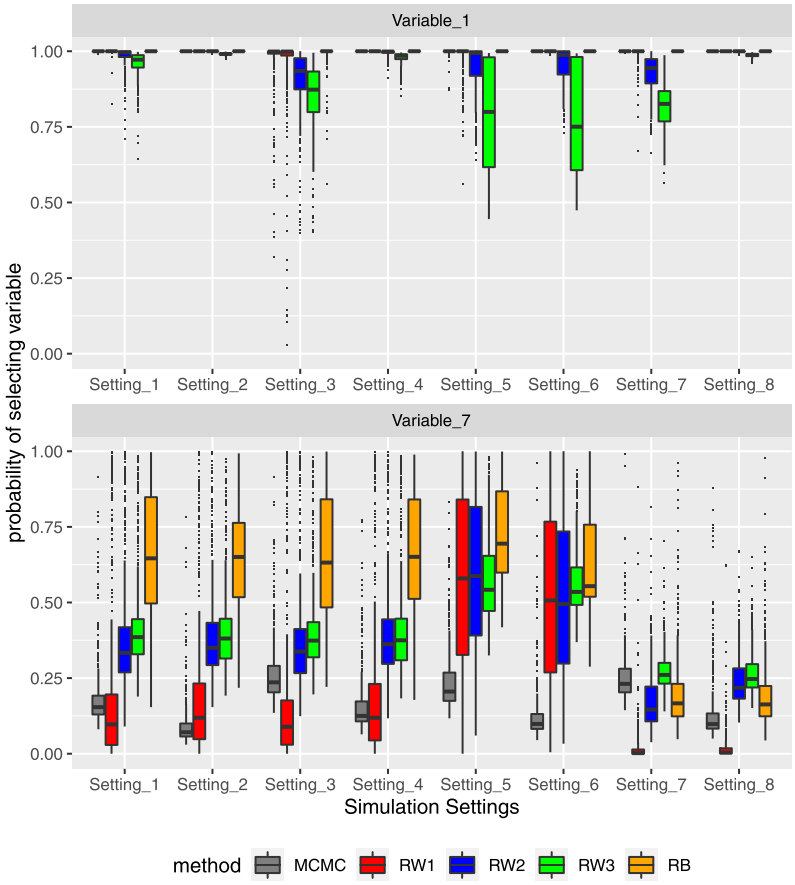

FIG 1. *Simulation: Sampling distribution of conditional (on data) probabilities of selecting β_1 and β_7 among $T = 500$ simulated data sets in 8 simulation settings by the 5 methods: MCMC via Bayesian LASSO, two-step random-weighting approach using weighting schemes (1.4) (denoted RW1), (1.5) (denoted RW2) and (1.6) (denoted RW3), and LASSO residual bootstrap (denoted RB).*

t^{th} simulated data set, where $j = 1, \cdots, p$, and $b = 1, \cdots, B$, and $t = 1, \cdots, T$. To be precise, we have

$$\left\{ \widehat{\beta}^{(b,t)}_{j(\text{MCMC})}, \widehat{\beta}^{(b,t)}_{j(\text{RW1})}, \widehat{\beta}^{(b,t)}_{j(\text{RW2})}, \widehat{\beta}^{(b,t)}_{j(\text{RW3})}, \widehat{\beta}^{(b,t)}_{j(\text{RB})} \right\}$$

that correspond to the sampled/bootstrapped $\boldsymbol{\beta}$'s of the five aforementioned methods, but for brevity we drop the subscripts whenever it does not cause any confusion, since each method is subject to the same performance evaluation. We then assess the performances of each of these five methods – BLASSO, RW1, RW2, RW3 and RB – in each of the 8 simulation settings using the following comparison criteria:

TABLE 2

Empirical coverage $\hat{q}_j$ and average width $\hat{l}_j$ (in parentheses) of the two-sided 90% CI for the first 10 variables in Simulation Setting 8, using the five approaches: MCMC via BLASSO, two-step random-weighting approach using weighting schemes (1.4) (denoted RW1), (1.5) (denoted RW2) and (1.6) (denoted RW3), and LASSO residual bootstrap (denoted RB).

$\beta_{0,j}$	MCMC	RW1	RW2	RW3	RB
1.00	0.918 (0.161)	0.878 (0.152)	0.882 (0.152)	0.906 (0.16)	0.344 (0.153)
1.25	0.908 (0.169)	0.88 (0.158)	0.876 (0.159)	0.904 (0.168)	0.588 (0.16)
1.50	0.894 (0.168)	0.864 (0.158)	0.868 (0.158)	0.886 (0.165)	0.578 (0.16)
1.75	0.918 (0.168)	0.886 (0.159)	0.892 (0.159)	0.9 (0.165)	0.596 (0.16)
2.00	0.922 (0.168)	0.894 (0.159)	0.882 (0.159)	0.898 (0.164)	0.556 (0.16)
2.25	0.886 (0.161)	0.866 (0.151)	0.872 (0.152)	0.874 (0.157)	0.35 (0.153)
0.00	1 (0.04)	1 (0.016)	1 (0.096)	1 (0.099)	0.998 (0.023)
0.00	1 (0.041)	0.998 (0.018)	1 (0.097)	1 (0.1)	1 (0.024)
0.00	1 (0.04)	1 (0.015)	1 (0.097)	1 (0.099)	1 (0.023)
0.00	0.998 (0.04)	1 (0.015)	1 (0.097)	1 (0.1)	1 (0.023)

- Conditional (on data) probability of selecting the j^{th} variable where $j = 1, \cdots, p$. Specifically, for each simulated data set $t = 1, \cdots, T$, we keep track of

$$\hat{p}_j^{(t)} := \frac{1}{B} \left| \left\{ b : \widehat{\beta}_j^{(b,t)} \neq 0 \right\} \right|.$$

  We note that the computation of $\hat{p}_j^{(t)}$ is sensible because all the five methods (including BLASSO with RJMCMC implementation) induce sparsity in the sampled/bootstrapped $\boldsymbol{\beta}$'s.
- Coverage and average width of the two-sided 90% credible/confidence interval (CI) for the j^{th} variable where $j = 1, \cdots, p$. Specifically, denote $\hat{r}_{0.05,j}^{(t)}$ and $\hat{r}_{0.95,j}^{(t)}$ as the 5^{th} percentile and 95^{th} percentile of the empirical distribution of $\{\widehat{\beta}_j^{(b,t)}\}_{1 \leq b \leq B}$. Then, the average width (across $T = 500$ simulated data sets) of the two-sided 90% CI for the j^{th} variable is computed as

$$\hat{l}_j := \frac{1}{T} \sum_{t=1}^{T} \left(\hat{r}_{0.95,j}^{(t)} - \hat{r}_{0.05,j}^{(t)} \right),$$

and its corresponding empirical coverage is calculated as

$$\hat{q}_j := \frac{1}{T}\left|\left\{t : \hat{r}^{(t)}_{0.05,j} \le \beta_{0,j} \le \hat{r}^{(t)}_{0.95,j}\right\}\right|.$$

Firstly, as expected, performance improves with larger sample size n, such as higher coverage probabilities and narrower CI's. Secondly, we note that the two-step random-weighting approach, especially weighting schemes (1.4) and (1.5) – denoted RW1 and RW2, outperforms the LASSO residual bootstrap (denoted RB) in all performance measures.

In Figure 1, we show the sampling distributions of $\left\{\hat{p}^{(t)}_1\right\}_{1\le t\le T}$ and $\left\{\hat{p}^{(t)}_7\right\}_{1\le t\le T}$ among the $T = 500$ simulated data sets in the 8 simulation settings for all the five methods. Recall that the first variable corresponds to $\beta_{0,1} = 1$ and the seventh variable corresponds to $\beta_{0,7} = 0$. Sampling distribution of conditional (on data) probabilities of selecting other relevant predictors is similar to that of the first variable, and sampling distribution of conditional probabilities of selecting other irrelevant predictors is similar to that of the seventh variable. In all 8 simulation settings, all methods almost always select the first variable, except for RW3 in Simulation Settings 5 and 6, due to the violation of condition (3.1). However, similar to MCMC, the two-step random-weighting schemes (especially RW1) have lower conditional probabilities of selecting the seventh variable (which is an irrelevant predictor) than the LASSO RB. This illustrates that the two-step random-weighting approach is more capable of discarding irrelevant variables as compared to LASSO residual bootstrap. Only in Simulation Settings 5 and 6 do we see similarly high conditional probabilities of selecting the seventh variable among RW1, RW2, RW3 and RB, due to violation of condition (3.1).

Empirical coverage and average width of the two-sided 90% CI's for relevant predictors (i.e. $\beta_{0,j} \neq 0$) paint a similar story. For illustration, the empirical coverage $\hat{q}_j$ and average width $\hat{l}_j$ (in parentheses) of the two-sided 90% CI for the first 10 variables, i.e. for $j = 1, \cdots, 10$, in Simulation Setting 8, are tabulated in Table 2. Generally, average widths of CI's are similar among all five methods in all but two simulation settings, where RW3 has much wider 90% CI's in Simulation Settings 5 and 6. Interestingly, empirical coverage for MCMC and random-weighting samples is similar and close to 90%, but the LASSO residual bootstrap samples always have the lowest empirical coverage, especially in Simulation Settings 7 and 8, where their empirical coverage is only around 30%–40%.

In addition, we obtain the total variation distance between empirical cumulative distribution function (ecdf) of MCMC samples and ecdf of samples produced by one of the other four methods – the two-step random-weighting (RW1, RW2 and RW3) and residual bootstrap (RB). The intent is to assess how well the random-weighting methods approximate the MCMC-approximated posterior. Specifically, for the j^{th} variable in the t^{th} simulated data set, let

$$\hat{F}^{(t)}_{j(MCMC)} = \text{ ecdf of } \left\{\widehat{\beta}^{(b,t)}_{j(MCMC)}\right\}_{1\le b\le B},$$

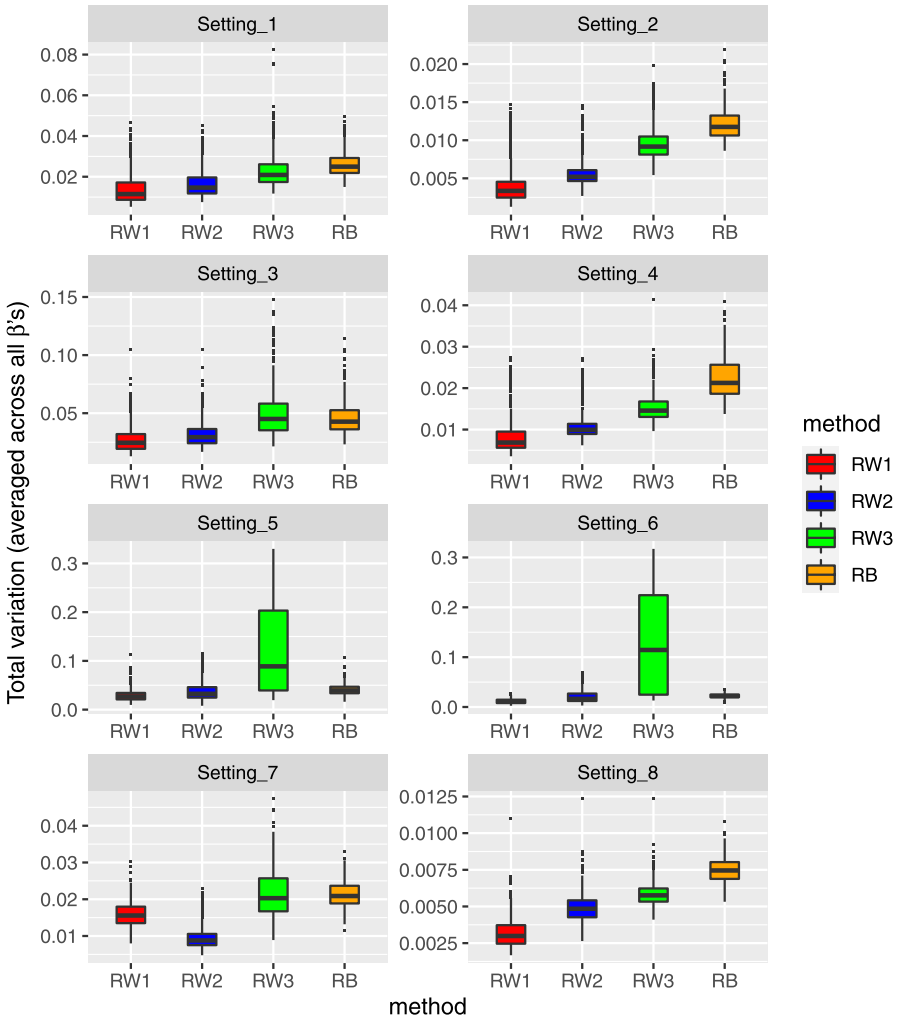

FIG 2. *Simulation: Sampling distribution of total variation distance between the random-weighting distribution and a Bayesian posterior (averaged across all β's) among $T = 500$ simulated data sets in 8 simulation settings between ecdf of MCMC samples and ecdf of samples from each of the 4 methods: two-step random-weighting approach using weighting schemes (1.4) (denoted RW1), (1.5) (denoted RW2) and (1.6) (denoted RW3), and LASSO residual bootstrap (denoted RB).*

and let $\hat{F}^{(t)}_{j(.)}$ be the ecdf of samples produced by one of the other 4 methods: RW1, RW2, RW3 or RB. Note that the ecdf's are easily obtained via the function `ecdf` in R `base` package (R Core Team, 2019). Then, for each of the 4 methods, we keep track of the total variation (averaged across all p variables) for each

TABLE 3
Variables in Boston Housing Data Set

Abbreviation	Variable
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
Black	proportion Black residents by town
lstat	lower status of the population (percent)

simulated data set $t = 1, \cdots, T$:

$$\mathrm{TV}^{(t)} = \frac{1}{p}\sum_{j=1}^{p}\frac{1}{2}\sum_{\omega\in\Omega}\left|\hat{F}^{(t)}_{j(MCMC)}(\omega) - \hat{F}^{(t)}_{j(.)}(\omega)\right|,$$

where the inner summation is approximated using a trapezoidal rule with an interval width of 0.001.

Figure 2 displays the sampling distribution of total variation distance between the random-weighting distribution and a Bayesian posterior (averaged across all β's), $\{TV^{(t)}\}_{1\le t\le T}$, among the $T = 500$ simulated data sets in the 8 simulation settings for the 4 methods: RW1, RW2, RW3 and RB. Generally, larger sample size n leads to smaller total variations. Moreover, in all simulation settings, RW1 and RW2 have smaller total variations than that of RB, which illustrates the viability of the two-step random-weighting samples to approximate posterior inference. RW3 has larger total variations especially in Settings 5 and 6, where the strong irrepresentable condition (3.1) is violated. This illustrates the need for restrictive regularity assumption for weighting scheme (1.6) that we highlighted in part (c) of Theorem 3.1.

4.2. Benchmark data example

To further illustrate the two-step random-weighting methodology, we apply it to the often-analyzed Boston Housing data set, which is available in the R package `MASS` (Venables and Ripley, 2002). Data from $n = 506$ housing prices in the suburbs of Boston are available, with response the median value of owner-occupied homes in $1000's, and with 13 variables ($p = 13$) listed in Table 3.

Again, we apply Bayesian LASSO, the random-weighting approach for all three weighting schemes (1.4), (1.5) and (1.6) according to Algorithm 2, as well as the parametric residual bootstrap method (Knight and Fu, 2000) with $B = 1000$. We use the same prior specifications as well as RJMCMC implementation for Bayesian LASSO as we did in our simulation studies. For the

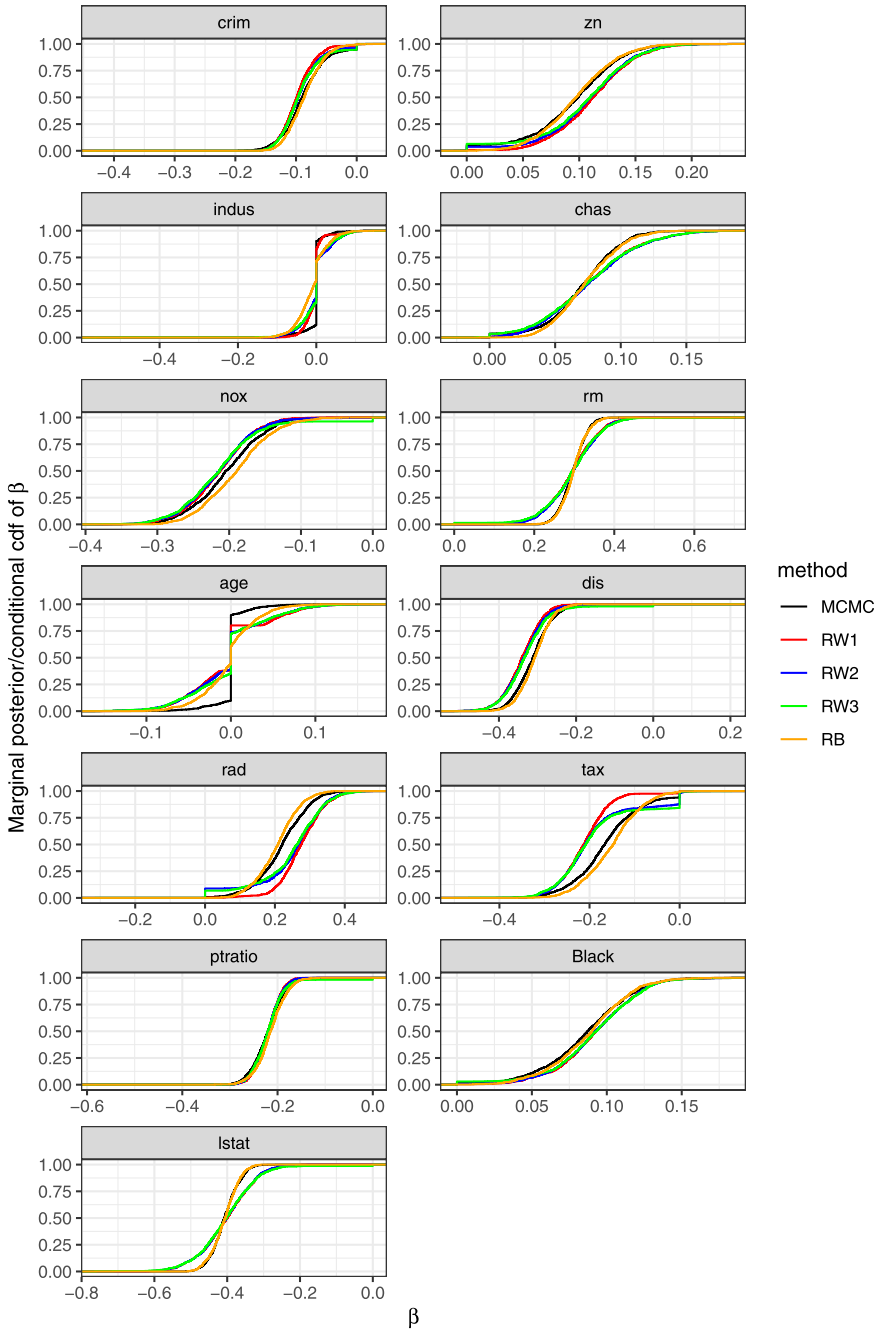

FIG 3. *Boston Housing data example: Marginal posterior/conditional distribution plots for* $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{13})'$ *sampled from the 5 methods – MCMC via Bayesian LASSO, the two-step random-weighting approach using weighting schemes (1.4) (1.5) and (1.6) (denoted RW1, RW2 and RW3 respectively), as well as the parametric residual bootstrap (denoted RB).*

random-weighting approach, random weights are drawn from a standard exponential distribution, and the regularization parameter is chosen with cross-validation using Liu and Yu (2013)'s unweighted LASSO+LS procedure (i.e. 2-step cross-validation). Meanwhile, for residual bootstrap, its regularization parameter is chosen via cross-validation using standard LASSO.

Figure 3 shows the marginal posterior distributions of $\boldsymbol{\beta}$'s sampled from MCMC as well as the marginal conditional (on data) distributions of $\boldsymbol{\beta}$'s obtained from the random-weighting methods and the parametric residual bootstrap. For most of the coefficients, there is very good agreement among the methods. One notable feature is that the parametric residual bootstrap approach induces the least sparsity among all five methods for variables `indus` and `age`. In addition, Bayesian LASSO appears to introduce slightly more sparsity than the random-weighting schemes for the variable `age`. Besides that, random-weighting with different penalty weights (1.6) appears to produce lower outliers for variables `crim`, `indus` and `ptratio`.

5. Discussion

The findings above extend what is known about asymptotic conditional sampling distribution of random-weighting solutions in LASSO regression, and thereby contribute to our understanding of uncertainty quantification in penalized estimation settings. Because random weighting is readily deployed in contemporary applications involving large-scale optimization, further work is warranted that sheds more light on the random-weighting approach and its links with bootstrap and Bayesian approaches.

Connection to bayes

Our foray into random-weighting asymptotics was motivated in part by renewed interest in the algorithm from the perspective of Bayesian nonparametric learning and generalized Bayesian analysis: Bissiri, Holmes and Walker (2016); Lyddon, Walker and Holmes (2018); Lyddon, Holmes and Walker (2019); Fong, Lyddon and Holmes (2019); Pompe (2021). We have not pursued those connections here, considering for example technical difficulties in working the appropriate prior distributions, but rather have focused on asymptotic conditional sampling theory.

Perturbation bootstrap (in general)

Whilst the random-weighting approach has a Bayesian justification, its resemblance to existing bootstrap algorithms, especially the perturbation bootstrap, warrants a comparison with non-Bayesian bootstrap literature. The (naive) perturbation bootstrap was introduced by Jin, Ying and Wei (2001) as a method to estimate sampling distributions of estimators related to U-process-structured

objective functions. Chatterjee and Bose (2005) established first-order distributional consistency of a generalized perturbation bootstrap technique in M-estimation where they allowed both $n \to \infty$ and $p_n \to \infty$. That paper also pointed out that for broader classes of models, the generalized bootstrap method is not second-order accurate without appropriate bias-correction and studentization. In particular, the work in (naive) perturbation bootstrap resembles the Bayesian NPL objective function (Fong, Lyddon and Holmes, 2019). Subsequently, Minnier, Tian and Cai (2011) proved the first-order distributional consistency of the perturbation bootstrap for Zou (2006)'s Adaptive LASSO (ALasso) and Fan and Li (2001)'s smoothly clipped absolute deviation (SCAD) under fixed-p setting in order to construct accurate confidence regions for ALasso and SCAD estimators. Again, their work has the flavor of Bayesian Loss-NPL (Fong, Lyddon and Holmes, 2019) where the loss function is either ALasso or SCAD. More recently, Das, Gregory and Lahiri (2019) extended the work of Minnier, Tian and Cai (2011) by introducing a suitably Studentized version of modified perturbation bootstrap ALasso estimator that achieves second-order correctness in distributional consistency even when $p_n \to \infty$.

Bootstrapping for LASSO

Various bootstrap techniques have been considered to construct confidence regions for standard LASSO estimators in (1.2) under different model settings, including fixed or random design, as well as homoscedastic or heteroscedastic errors $\boldsymbol{\epsilon}$. Knight and Fu (2000) first considered the residual bootstrap under fixed design and homoscedastic error. Chatterjee and Lahiri (2010) presented a rigorous proof for the heuristic discussion of Knight and Fu (2000)'s Section 4 to show that the LASSO residual bootstrap samples fail to be distributionally consistent unless $\boldsymbol{\beta}_0$ is not sparse, for which Knight and Fu (2000) invoked the Skorokhod's argument. Subsequently, Chatterjee and Lahiri (2011b) rectified the shortcoming by proposing a modified residual bootstrap method by thresholding the Lasso estimator. Meanwhile, Camponovo (2015) proposed a modified paired-bootstrap technique and established its distributional consistency to approximate the distribution of Lasso estimators in linear models with random design and heteroscedastic errors. Recently, Das and Lahiri (2019) considered the perturbation bootstrap method for Lasso estimators under both fixed and random designs with heteroscedastic errors. Since centering on the thresholded Lasso estimator (c.f. Chatterjee and Lahiri, 2011b) resulted in distributional inconsistency of the naive perturbation bootstrap, Das and Lahiri (2019) proceeded with a suitably Studentized version of modified perturbation bootstrap (c.f. Das, Gregory and Lahiri (2019)) to rectify the shortcoming.

Comparison and contribution of our paper

Interestingly, the setup of naive perturbation bootstrap in Das and Lahiri (2019) mimics the proposed random-weighting approach (1.3) in LASSO regression

with weighting scheme (1.4), but there remain some differences in our approach. Das and Lahiri (2019) also considered heteroscedastic error term $\boldsymbol{\epsilon}$, which we do not consider in this paper. Meanwhile, the weighting schemes considered in this paper are slightly more flexible, since we also consider the cases where independent random weights are also assigned on the LASSO penalty term in weighting schemes (1.5) and (1.6). The random weights in Das and Lahiri (2019)'s perturbation bootstrap are restricted to independent draws from distribution with $\sigma_W^2 = \mu_W^2$, whereas we consider any positive random weights with finite fourth moment. Furthermore, our extended random-weighting framework in Section 3.2 attains conditional sparse normality property under growing p_n setting, whereas Das and Lahiri (2019)'s (modified) perturbation bootstrap method achieves distributional consistency under fixed dimensional ($p_n = p$) setting.

Appendix A

We present the key steps of the proofs for all the theorems and proposition in this paper. More detailed derivations are furnished in Ng (2022). Many subsequent proofs rely on this following result.

Lemma A.1. *Let $U_1, U_2, \cdots$ be any i.i.d. random variables with $\mathbb{E}(U_i) = 0$ and $\mathbb{E}[(U_i)^2] = \sigma^2 < \infty$. Then for any bounded sequence of real numbers $\{k_i\}$ and for any $\frac{1}{2} < c < 1$,*

$$\frac{1}{n^c} \sum_{i=1}^{n} k_i U_i \xrightarrow{a.s.} 0.$$

Proof. This lemma is a slight generalization of Theorem 2.5.8 of Durrett (2010). Apply the same techniques in Durrett (2010)'s proof to obtain the result; otherwise, see Ng (2022) for more details. □

Lemma A.2. *Assume assumptions (2.2) and (2.3). Then,*

$$\left\| \left(C^w_{n(11)} \right)^{-1} \right\|_2 = O_p(1).$$

Proof. Due to assumptions (2.2) and (2.3) and that q is fixed, $C_{n(11)}$ is invertible for all n. It is also easy to verify that $C^w_{n(11)}$ is invertible for every n. Next,

$$C^w_{n(11)} = C_{n(11)} + \frac{1}{n} X'_{(1)} (D_n - \mu_W I_n) X_{(1)}$$

where the Strong Law of Large Numbers ensures that

$$\frac{1}{n} X'_{(1)} (D_n - \mu_W I_n) X_{(1)} \xrightarrow{\text{a.s.}} \mathbf{0}$$

due to assumption (2.2). Since $C_{n(11)}$ is invertible for all n, we have

$$\left\| \left(C^w_{n(11)} \right)^{-1} \right\|_2 = \left\| \left(C_{n(11)} + o(1) \right)^{-1} \right\|_2 = \mathcal{O}(1) \ a.s.$$

In fact, if we assume $C_{n(11)} \to C_{11}$ for some nonsingular matrix C_{11} in Lemma A.2, then by the Strong Law of Large Numbers and Continuous Mapping Theorem,

$$\left(C^w_{n(11)}\right)^{-1} \xrightarrow{\text{a.s.}} \frac{1}{\mu_W} C_{11}^{-1}. \quad \square$$

Lemma A.3. *Assume assumptions (2.2) and (2.3). For any* $\frac{1}{2} < c_1 < 1$, *if* $\exists$ $0 \le c_3 < 2c_1 - 1$ *for which* $p_n = \mathcal{O}(n^{c_3})$, *then*

$$\left\| n^{1-c_1} \widetilde{C}^w_n \right\|_2 = o_p(1).$$

Proof. Let

$$H = X_{(1)} C^{-1}_{n(11)} C_{n(12)} - X_{(2)}.$$

Then

$$n^{1-c_1} \widetilde{C}^w_n = \frac{1}{n^{c_1}} H'(\mu_W I_n - D_n) X_{(1)} \left(C^w_{n(11)}\right)^{-1}.$$

Due to assumptions (2.2) and (2.3) and that q is fixed, every element of the matrix H is bounded. Let h_{ij} and x_{ij} be the $(i,j)^{th}$ element of H and $X_{(1)}$ respectively. For $0 \le c_3 < 2c_1 - 1$, by Lemma A.1,

$$\frac{1}{n^{c_1 - \frac{c_3}{2}}} \sum_{i=1}^{n} h_{k,i} x_{i,l} (W_i - \mu_W) \xrightarrow{\text{a.s.}} 0$$

for every $k = 1, \cdots, p_n - q$ and $l = 1, \cdots, q$. Thus,

$$\begin{aligned}
&\left\| \frac{1}{n^{c_1}} H'(\mu_W I_n - D_n) X_{(1)} \right\|_2^2 \\
&\le \left\| \frac{1}{n^{c_1}} H'(\mu_W I_n - D_n) X_{(1)} \right\|_F^2 \\
&= \sum_{k=1}^{p_n - q} \sum_{l=1}^{q} \left[\frac{1}{n^{\frac{c_3}{2}}} \times \frac{1}{n^{c_1 - \frac{c_3}{2}}} \sum_{i=1}^{n} h_{k,i} x_{i,l} (\mu_W - W_i) \right]^2 \\
&= \mathcal{O}(p_n) \times o\left(n^{-c_3}\right) = o(1) \quad a.s..
\end{aligned}$$

Finally, by Lemma A.2,

$$\left\| n^{1-c_1} \widetilde{C}^w_n \right\|_2 \le \left\| \frac{1}{n^{c_1}} H'(\mu_W I_n - D_n) X_{(1)} \right\|_2 \left\| \left(C^w_{n(11)}\right)^{-1} \right\|_2 = o_p(1). \quad \square$$

Lemma A.4. *Suppose that* $p_n = p$ *is fixed. Assume (2.2) and (2.4). Then, as* $n \to \infty$,

$$\frac{\mu_W}{n} X' D_n X \xrightarrow{a.s.} \mu_W C.$$

Proof. Due to assumption (2.2), the Strong Law of Large Numbers gives

$$\frac{1}{n}X'(D_n - \mu_W I_n)X = \frac{1}{n}\sum_{i=1}^{n}(W_i - \mu_W)\boldsymbol{x}_i\boldsymbol{x}_i' \xrightarrow{\text{a.s.}} \mathbf{0},$$

where $\boldsymbol{x}_i$ is the i^{th} row of X. Then, due to assumption (2.4),

$$\frac{1}{n}X'D_nX = \frac{1}{n}X'(D_n - \mu_W I_n)X + \frac{\mu_W}{n}X'X \xrightarrow{\text{a.s.}} \mathbf{0} + \mu_W C = \mu_W C. \qquad \square$$

An immediate consequence of Lemma A.4 is that when p is fixed,

$$C^w_{n(ij)} \xrightarrow{\text{a.s.}} \mu_W C_{ij} \quad \forall\, i, j = 1, 2.$$

We remind readers that in this paper, we consider a common probability space $P = P_D \times P_W$, which correspond to the two sources of randomness $(\boldsymbol{\epsilon}, \boldsymbol{W})$. Note that the product probability space highlights the fact that the random weights $\boldsymbol{W}$ are drawn independently from the data D. The rest of the proofs deals with convergence of conditional probabilities/distributions (given data, i.e. given $\mathcal{F}_n$) for expressions containing $\boldsymbol{\epsilon}$, where the convergence takes place almost surely under P_D (i.e. for almost every data set). See Mason and Newton (1992) for relevant background.

Lemma A.5. *Assume (2.1). Then*

$$\frac{\boldsymbol{\epsilon}'D_n\boldsymbol{\epsilon}}{n} \xrightarrow{\text{c.p.}} \mu_W\sigma_\epsilon^2 \quad a.s.\ P_D.$$

Proof. Clearly,

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \to \sigma_\epsilon^2 \quad a.s.\ P_D.$$

Due to assumption (2.1),

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^4 = \mathcal{O}(1) \quad a.s.\ P_D,$$

which leads to

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(\epsilon_i^4 W_i^2|\mathcal{F}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\epsilon_i^4\mathbb{E}(W_i^2) = \frac{\sigma_W^2 + \mu_W^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^4\right) \to 0\ a.s.\ P_D.$$

Hence, by the Weak Law of Large Numbers (e.g., Theorem 1.14(ii) of Shao (2003)),

$$\frac{\boldsymbol{\epsilon}'D_n\boldsymbol{\epsilon}}{n} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2(W_i - \mu_W) + \frac{\mu_W}{n}\sum_{i=1}^{n}\epsilon_i^2 \xrightarrow{\text{c.p.}} 0 + \mu_W\sigma_\epsilon^2 = \mu_W\sigma_\epsilon^2 \quad a.s.\ P_D. \qquad \square$$

Lemma A.6. *Assume (2.1), (2.2) and (2.3). Then for any $c > 0$,*

$$\frac{1}{n^c}\boldsymbol{Z}^w_{n(1)} = o_p(1) \quad a.s.\ P_D.$$

Proof. Let x_{ij} be the $(i,j)^{th}$ element of $X_{(1)}$. Then, we can rewrite

$$\begin{aligned}\left(\frac{1}{n^c}\left\|\boldsymbol{Z}^w_{n(1)}\right\|_2\right)^2 &= \frac{1}{n^{2c}}\sum_{j=1}^{q}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i-\mu_W)+\frac{\mu_W}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i x_{ji}\right)^2\\ &= \sum_{j=1}^{q}\left(\frac{1}{n^{\frac{1}{2}+c}}\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i-\mu_W)+\frac{\mu_W}{n^{\frac{1}{2}+c}}\sum_{i=1}^{n}\epsilon_i x_{ji}\right)^2.\end{aligned}$$

Due to assumptions (2.1) and (2.2) and that F_W has finite fourth moment, we could deploy the Lindeberg's Central Limit Theorem

$$\frac{\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i-\mu_W)}{\sqrt{\sigma_W^2\sum_{i=1}^{n}\epsilon_i^2 x_{ji}^2}} \xrightarrow{\text{c.d.}} N(0,1) \quad a.s.\ P_D;$$

see Ng (2022) for more details on how the Liapounov sufficient condition is satisfied a.s. P_D. Subsequently, for all $j = 1, \cdots, q$,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i-\mu_W)\\ &= \sqrt{\frac{\sigma_W^2}{n}\sum_{i=1}^{n}\epsilon_i^2 x_{ji}^2}\times\frac{\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i-\mu_W)}{\sqrt{\sigma_W^2\sum_{i=1}^{n}\epsilon_i^2 x_{ji}^2}}\\ &= \mathcal{O}_p(1) \quad a.s.\ P_D,\end{aligned}$$

and hence,

$$\frac{1}{n^{\frac{1}{2}+c}}\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i-\mu_W) = o_p(1) \quad a.s.\ P_D.$$

Finally, by assumption (2.2) and Lemma A.1,

$$\frac{\mu_W}{n^{\frac{1}{2}+c}}\sum_{i=1}^{n}\epsilon_i x_{ji} \to 0 \quad a.s.\ P_D$$

for all $j = 1, \cdots, q$. Since q is fixed,

$$\left(\frac{1}{n^c}\left\|\boldsymbol{Z}^w_{n(1)}\right\|_2\right)^2 = o_p(1) \quad a.s.\ P_D,$$

and the result follows. □

If we assume that $C_{n(11)} \to C_{11}$ for some nonsingular matrix C_{11} in Lemma A.6, notations could be simplified in the preceding proof by using Cramer-Wold device. We point out to readers that the $C_{n(11)} \to C_{11}$ assumption is required in Theorem 3.4 but not in Theorem 3.1. The following proof contains some interim results that will be utilized in the proof of Theorem 3.4.

Specifically, let $\boldsymbol{x}_{i(1)}$ be the i^{th} row of $X_{(1)}$. Then, for every $\boldsymbol{z} \in \mathbb{R}^q$,

$$\begin{aligned}
&\boldsymbol{z}' \left[\frac{1}{\sqrt{n}} X'_{(1)} (D_n - \mu_W I_n) \boldsymbol{\epsilon} \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i (W_i - \mu_W) \boldsymbol{z}' \boldsymbol{x}_{i(1)} \\
&= \sqrt{\frac{\sigma_W^2}{n} \sum_{i=1}^{n} \epsilon_i^2 \left(\boldsymbol{z}' \boldsymbol{x}_{i(1)} \right)^2} \times \frac{\sum_{i=1}^{n} \epsilon_i (W_i - \mu_W) \boldsymbol{z}' \boldsymbol{x}_{i(1)}}{\sqrt{\sigma_W^2 \sum_{i=1}^{n} \epsilon_i^2 \left(\boldsymbol{z}' \boldsymbol{x}_{i(1)} \right)^2}}.
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \left(\boldsymbol{z}' \boldsymbol{x}_{i(1)} \right)^2 &= \boldsymbol{z}' \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \boldsymbol{x}_{i(1)} \boldsymbol{x}'_{i(1)} \right) \boldsymbol{z} \\
&= \boldsymbol{z}' \left(\sigma_\epsilon^2 C_{n(11)} + \frac{1}{n} \sum_{i=1}^{n} \left(\epsilon_i^2 - \sigma_\epsilon^2 \right) \boldsymbol{x}_{i(1)} \boldsymbol{x}'_{i(1)} \right) \boldsymbol{z} \\
&\to \boldsymbol{z}' \left(\sigma_\epsilon^2 C_{11} \right) \boldsymbol{z} \quad a.s.\ P_D
\end{aligned}$$

due to assumption (2.2) and the Strong Law of Large Numbers. Next, by assuming (2.1) and (2.2) and that F_W has finite fourth moment, we could deploy the Lindeberg's Central Limit Theorem

$$\frac{\sum_{i=1}^{n} \epsilon_i (W_i - \mu_W) \boldsymbol{z}' \boldsymbol{x}_{i(1)}}{\sqrt{\sigma_W^2 \sum_{i=1}^{n} \epsilon_i^2 \left(\boldsymbol{z}' \boldsymbol{x}_{i(1)} \right)^2}} \xrightarrow{\text{c.d.}} N(0,1) \quad a.s.\ P_D;$$

see Ng (2022) for more details on how the Liapounov sufficient condition is satisfied a.s. P_D. Then, by Slutsky's Theorem, for every $\boldsymbol{z} \in \mathbb{R}^q$,

$$\boldsymbol{z}' \left[\frac{1}{\sqrt{n}} X'_{(1)} (D_n - \mu_W I_n) \boldsymbol{\epsilon} \right] \xrightarrow{\text{c.d.}} N\left(0 \ , \ \boldsymbol{z}' \left(\sigma_W^2 \sigma_\epsilon^2 C_{11} \right) \boldsymbol{z} \right).$$

and by Cramer-Wold device,

$$\frac{1}{\sqrt{n}} X'_{(1)} (D_n - \mu_W I_n) \boldsymbol{\epsilon} \xrightarrow{\text{c.d.}} N_q \left(\mathbf{0} \ , \ \sigma_W^2 \sigma_\epsilon^2 C_{11} \right),$$

Since assumption (2.2) and Lemma A.1 ensure that for any $c > 0$,

$$\frac{1}{n^{\frac{1}{2}+c}} X'_{(1)} \boldsymbol{\epsilon} \to \mathbf{0} \quad a.s.\ P_D,$$

we finally have

$$\frac{1}{n^c}\boldsymbol{Z}^w_{n(1)} = \frac{1}{n^c}\left[\frac{1}{\sqrt{n}}X'_{(1)}(D_n - \mu_W I_n)\boldsymbol{\epsilon}\right] + \frac{\mu_W}{n^{\frac{1}{2}+c}}X'_{(1)}\boldsymbol{\epsilon} = o_p(1) \quad a.s.\ P_D.$$

Lemma A.7. *Assume (2.1), (2.2) and (2.3).*

(a) If there exists $\frac{1}{2} < c_1 < c_2 < 1.5 - c_1$ *and* $0 \leq c_3 < 2(c_2 - c_1)$ *for which* $p_n = \mathcal{O}(n^{c_3})$*, then*

$$\frac{1}{n^{c_2-\frac{1}{2}}}\left\|\boldsymbol{Z}^w_{n(3)}\right\|_2 = o_p(1) \quad a.s.\ P_D.$$

(b) If there exists $\frac{1}{2} < c_1 < c_2 < 1.5 - c_1$ *and* $0 \leq c_3 < \frac{2}{3}(c_2 - c_1)$ *for which* $p_n = \mathcal{O}(n^{c_3})$*, then*

$$\frac{p_n - q}{n^{c_2-\frac{1}{2}}}\left\|\boldsymbol{Z}^w_{n(3)}\right\|_2 = o_p(1) \quad a.s.\ P_D.$$

Proof. By using assumptions (2.1) and (2.2) and that F_W has finite fourth moment, we could deploy the Lindeberg's Central Limit Theorem

$$\frac{\sum_{i=1}^n h_{ji}\epsilon_i(W_i - \mu_W)}{\sqrt{\sigma_W^2\sum_{i=1}^n h_{ji}^2\epsilon_i^2}} \xrightarrow{\text{c.d.}} N(0,1) \quad a.s.\ P_D;$$

see Ng (2022) for more details on how the Liapounov sufficient condition is satisfied a.s. P_D. Thus, for all $j = 1, \cdots, p_n - q$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n h_{ji}\epsilon_i(W_i - \mu_W) \\
&= \sqrt{\frac{\sigma_W^2}{n}\sum_{i=1}^n h_{ji}^2\epsilon_i^2} \times \frac{\sum_{i=1}^n h_{ji}\epsilon_i(W_i - \mu_W)}{\sqrt{\sigma_W^2\sum_{i=1}^n h_{ji}^2\epsilon_i^2}} \\
&= O_p(1) \quad a.s.\ P_D,
\end{aligned}$$

which leads to

$$\frac{1}{n^{c_1}}\sum_{i=1}^n h_{ji}\epsilon_i(W_i - \mu_W) = o_p(1) \quad a.s.\ P_D,$$

whereas Lemma A.1 ensures that

$$\frac{1}{n^{c_1}}\sum_{i=1}^n h_{ji}\epsilon_i \to 0 \quad a.s.\ P_D.$$

Therefore, for part (a) of Lemma A.7,

$$\left(\frac{1}{n^{c_2-\frac{1}{2}}}\left\|\boldsymbol{Z}^w_{n(3)}\right\|_2\right)^2$$

$$
\begin{aligned}
&= \frac{n^{2c_1-1}}{n^{2c_2-1}} \sum_{j=1}^{p_n-q} \left(\frac{1}{n^{c_1}} \sum_{i=1}^{n} h_{ji}\epsilon_i(W_i - \mu_W) + \frac{1}{n^{c_1}} \sum_{i=1}^{n} h_{ji}\epsilon_i \right)^2 \\
&= \mathcal{O}\left(n^{2(c_1-c_2)}\right) \times o_p\left(n^{c_3}\right) \quad a.s.\ P_D \\
&= o_p(1) \quad a.s.\ P_D
\end{aligned}
$$

since $c_3 < 2(c_2 - c_1)$.

For part (b) of Lemma A.7,

$$
\begin{aligned}
&\left(\frac{p_n - q}{n^{c_2-\frac{1}{2}}} \left\| \boldsymbol{Z}^w_{n(3)} \right\|_2 \right)^2 \\
&= \mathcal{O}\left(n^{2(c_1-c_2+c_3)}\right) \times o_p\left(n^{c_3}\right) \quad a.s.\ P_D \\
&= o_p(1) \quad a.s.\ P_D
\end{aligned}
$$

since $c_3 < \frac{2}{3}(c_2 - c_1)$. □

Lemma A.8. *Assume (2.2) and that $p_n = p$ is fixed. Then*

$$\frac{1}{n} X' D_n \boldsymbol{\epsilon} \xrightarrow{c.p.} \mathbf{0} \quad a.s.\ P_D.$$

Proof. Let $\boldsymbol{x}_i$ and x_{ij} be the i^{th} row and $(i,j)^{th}$ element of X respectively. Due to assumption (2.2),

$$\frac{1}{n} X' \boldsymbol{\epsilon} \to \mathbf{0} \quad a.s.\ P_D,$$

and for all $j = 1, \cdots, p$,

$$
\begin{aligned}
\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left(x_{ji}^2 \epsilon_i^2 W_i^2 \middle| \mathcal{F}_n \right) &= \frac{1}{n^2} \sum_{i=1}^{n} x_{ji}^2 \epsilon_i^2 \mathbb{E}(W_i^2) \\
&\leq \frac{M_1^2(\sigma_W^2 + \mu_W^2)}{n} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \right) \\
&\to 0 \quad a.s.\ P_D.
\end{aligned}
$$

Hence, by the Weak Law of Large Numbers (e.g., Theorem 1.14(ii) of Shao (2003)),

$$\frac{1}{n} X'(D_n - \mu_W I_n)\boldsymbol{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (W_i - \mu_W) \boldsymbol{x}_i \xrightarrow{\text{c.p.}} \mathbf{0} \quad a.s.\ P_D.$$

Finally,

$$\frac{X' D_n \boldsymbol{\epsilon}}{n} = \frac{1}{n} X'(D_n - \mu_W I_n)\boldsymbol{\epsilon} + \frac{\mu_W}{n} X' \boldsymbol{\epsilon} \xrightarrow{\text{c.p.}} \mathbf{0} \quad a.s.\ P_D.$$ □

Lemma A.9. *Suppose that $p_n = p$ is fixed. Assume (2.1), (2.2), (2.4), and*

$$\frac{1}{\sqrt{n}} X' \boldsymbol{e}_n \to \mathbf{0} \quad a.s.\ P_D,$$

where $\boldsymbol{e}_n$ is the residual of the strongly consistent estimator $\widehat{\boldsymbol{\beta}}_n^{SC}$ of the linear model (1.1). Then,

$$\frac{1}{\sqrt{n}} X' D_n \boldsymbol{e}_n \xrightarrow{c.d.} N_p\left(\mathbf{0}, \sigma_W^2 \sigma_\epsilon^2 C\right) \quad a.s.\ P_D.$$

Proof. Due to assumption (2.4),

$$\frac{\sigma_\epsilon^2}{n} X'X \to \sigma_\epsilon^2 C.$$

Since $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}$ is a strongly consistent estimator of $\boldsymbol{\beta}$ in (1.1), we have

$$\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}} - \boldsymbol{\beta}_0\right) \to \mathbf{0} \quad a.s.\ P_D.$$

Let $\boldsymbol{x}_i$ be the i^{th} row of X, and let e_i be the i^{th} element of $\boldsymbol{e}_n$. Due to assumption (2.2) and Lemma A.1 and the fact that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}$ is strongly consistent,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (e_i^2 - \sigma_\epsilon^2)\boldsymbol{x}_i\boldsymbol{x}_i' &= \frac{1}{n}\sum_{i=1}^n \left(\left[\boldsymbol{x}_i'\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}\right) + \epsilon_i\right]^2 - \sigma_\epsilon^2\right)\boldsymbol{x}_i\boldsymbol{x}_i' \\
&= \frac{1}{n}\sum_{i=1}^n (\epsilon_i^2 - \sigma_\epsilon^2)\boldsymbol{x}_i\boldsymbol{x}_i' \\
&\quad + \frac{2}{n}\sum_{i=1}^n \epsilon_i\left[\boldsymbol{x}_i'\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}\right)\right]\boldsymbol{x}_i\boldsymbol{x}_i' \\
&\quad + \frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{x}_i'\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}\right)\right]^2\boldsymbol{x}_i\boldsymbol{x}_i' \\
&\to \mathbf{0} \quad a.s.\ P_D,
\end{aligned}$$

which leads to

$$\frac{1}{n}\sum_{i=1}^n e_i^2\boldsymbol{x}_i\boldsymbol{x}_i' = \frac{1}{n}\sum_{i=1}^n (e_i^2 - \sigma_\epsilon^2)\boldsymbol{x}_i\boldsymbol{x}_i' + \frac{\sigma_\epsilon^2}{n}X'X \to \sigma_\epsilon^2 C \quad a.s.\ P_D. \tag{A.1}$$

Now for every $\boldsymbol{z} \in \mathbb{R}^p$, consider

$$\begin{aligned}
&\boldsymbol{z}'\left[\frac{1}{\sqrt{n}}X'(D_n - \mu_W I_n)\boldsymbol{e}_n\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i(W_i - \mu_W)(\boldsymbol{z}'\boldsymbol{x}_i)
\end{aligned}$$

$$= \sqrt{\frac{\sigma_W^2}{n}\sum_{i=1}^{n} e_i^2(\boldsymbol{z}'\boldsymbol{x}_i)^2} \times \frac{\sum_{i=1}^{n} e_i(W_i-\mu_W)(\boldsymbol{z}'\boldsymbol{x}_i)}{\sqrt{\sigma_W^2\sum_{i=1}^{n} e_i^2(\boldsymbol{z}'\boldsymbol{x}_i)^2}}.$$

From (A.1), we have

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2(\boldsymbol{z}'\boldsymbol{x}_i)^2 = \boldsymbol{z}'\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\boldsymbol{z} \to \boldsymbol{z}'\left(\sigma_\epsilon^2 C\right)\boldsymbol{z} \quad a.s.\ P_D.$$

Due to assumptions (2.1) and (2.2), as well as the fact that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}$ is strongly consistent and F_W has finite fourth moment, we could deploy the Lindeberg's Central Limit Theorem

$$\frac{\sum_{i=1}^{n} e_i(W_i-\mu_W)(\boldsymbol{z}'\boldsymbol{x}_i)}{\sqrt{\sigma_W^2\sum_{i=1}^{n} e_i^2(\boldsymbol{z}'\boldsymbol{x}_i)^2}} \xrightarrow{\text{c.d.}} N(0,1) \quad a.s.\ P_D;$$

see Ng (2022) for more details on how the Liapounov sufficient condition is satisfied a.s. P_D. Then, by Slutsky's Theorem, for every $\boldsymbol{z} \in \mathbb{R}^p$,

$$\boldsymbol{z}'\left[\frac{1}{\sqrt{n}}X'(D_n-\mu_W I_n)\boldsymbol{e}_n\right] \xrightarrow{\text{c.d.}} N\left(0\ ,\ \boldsymbol{z}'\left(\sigma_W^2\sigma_\epsilon^2 C\right)\boldsymbol{z}\right) \quad a.s.\ P_D,$$

and by Cramer-Wold device,

$$\frac{1}{\sqrt{n}}X'(D_n-\mu_W I_n)\boldsymbol{e}_n \xrightarrow{\text{c.d.}} N_p\left(\boldsymbol{0}\ ,\ \sigma_W^2\sigma_\epsilon^2 C\right) \quad a.s.\ P_D.$$

Finally,

$$\frac{1}{\sqrt{n}}X'D_n\boldsymbol{e}_n \xrightarrow{\text{c.d.}} N_p\left(\boldsymbol{0}\ ,\ \sigma_W^2\sigma_\epsilon^2 C\right) \quad a.s.\ P_D$$

since by assumption (3.2),

$$\frac{\mu_W}{\sqrt{n}}X'\boldsymbol{e}_n \to \boldsymbol{0} \quad a.s.\ P_D. \qquad \square$$

We are now ready to prove the main results presented in the main text. The proof of Proposition 3.1 is similar to that of Proposition 1 of Zhao and Yu (2006).

Proof of Proposition 3.1. First, we note that since $\mathrm{rank}(X) = p_n$, where $p_n \leq n$, the solution to (1.3) is unique by Osborne, Presnell and Turlach (2000) and Tibshirani (2013). We begin with weighting scheme (1.6). Results for the other two simpler weighting schemes could then be easily inferred.

$$\widehat{\boldsymbol{\beta}}_n^w = \arg\min_{\boldsymbol{\beta}}\left\{\frac{1}{n}(Y-X\boldsymbol{\beta})'D_n(Y-X\boldsymbol{\beta}) + \frac{\lambda_n}{n}\sum_{j=1}^{p_n} W_{0,j}|\beta_j|\right\}$$

$$= \arg\min_{\boldsymbol{\beta}}\left\{\frac{1}{n}[\boldsymbol{\epsilon}-X(\boldsymbol{\beta}-\boldsymbol{\beta}_0)]'D_n[\boldsymbol{\epsilon}-X(\boldsymbol{\beta}-\boldsymbol{\beta}_0)]\right.$$

$$+\frac{\lambda_n}{n}\sum_{j=1}^{p_n} W_{0,j}|\beta_{0,j}+\beta_j-\beta_{0,j}|\Bigg\}.$$

Therefore,

$$\begin{aligned}
&(\widehat{\boldsymbol{\beta}}_n^w-\boldsymbol{\beta}_0)\\
&=\arg\min_{\boldsymbol{u}_n}\left\{\frac{1}{n}(\boldsymbol{\epsilon}-X\boldsymbol{u}_n)'D_n(\boldsymbol{\epsilon}-X\boldsymbol{u}_n)+\frac{\lambda_n}{n}\sum_{j=1}^{p_n}W_{0,j}|\beta_{0,j}+u_{n,j}|\right\}\\
&=\arg\min_{\boldsymbol{u}_n}\left\{\boldsymbol{u}_n'\left(\frac{X'D_nX}{n}\right)\boldsymbol{u}_n-2\boldsymbol{u}_n'\left(\frac{X'D_n\boldsymbol{\epsilon}}{n}\right)+\frac{\boldsymbol{\epsilon}'D_n\boldsymbol{\epsilon}}{n}\right.\\
&\qquad\left.+\frac{\lambda_n}{n}\sum_{j=1}^{p_n}W_{0,j}|\beta_{0,j}+u_{n,j}|\right\}.
\end{aligned}$$

The term $(\boldsymbol{\epsilon}'D_n\boldsymbol{\epsilon})/n$ could be dropped since for every n, it does not contain $\boldsymbol{u}_n$ and Lemma A.5 ensures that it converges in conditional probability to a finite limit. Differentiating the first two terms with respect to $\boldsymbol{u}_n$ yields

$$\frac{1}{n}\left\{2X'D_nX\boldsymbol{u}_n-2X'D_n\boldsymbol{\epsilon}\right\}=\frac{1}{n}\left\{2\sqrt{n}\left[C_n^w\left(\sqrt{n}\boldsymbol{u}_n\right)-\boldsymbol{Z}_n^w\right]\right\}.$$

For $j=1,\cdots,p_n$, considering sub-differentials of the penalty term with respect to $u_{n,j}$ yields

$$\begin{aligned}
&\begin{cases}\frac{\lambda_n}{n}W_{0,j}\times\operatorname{sgn}\left(\beta_{0,j}+u_{n,j}\right) & \text{for } \beta_{0,j}+u_{n,j}\neq 0\\ \frac{\lambda_n}{n}W_{0,j}\times[-1,1] & \text{for } \beta_{0,j}+u_{n,j}=0\end{cases}\\
&=\begin{cases}\frac{\lambda_n}{n}W_{0,j}\times\operatorname{sgn}\left(\widehat{\beta}_{n,j}^w\right) & \text{for } \widehat{\beta}_{n,j}^w\neq 0\\ \frac{\lambda_n}{n}W_{0,j}\times[-1,1] & \text{for } \widehat{\beta}_{n,j}^w=0\end{cases}
\end{aligned}$$

Note that $\widehat{\boldsymbol{\beta}}_n^w=\widehat{\boldsymbol{u}}_n+\boldsymbol{\beta}_0$, which can be partitioned into

$$\widehat{\boldsymbol{\beta}}_n^w=\begin{bmatrix}\widehat{\boldsymbol{\beta}}_{n(1*)}^w\\ \widehat{\boldsymbol{\beta}}_{n(2*)}^w\end{bmatrix},$$

where $\widehat{\boldsymbol{\beta}}_{n(1*)}^w$ consists of non-zero elements of $\widehat{\boldsymbol{\beta}}_n^w$, and $\widehat{\boldsymbol{\beta}}_{n(2*)}^w=\mathbf{0}$. The asterisk here is to distinguish the partition of random-weighting samples $\widehat{\boldsymbol{\beta}}_n^w$ from the true partition of $\boldsymbol{\beta}_0$. It follows that

$$\begin{aligned}
&2\sqrt{n}\left[C_n^w\left(\sqrt{n}\widehat{\boldsymbol{u}}_n\right)-\boldsymbol{Z}_n^w\right]\\
&=2\sqrt{n}\left\{\begin{bmatrix}C_{n(11*)}^w & C_{n(12*)}^w\\ C_{n(21*)}^w & C_{n(22*)}^w\end{bmatrix}\times\sqrt{n}\begin{bmatrix}\widehat{\boldsymbol{u}}_{n(1*)}\\ \widehat{\boldsymbol{u}}_{n(2*)}\end{bmatrix}-\begin{bmatrix}\boldsymbol{Z}_{n(1*)}^w\\ \boldsymbol{Z}_{n(2*)}^w\end{bmatrix}\right\}.
\end{aligned}$$

Note that $\widehat{\boldsymbol{u}}_{n(2*)}$ does not necessarily equal to $\mathbf{0}$ unless the partition of the random-weighting samples $\widehat{\boldsymbol{\beta}}^w_n$ coincides with the true partition of $\boldsymbol{\beta}_0$. As a consequence of the Karush-Kuhn-Tucker (KKT) conditions, we have

$$C^w_{n(11*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1*)}\right] + C^w_{n(12*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(2*)}\right] - \boldsymbol{Z}^w_{n(1*)} = -\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left(\widehat{\boldsymbol{\beta}}^w_{n(1*)}\right) \tag{A.2}$$

and

$$\left|C^w_{n(21*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1*)}\right] + C^w_{n(22*)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(2*)}\right] - \boldsymbol{Z}^w_{n(2*)}\right| \leq \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(2)} \tag{A.3}$$

element-wise. Meanwhile, we also note that

$$\begin{aligned}\left\{\left|\widehat{\boldsymbol{u}}_{n(1)}\right| < \left|\boldsymbol{\beta}_{0(1)}\right|\right\} &= \left\{\widehat{\boldsymbol{u}}_{n(1)} < \left|\boldsymbol{\beta}_{0(1)}\right|\right\} \bigcap \left\{\widehat{\boldsymbol{u}}_{n(1)} > -\left|\boldsymbol{\beta}_{0(1)}\right|\right\} \\ &= \left\{\widehat{\boldsymbol{\beta}}^w_{n(1)} < \boldsymbol{\beta}_{0(1)} + \left|\boldsymbol{\beta}_{0(1)}\right|\right\} \bigcap \left\{\widehat{\boldsymbol{\beta}}^w_{n(1)} > \boldsymbol{\beta}_{0(1)} - \left|\boldsymbol{\beta}_{0(1)}\right|\right\},\end{aligned}$$

where all inequalities hold element-wise. Thus, $\widehat{\boldsymbol{\beta}}^w_{n(1)} < 0$ element-wise if $\boldsymbol{\beta}_{0(1)} < 0$ element-wise, and vice versa. In other words,

$$\left\{\operatorname{sgn}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)}\right) = \operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right)\right\} \supseteq \left\{\left|\widehat{\boldsymbol{u}}_{n(1)}\right| < \left|\boldsymbol{\beta}_{0(1)}\right| \text{ element-wise}\right\}. \tag{A.4}$$

Therefore, by (A.2), (A.3), (A.4), and uniqueness of solution for the random-weighting setup (1.3), if there exists $\widehat{\boldsymbol{u}}_n$ such that the following equation and inequalities hold:

$$C^w_{n(11)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}^w_{n(1)} = -\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \tag{A.5}$$

$$-\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(2)} \leq C^w_{n(21)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}^w_{n(2)} \leq \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(2)} \text{ element-wise} \tag{A.6}$$

$$\left|\widehat{\boldsymbol{u}}_{n(1)}\right| < \left|\boldsymbol{\beta}_{0(1)}\right| \quad \text{element-wise,} \tag{A.7}$$

then we have $\operatorname{sgn}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)}\right) = \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]$ and $\widehat{\boldsymbol{u}}_{n(2)} = \widehat{\boldsymbol{\beta}}^w_{n(2)} = \boldsymbol{\beta}_{0(2)} = \mathbf{0}$, ie.

$$\widehat{\boldsymbol{\beta}}^w_n \overset{s}{=} \boldsymbol{\beta}_0,$$

and

$$\begin{aligned}&P\left(\widehat{\boldsymbol{\beta}}^w_n \overset{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n\right) \\ &\geq P\Bigg(\left\{\left|C^w_{n(21)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}^w_{n(2)}\right| \leq \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(2)} \text{ element-wise}\right\} \\ &\qquad \bigcap \left\{C^w_{n(11)}\left[\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)}\right] - \boldsymbol{Z}^w_{n(1)} = -\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]\right\}\end{aligned}$$

$$\bigcap \left\{ \left| \widehat{\boldsymbol{u}}_{n(1)} \right| < \left| \boldsymbol{\beta}_{0(1)} \right| \text{ element-wise} \right\} \Big| \mathcal{F}_n \Big).$$

Now we proceed to simplify these equation and inequalities (A.5), (A.6) and (A.7). Equation (A.5) can be re-written as

$$\sqrt{n}\widehat{\boldsymbol{u}}_{n(1)} = \left(C^w_{n(11)} \right)^{-1} \left[\boldsymbol{Z}^w_{n(1)} - \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right]. \tag{A.8}$$

Substituting inequality (A.7) into equation (A.8) above leads to A^w_n. Replace the expression

$$\boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right]$$

in equation (A.8) with $W_0 \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right]$ and $\operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right]$ for weighting schemes (1.5) and (1.4) respectively to obtain A^w_n.

Next, substituting equation (A.8) into inequality (A.6) and simple arithmetic yield

$$\begin{aligned} \widetilde{B}^w_n \equiv \Bigg\{ & \left| \widetilde{C}^w_n \boldsymbol{Z}^w_{n(1)} + \boldsymbol{Z}^w_{n(3)} - \frac{\lambda_n}{2\sqrt{n}} C^w_{n(21)} \left(C^w_{n(11)} \right)^{-1} \boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right| \\ & - \frac{\lambda_n}{2\sqrt{n}} \left| C_{n(21)} C^{-1}_{n(11)} \boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right| \\ & \leq \frac{\lambda_n}{2\sqrt{n}} \left(\boldsymbol{W}_{0(2)} - \left| C_{n(21)} C^{-1}_{n(11)} \boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right| \right) \text{ element-wise} \Bigg\} \end{aligned}$$

for weighting scheme (1.6). Now, observe that $B^w_n \subseteq \widetilde{B}^w_n$, since (LHS of B^w_n) $\geq$ (LHS of $\widetilde{B}^w_n$) element-wise. Thus,

$$P\left(\widehat{\boldsymbol{\beta}}^w_n \overset{s}{=} \boldsymbol{\beta}_0 \Big| \mathcal{F}_n \right) \geq P\left(A^w_n \cap \widetilde{B}^w_n \big| \mathcal{F}_n \right) \geq P\left(A^w_n \cap B^w_n \big| \mathcal{F}_n \right).$$

For weighting scheme (1.5),

$$\begin{aligned} \widetilde{B}^w_n \equiv \Bigg\{ & \left| \widetilde{C}^w_n \boldsymbol{Z}^w_{n(1)} + \boldsymbol{Z}^w_{n(3)} - \frac{\lambda_n W_0}{2\sqrt{n}} C^w_{n(21)} \left(C^w_{n(11)} \right)^{-1} \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right| \\ & - \frac{\lambda_n W_0}{2\sqrt{n}} \left| C_{n(21)} C^{-1}_{n(11)} \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right| \\ & \leq \frac{\lambda_n W_0}{2\sqrt{n}} \left(\mathbf{1}_{p_n - q} - \left| C_{n(21)} C^{-1}_{n(11)} \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)} \right] \right| \right) \text{ element-wise} \Bigg\}. \end{aligned} \tag{A.9}$$

Now, observe that $B^w_n \subseteq \widetilde{B}^w_n$, since (LHS of B^w_n) $\geq$ (LHS of $\widetilde{B}^w_n$) element-wise, whereas (RHS of B^w_n) $\leq$ (RHS of $\widetilde{B}^w_n$) element-wise due to the Irrepresentable condition (3.1). Therefore,

$$P\left(\widehat{\boldsymbol{\beta}}^w_n \overset{s}{=} \boldsymbol{\beta}_0 \Big| \mathcal{F}_n \right) \geq P\left(A^w_n \cap \widetilde{B}^w_n \big| \mathcal{F}_n \right) \geq P\left(A^w_n \cap B^w_n \big| \mathcal{F}_n \right).$$

For weighting scheme (1.4), substitute $W_0 = 1$ in (A.9) and the result follows. □

Proof of Theorem 3.1. From Proposition 3.1,

$$\begin{aligned} P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \stackrel{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n\right) &\geq P\left(A_n^w \bigcap B_n^w \middle| \mathcal{F}_n\right) \\ &= 1 - P\left[(A_n^w)^c \bigcup (B_n^w)^c \middle| \mathcal{F}_n\right] \\ &\geq 1 - \left\{P\left[(A_n^w)^c \middle| \mathcal{F}_n\right] + P\left[(B_n^w)^c \middle| \mathcal{F}_n\right]\right\}. \end{aligned}$$

We now investigate the conditional probabilities $P\left[(A_n^w)^c \middle| \mathcal{F}_n\right]$ and $P\left[(B_n^w)^c \middle| \mathcal{F}_n\right]$ separately. All three weighting schemes (1.4), (1.5) and (1.6) share very similar $P\left[(A_n^w)^c \middle| \mathcal{F}_n\right]$. We start off with the most general version (1.6) of the weighting schemes. Results for the other two simpler weighting schemes could then be easily inferred. For ease of notation, let

$$\boldsymbol{z}_n = [z_{n,1}, \cdots, z_{n,q}]' := \left(C_{n(11)}^w\right)^{-1} \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n}{2\sqrt{n}} \boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]\right).$$

Note that

$$\frac{\lambda_n}{2n} \boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right] \xrightarrow{p} \mathbf{0}.$$

Hence, by Lemmas A.2 and A.6,

$$\begin{aligned} P\left[(A_n^w)^c \middle| \mathcal{F}_n\right] &= P\left(\bigcup_{j=1}^{q} \left\{ |z_{n,j}| > \sqrt{n} \left|\beta_{0,j}\right| \right\} \middle| \mathcal{F}_n\right) \\ &\leq \sum_{j=1}^{q} P\left(\frac{1}{\sqrt{n}} |z_{n,j}| > \left|\beta_{0,j}\right| \middle| \mathcal{F}_n\right) \\ &\to 0 \quad a.s.\ P_D, \end{aligned}$$

because for all $j = 1, \cdots, q$, we have $|\beta_{0,j}| > 0$ but

$$\frac{1}{\sqrt{n}} |z_{n,j}| = o_p(1) \quad a.s.\ P_D.$$

For weighting schemes (1.5) and (1.4), replace the expression

$$\boldsymbol{W}_{0(1)} \circ \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]$$

with $W_0 \operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]$ and $\operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]$ respectively to obtain the same result

$$P\left[(A_n^w)^c \middle| \mathcal{F}_n\right] \to 0 \quad a.s.\ P_D.$$

We now turn our attention to $P\left[(B_n^w)^c \middle| \mathcal{F}_n\right]$, where weighting scheme (1.6) is markedly different – and derived separately – from weighting schemes (1.4) and (1.5). We first consider weighting scheme (1.5), and then infer the result for weighting scheme (1.4) as a special case. For ease of notation, define

$$\boldsymbol{\zeta}_n = [\zeta_{n,1}, \cdots, \zeta_{n,p_n-q}]' := \boldsymbol{Z}_{n(3)}^w,$$

$$\boldsymbol{\nu}_n = [\nu_{n,1}, \cdots, \nu_{n,p_n-q}]' := \widetilde{C}_n^w \left(\boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n W_0}{2\sqrt{n}} \mathrm{sgn}\left[\boldsymbol{\beta}_{0(1)}\right] \right).$$

Then, for any $\xi > 0$,

$$\begin{aligned}
& P\left[(B_n^w)^c \,\middle|\mathcal{F}_n\right] \\
&= P\left(\bigcup_{j=1}^{p_n-q} \left\{ |\zeta_{n,j} + \nu_{n,j}| > \frac{\lambda_n}{2\sqrt{n}} \eta_j \right\} \middle| \mathcal{F}_n \right) \\
&\le P\left(\bigcup_{j=1}^{p_n-q} \left\{ |\zeta_{n,j}| > \frac{\lambda_n}{2\sqrt{n}} \eta_j - \xi \right\} \middle| \mathcal{F}_n \right) + P\left(\bigcup_{j=1}^{p_n-q} \{ |\nu_{n,j}| > \xi \} \middle| \mathcal{F}_n \right) \\
&\le \quad P\left(\bigcup_{j=1}^{p_n-q} \left\{ |\zeta_{n,j}| > \frac{\lambda_n W_0}{2\sqrt{n}} \eta_j - \xi \right\} \middle| \mathcal{F}_n \right) + P\left(\|\boldsymbol{\nu}_n\|_2 > \xi \middle| \mathcal{F}_n \right).
\end{aligned}$$

Since

$$\frac{\lambda_n W_0}{n^{1.5-c_1}} \mathrm{sgn}\left[\boldsymbol{\beta}_{0(1)}\right] = o_p(1),$$

we have, by Lemmas A.3 and A.6,

$$\|\boldsymbol{\nu}_n\|_2 \le \left\| n^{1-c_1} \widetilde{C}_n^w \right\|_2 \left\| \frac{1}{n^{1-c_1}} \boldsymbol{Z}_{n(1)}^w - \frac{\lambda_n W_0}{2n^{1.5-c_1}} \mathrm{sgn}\left[\boldsymbol{\beta}_{0(1)}\right] \right\|_2 = o_p(1) \quad a.s.\ P_D,$$

and thus,

$$P\left(\|\boldsymbol{\nu}_n\|_2 > \xi \middle| \mathcal{F}_n \right) = o(1) \quad a.s.\ P_D.$$

Now, let

$$\eta_* = \min_{1 \le j \le p_n - q} \eta_j,$$

and note that $0 < \eta_* \le 1$ from assumption (3.1). Then,

$$\begin{aligned}
& P\left(\bigcup_{j=1}^{p_n-q} \left\{ |\zeta_{n,j}| > \frac{\lambda_n W_0}{2\sqrt{n}} \eta_j - \xi \right\} \middle| \mathcal{F}_n \right) \\
&\le P\left(\left\| \boldsymbol{\zeta}_n \right\|_2 > \frac{\lambda_n W_0}{2\sqrt{n}} \eta_* - \xi \middle| \mathcal{F}_n \right) \\
&= P\left(\frac{1}{n^{c_2 - \frac{1}{2}}} \left(\left\| \boldsymbol{\zeta}_n \right\|_2 + \xi \right) > \frac{\lambda_n W_0}{2n^{c_2}} \eta_* \middle| \mathcal{F}_n \right) \\
&= o(1) \quad a.s.\ P_D,
\end{aligned}$$

because

$$\frac{\lambda_n W_0}{2n^{c_2}} \eta_* = \mathcal{O}_p(1)$$

whereas part (a) of Lemma A.7 ensures that

$$\frac{1}{n^{c_2-\frac{1}{2}}}\left(\left\|\boldsymbol{\zeta}_n\right\|_2+\xi\right)=o_p(1)\quad a.s.\ P_D.$$

Thus, for weighting scheme (1.5), we have just shown that

$$P\left[\left(B_n^w\right)^c\left|\mathcal{F}_n\right.\right]=o(1)\quad a.s.\ P_D.$$

For weighting scheme (1.4), take $W_0=1$ and repeat the preceding steps to obtain the same result.

Now, for weighting scheme (1.6), define

$$\begin{aligned}\boldsymbol{\nu}_n=\left[\nu_{n,1},\cdots,\nu_{n,p_n-q}\right]'&:=\widetilde{C}_n^w\left(\boldsymbol{Z}_{n(1)}^w-\frac{\lambda_n}{2\sqrt{n}}\boldsymbol{W}_{0(1)}\circ\operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]\right),\\ \boldsymbol{\gamma}_n=\left[\gamma_{n,1},\cdots,\gamma_{n,p_n-q}\right]'&:=C_{n(21)}C_{n(11)}^{-1}\boldsymbol{W}_{0(1)}\circ\operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right].\end{aligned}$$

and for any $\xi>0$,

$$\begin{aligned}&P\left[\left(B_n^w\right)^c\left|\mathcal{F}_n\right.\right]\\&=P\left(\bigcup_{j=1}^{p_n-q}\left\{\left|\zeta_{n,j}+\nu_{n,j}\right|>\frac{\lambda_n}{2\sqrt{n}}\left(W_{0(2),j}-\left|\gamma_{n,j}\right|\right)\right\}\middle|\mathcal{F}_n\right)\\&\le P\left(\bigcup_{j=1}^{p_n-q}\left\{\left|\zeta_{n,j}\right|>\frac{\lambda_n}{2\sqrt{n}}\left(W_{0(2),j}-\left|\gamma_{n,j}\right|\right)-\xi\right\}\middle|\mathcal{F}_n\right)+P\left(\left\|\boldsymbol{\nu}_n\right\|_2>\xi\middle|\mathcal{F}_n\right).\end{aligned}$$

Again,

$$\frac{\lambda_n}{n^{1.5-c_1}}\boldsymbol{W}_{0(1)}\circ\operatorname{sgn}\left[\boldsymbol{\beta}_{0(1)}\right]=o_p(1),$$

so, by Lemmas A.3 and A.6,

$$P\left(\left\|\boldsymbol{\nu}_n\right\|_2>\xi\middle|\mathcal{F}_n\right)=o(1)\quad a.s.\ P_D.$$

Notice how the penalty weights $\boldsymbol{W}_{0(1)}$ and $\boldsymbol{W}_{0(2)}$ upend the strong irrepresentable condition (3.1). Specifically,

$$P\left(W_{0(2),j}-\left|\gamma_{n,j}\right|<0\right)>0,$$

which then renders the probability bound to be unhelpful. Instead, notice that from the strong irrepresentable condition (3.1),

$$\gamma_{n,j}\le(1-\eta_*)\times\max_{1\le j\le q}W_{0(1),j}$$

for all $j=1,\cdots,q$. We focus on the more restrictive case where

$$\eta_*=1\Longleftrightarrow\boldsymbol{\eta}=\mathbf{1}_{p_n-q},$$

which leads to a more meaningful probability bound. Then, $\gamma_{n,j} = 0$ for all $j = 1, \cdots, q$, and

$$\begin{aligned} & P\left(\bigcup_{j=1}^{p_n-q}\left\{|\zeta_{n,j}| > \frac{\lambda_n}{2\sqrt{n}} W_{0(2),j} - \xi\right\}\middle|\mathcal{F}_n\right) \\ & \leq P\left(\left\|\boldsymbol{\zeta}_n\right\|_2 > \frac{\lambda_n}{2\sqrt{n}}\left(\min_{1\leq j\leq p_n-q} W_{0(2),j}\right) - \xi\middle|\mathcal{F}_n\right) \\ & = P\left(\frac{1}{n^{c_2-\frac{1}{2}}}\left(\left\|\boldsymbol{\zeta}_n\right\|_2 + \xi\right) > \frac{\lambda_n}{2n^{c_2}}\left(\min_{1\leq j\leq p_n-q} W_{0(2),j}\right)\middle|\mathcal{F}_n\right) \end{aligned}$$

For the case of exponential random weights

$$F_W(w) = 1 - e^{-\theta_w w}$$

for some $\theta_w > 0$, we immediately have

$$\left(\min_{1\leq j\leq p_n-q} W_{0(2)j}\right) \sim \text{Exp}\left((p_n - q)\theta_w\right).$$

Then, by part (b) of Lemma A.7,

$$\begin{aligned} & P\left(\frac{1}{n^{c_2-\frac{1}{2}}}\left(\left\|\boldsymbol{\zeta}_n\right\|_2 + \xi\right) > \frac{\lambda_n}{2n^{c_2}}\left(\min_{1\leq j\leq p_n-q} W_{0(2),j}\right)\middle|\mathcal{F}_n\right) \\ & = P\left(W < \theta_w \frac{2n^{c_2}}{\lambda_n}\frac{p_n - q}{n^{c_2-\frac{1}{2}}}\left(\left\|\boldsymbol{\zeta}_n\right\|_2 + \xi\right)\middle|\mathcal{F}_n\right) \text{ where } W \sim \text{Exp}(1) \\ & = o(1) \quad a.s.\ P_D, \end{aligned}$$

and we have just shown that

$$P\left[\left(B_n^w\right)^c\middle|\mathcal{F}_n\right] = o(1) \quad a.s.\ P_D$$

for weighting scheme (1.6).

Finally,

$$\begin{aligned} & P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \stackrel{s}{=} \boldsymbol{\beta}_0\middle|\mathcal{F}_n\right) \\ & \geq 1 - \left\{P\left[\left(A_n^w\right)^c\middle|\mathcal{F}_n\right] + P\left[\left(B_n^w\right)^c\middle|\mathcal{F}_n\right]\right\} \\ & = 1 - o(1) \quad a.s.\ P_D \end{aligned}$$

for all three weighting schemes (1.4), (1.5) and (1.6). □

Proof of Theorem 3.2. From the proof of Proposition 3.1,

$$(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0)$$

$$= \arg\min_{\boldsymbol{u}} \left\{ \boldsymbol{u}' \left(\frac{X' D_n X}{n} \right) \boldsymbol{u} - 2\boldsymbol{u}' \left(\frac{X' D_n \boldsymbol{\epsilon}}{n} \right) + \frac{\boldsymbol{\epsilon}' D_n \boldsymbol{\epsilon}}{n} + \frac{\lambda_n}{n} \sum_{j=1}^{p} W_{0,j} |\beta_{0,j} + u_{n,j}| \right\}$$

$$:= \arg\min_{\boldsymbol{u}} \quad g_n(\boldsymbol{u}).$$

By Lemmas A.4, A.5 and A.8, for $\frac{\lambda_n}{n} \to \lambda_0 \in [0, \infty)$, Slutsky Theorem gives

$$g_n(\boldsymbol{u}) \xrightarrow{\text{c.d.}} g(\boldsymbol{u}) + \mu_W \sigma_\epsilon^2 \quad a.s.\ P_D.$$

Note that for weighting schemes (1.5) and (1.6), $g(\boldsymbol{u})$ is a random function as it contains random weights. Since $g_n(\boldsymbol{u})$ is convex and $g(\boldsymbol{u})$ has a unique minimum, it follows from Geyer (1996) that

$$\arg\min_{\boldsymbol{u}} g_n(\boldsymbol{u}) \xrightarrow{\text{c.d.}} \arg\min_{\boldsymbol{u}} \left\{ g(\boldsymbol{u}) + \mu_W \sigma_\epsilon^2 \right\} = \arg\min_{\boldsymbol{u}} g(\boldsymbol{u}) \quad a.s.\ P_D.$$

For weighting schemes (1.4), $g(\boldsymbol{u})$ is not a random function. Instead, we note that since $g_n(\boldsymbol{u})$ is convex, it follows from pointwise convergence of conditional probability that

$$\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0 = \mathcal{O}_p(1).$$

For any compact set K, by applying the Convexity Lemma (Pollard, 1991),

$$\sup_{\boldsymbol{u} \in K} \left| g_n(\boldsymbol{u}) - g(\boldsymbol{u}) - \mu_W \sigma_\epsilon^2 \right| \xrightarrow{\text{c.p.}} 0 \quad a.s.\ P_D.$$

Therefore,

$$\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0 \right) = \arg\min_{\boldsymbol{u}} g_n(\boldsymbol{u}) \xrightarrow{\text{c.p.}} \arg\min_{\boldsymbol{u}} g(\boldsymbol{u}) \quad a.s.\ P_D.$$

Finally, for all three weighting schemes, if $\lambda_0 = 0$, $\arg\min_{\boldsymbol{u}} g(\boldsymbol{u}) = \mathbf{0}$, i.e.

$$\widehat{\boldsymbol{\beta}}_n^w \xrightarrow{\text{c.p.}} \boldsymbol{\beta}_0 \quad a.s.\ P_D. \qquad \square$$

Proof of Theorem 3.3. Let $\boldsymbol{e}_n$ be the residual that corresponds to the strongly consistent estimator $\widehat{\boldsymbol{\beta}}_n^{\text{SC}}$ of the linear regression model (1.1), and define

$$Q_n(\boldsymbol{z}) := \left\| D_n^{\frac{1}{2}} (\boldsymbol{y} - X\boldsymbol{z}) \right\|_2^2 + \lambda_n \sum_{j=1}^{p} W_{0,j} |z_j|,$$

which leads to

$$Q_n \left(\widehat{\boldsymbol{\beta}}_n^{\text{SC}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right)$$

$$= \left\| D_n^{\frac{1}{2}} \left(\boldsymbol{e}_n - \frac{1}{\sqrt{n}} X\boldsymbol{u} \right) \right\|_2^2 + \lambda_n \sum_{j=1}^{p} W_{0,j} \left| \widehat{\beta}_{n,j}^{\mathrm{SC}} + \frac{1}{\sqrt{n}} u_j \right|,$$

and

$$Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}\right) = \left\| D_n^{\frac{1}{2}} \left(Y - X\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}} \right) \right\|_2^2 + \lambda_n \sum_{j=1}^{p} W_{0,j} \left| \widehat{\beta}_{n,j}^{\mathrm{SC}} \right|$$
$$= \left\| D_n^{\frac{1}{2}} \boldsymbol{e}_n \right\|_2^2 + \lambda_n \sum_{j=1}^{p} W_{0,j} \left| \widehat{\beta}_{n,j}^{\mathrm{SC}} \right|.$$

Now, define

$$V_n(\boldsymbol{u}) := Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right) - Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}} \right),$$

and note that

$$\arg\min_{\boldsymbol{u}} V_n(\boldsymbol{u}) = \arg\min_{\boldsymbol{u}} Q_n\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}} + \frac{1}{\sqrt{n}} \boldsymbol{u} \right) = \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{w} - \widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}} \right).$$

Notice that $V_n(\boldsymbol{u})$ can be simplified into

$$\boldsymbol{u}' \left(\frac{X' D_n X}{n} \right) \boldsymbol{u} - 2\boldsymbol{u}' \left(\frac{X' D_n \boldsymbol{e}_n}{\sqrt{n}} \right)$$
$$+ \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} W_{0,j} \left(\left| \sqrt{n}\widehat{\beta}_{n,j}^{\mathrm{SC}} + u_j \right| - \left| \sqrt{n}\widehat{\beta}_{n,j}^{\mathrm{SC}} \right| \right),$$

where its penalty term can be expanded into

$$\frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} W_{0,j} \; p_n(u_j)$$
$$:= \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} W_{0,j} \Big\{ \left| \sqrt{n} \left[\beta_{0,j} + \left(\widehat{\beta}_{n,j}^{\mathrm{SC}} - \beta_{0,j} \right) \right] + \mu_j \right|$$
$$- \left| \sqrt{n} \left[\beta_{0,j} + \left(\widehat{\beta}_{n,j}^{\mathrm{SC}} - \beta_{0,j} \right) \right] \right| \Big\}$$

For $\beta_{0,j} \neq 0$,

$$\left(\widehat{\beta}_{n,j}^{\mathrm{SC}} - \beta_{0,j} \right) \to 0 \quad a.s.\ P_D,$$

and hence $\sqrt{n}\beta_{0,j}$ dominates u_j for large n. Thus, it is easy to verify that $p_n(u_j)$ converges to $u_j \mathrm{sgn}\left(\beta_{0,j}\right)$ for all $j \in \{j : \beta_{0,j} \neq 0\}$. Thus, by Lemmas A.4 and A.9, if $q = p$, Slutsky Theorem ensures that

$$V_n(\boldsymbol{u}) \xrightarrow{\text{c.d.}} V(\boldsymbol{u}) := \mu_W \boldsymbol{u}' C \boldsymbol{u} - 2\boldsymbol{u}'\Psi + \lambda_0 \sum_{j=1}^{p} W_j \left[u_j \,\mathrm{sgn}(\beta_{0,j}) \right] \quad a.s.\ P_D,$$

where Ψ has a $N\left(\boldsymbol{0}, \sigma_W^2 \sigma_\epsilon^2 C\right)$ distribution, and

(i) $W_j = 1$ for all j under weighting scheme (1.4),
(ii) $W_j = W_0$ for all j, $W_0 \sim F_W$ and $W_0 \perp \Psi$ under weighting scheme (1.5),
(iii) $W_j \overset{iid}{\sim} F_W$ and $W_j \perp \Psi$ for all j under weighting scheme (1.6).

Since $V_n(\boldsymbol{u})$ is convex and $V(\boldsymbol{u})$ has a unique minimum, it follows from Geyer (1996) that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{\mathrm{SC}}\right) = \arg\min_{\boldsymbol{u}} V_n(\boldsymbol{u}) \xrightarrow{\text{c.d.}} \arg\min_{\boldsymbol{u}} V(\boldsymbol{u}) \quad a.s.\ P_D$$

when $q = p$. In particular, if $\lambda_0 = 0$,

$$\arg\min_{\boldsymbol{u}} V(\boldsymbol{u}) = \frac{1}{\mu_W} C^{-1}\Psi \sim N\left(\mathbf{0}, \frac{\sigma_W^2 \sigma_\epsilon^2}{\mu_W^2} C^{-1}\right).$$

However, if $0 < q < p$, then for $j \in \{j : \beta_{0,j} = 0\}$, $p_n(u_j)$ is back to

$$\left|\sqrt{n}\widehat{\beta}_{n,j}^{\mathrm{SC}} + \mu_j\right| - \left|\sqrt{n}\widehat{\beta}_{n,j}^{\mathrm{SC}}\right|,$$

which depends on the sample path of realized data. This necessitates the Skorokhod argument, thus leading to the penalty term in (3.3). □

We need the following lemma to prove Theorem 3.4:

Lemma A.10. *Consider Liu and Yu (2013)'s unweighted two-step LASSO+LS estimator $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$, with its corresponding set of selected variables denoted as $\widehat{S}_n$. Adopt assumptions (2.2), (2.3) and (3.1). If there exists $\frac{1}{2} < c_1 < c_2 < 1$ and $0 \le c_3 < 2(c_2 - c_1)$ for which $\lambda_n = \mathcal{O}(n^{c_2})$ and $p_n = \mathcal{O}(n^{c_3})$, then as $n \to \infty$,*

$$P\left(\widehat{S}_n = S_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D.$$

Proof. The first step (i.e. the variable selection step) of obtaining $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$ is effectively the standard LASSO procedure. Thus, by assumption (3.1), from the proof of Proposition 1 of Zhao and Yu (2006), we obtain

$$\left\{\widehat{S}_n = S_0\right\} \supseteq \{A_n \cap B_n\}$$

and thus

$$P\left(\widehat{S}_n = S_0 \middle| \mathcal{F}_n\right) \ge P\left(A_n \cap B_n \middle| \mathcal{F}_n\right),$$

where

$$A_n \equiv \left\{\left|C_{n(11)}^{-1}\frac{X_{(1)}'\boldsymbol{\epsilon}}{\sqrt{n}}\right| \le \sqrt{n}\left(\left|\boldsymbol{\beta}_{0(1)}\right| - \frac{\lambda_n}{2n}\left|C_{n(11)}^{-1}\mathrm{sgn}\left(\boldsymbol{\beta}_{0(1)}\right)\right|\right) \text{ element-wise}\right\}$$

$$B_n \equiv \left\{\left|\frac{1}{\sqrt{n}}\left[C_{n(21)}C_{n(11)}^{-1}X_{(1)}' - X_{(2)}'\right]\boldsymbol{\epsilon}\right| \le \frac{\lambda_n}{2\sqrt{n}}\boldsymbol{\eta} \text{ element-wise}\right\}.$$

Next, we want to show that

$$P\left(A_n^c\middle|\mathcal{F}_n\right) \to 0 \ \ a.s.\ P_D \quad \text{and} \quad P\left(B_n^c\middle|\mathcal{F}_n\right) \to 0 \ \ a.s.\ P_D$$

such that

$$P\left(\widehat{S}_n = S_0\middle|\mathcal{F}_n\right) \geq 1 - \left[P\left(A_n^c\middle|\mathcal{F}_n\right) + P\left(B_n^c\middle|\mathcal{F}_n\right)\right] \to 1 \quad a.s.\ P_D.$$

First, by assumptions (2.2) and (2.3), $C_{n(11)}^{-1} = \mathcal{O}(1)$ for all n, whereas

$$\frac{\lambda_n}{2n} C_{n(11)}^{-1} \text{sgn}\left(\boldsymbol{\beta}_{0(1)}\right) \to \mathbf{0}.$$

By Lemma A.1, for any $\frac{1}{2} < c' < 1$,

$$\frac{1}{n^{c'}} X_{(1)}'\boldsymbol{\epsilon} \to \mathbf{0} \quad a.s.\ P_D \quad \Longrightarrow \quad \frac{1}{n^{c'-\frac{1}{2}}}\left(C_{n(11)}^{-1}\frac{X_{(1)}'\boldsymbol{\epsilon}}{\sqrt{n}}\right) \to \mathbf{0} \quad a.s.\ P_D.$$

For ease of notation, let

$$\boldsymbol{z} = [z_1, \cdots, z_q]' := C_{n(11)}^{-1}\frac{X_{(1)}'\boldsymbol{\epsilon}}{\sqrt{n}}.$$

Then, for any $\frac{1}{2} < c' < 1$,

$$\begin{aligned} P\left(A_n^c\middle|\mathcal{F}_n\right) &\leq \sum_{j=1}^{q} P\left(|z_j| > \sqrt{n}\left[|\beta_{0,j}| + o(1)\right]\middle|\mathcal{F}_n\right) \\ &= \sum_{j=1}^{q} P\left(\frac{|z_j|}{n^{c'-\frac{1}{2}}} > n^{1-c'}\Big[|\beta_{0,j}| + o(1)\Big]\middle|\mathcal{F}_n\right) \\ &\to 0 \quad a.s.\ P_D. \end{aligned}$$

Next, using the same notations that we introduced in the proofs of Lemma A.7 and Theorem 3.1, let

$$H = X_{(1)}C_{n(11)}^{-1}C_{n(12)} - X_{(2)},$$

and let

$$\eta_* = \min_{1\leq j\leq p_n - q} \boldsymbol{\eta},$$

where assumption (3.1) ensures that $0 < \eta_* \leq 1$. Again, due to assumptions (2.2) and (2.3) and that q is fixed, every element in the matrix H is bounded. Let h_{ij} be the $(i,j)^{th}$ element of H. Again, by Lemma A.1, for all $j = 1, \cdots, p_n - q$,

$$\frac{1}{n^{c_1}}\sum_{i=1}^{n} h_{ji}\epsilon_i \to 0 \quad a.s.\ P_D$$

for $\frac{1}{2} < c_1 < 1$. Consequently, we have

$$\begin{aligned} P\left(B_n^c \middle| \mathcal{F}_n\right) &= P\left(\bigcup_{j=1}^{p_n-q}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_{ji}\epsilon_i\right| > \frac{\lambda_n}{2\sqrt{n}}\eta_j\right\}\middle| \mathcal{F}_n\right) \\ &\leq P\left(\left\|\frac{1}{\sqrt{n}}H'\boldsymbol{\epsilon}\right\|_2 > \frac{\lambda_n}{2\sqrt{n}}\eta_*\middle| \mathcal{F}_n\right) \\ &= P\left(\frac{1}{n^{c_2-\frac{1}{2}}}\left\|\frac{1}{\sqrt{n}}H'\boldsymbol{\epsilon}\right\|_2 > \frac{\lambda_n}{2n^{c_2}}\eta_*\middle| \mathcal{F}_n\right), \end{aligned}$$

where

$$\begin{aligned} \left(\frac{1}{n^{c_2-\frac{1}{2}}}\left\|\frac{1}{\sqrt{n}}H'\boldsymbol{\epsilon}\right\|_2\right)^2 &= \frac{n^{2c_1-1}}{n^{2c_2-1}}\sum_{j=1}^{p_n-q}\left(\frac{1}{n^{c_1}}\sum_{i=1}^{n} h_{ji}\epsilon_i\right)^2 \\ &= \mathcal{O}\left(\frac{1}{n^{2(c_2-c_1)}}\right) \times o\left(n^{c_3}\right) \ \ a.s.\ P_D \\ &= o(1) \ \ a.s.\ P_D \end{aligned}$$

because $c_3 < 2(c_2 - c_1)$ and $\frac{1}{2} < c_1 < c_2 < 1$, whereas

$$\frac{\lambda_n}{2n^{c_2}}\eta_* = \mathcal{O}(1).$$

Hence $P\left(B_n^c \middle| \mathcal{F}_n\right) \to 0$ almost surely under P_D and the result follows. □

Note that the constraints on c_1, c_2 and c_3 in Lemma A.10 cover the more restrictive constraints found in Theorem 3.1. Therefore, the result in Lemma A.10 still holds under the assumptions of Theorem 3.1.

The following version of Sherman–Morrison–Woodbury matrix-inversion identity (e.g., Equation (26) of Henderson and Searle (1981)) will come in handy later: For any square matrices A and B of conformal sizes where A is invertible, we have

$$(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1}\left(I + BA^{-1}\right)^{-1}. \tag{A.10}$$

Proof of Theorem 3.4. Since the first-step is in fact equivalent to the one-step procedure, Theorem 3.1 immediately gives us

$$P\left(\widehat{S}_n^w = S_0 \middle| \mathcal{F}_n\right) \geq P\left(\widehat{\boldsymbol{\beta}}_n^w \overset{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D,$$

while Lemma A.10 immediately gives us

$$P\left(\widehat{S}_n = S_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D.$$

Conditional on $\left\{\widehat{S}_n^w = S_0\right\}$ and $\left\{\widehat{S}_n = S_0\right\}$, since $Y = X_{(1)}\boldsymbol{\beta}_{0(1)} + \boldsymbol{\epsilon}$,

$$\widehat{\boldsymbol{\beta}}_{n(1)}^w - \widehat{\boldsymbol{\beta}}_{n(1)}^{LAS+LS}$$

$$
\begin{aligned}
&= \left(X'_{(1)}D_nX_{(1)}\right)^{-1}X'_{(1)}D_nY - \left(X'_{(1)}X_{(1)}\right)^{-1}X'_{(1)}Y \\
&= \left(X'_{(1)}D_nX_{(1)}\right)^{-1}X'_{(1)}D_n\boldsymbol{\epsilon} - \left(X'_{(1)}X_{(1)}\right)^{-1}X'_{(1)}\boldsymbol{\epsilon} \\
&= \left(C^w_{n(11)}\right)^{-1}\frac{X'_{(1)}(D_n-I_n)\boldsymbol{\epsilon}}{n} - \left[C^{-1}_{n(11)} - \left(C^w_{n(11)}\right)^{-1}\right]\frac{X'_{(1)}\boldsymbol{\epsilon}}{n},
\end{aligned}
$$

which leads to

$$
\begin{aligned}
&\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)} - \widehat{\boldsymbol{\beta}}^{LAS+LS}_{n(1)}\right) \\
&= \left(C^w_{n(11)}\right)^{-1}\frac{X'_{(1)}(D_n-I_n)\boldsymbol{\epsilon}}{\sqrt{n}} - \left[C^{-1}_{n(11)} - \left(C^w_{n(11)}\right)^{-1}\right]\frac{X'_{(1)}\boldsymbol{\epsilon}}{\sqrt{n}}.
\end{aligned}
$$

Based on the (alternative) proof of Lemma A.2, we have seen that

$$
\left(C^w_{n(11)}\right)^{-1} \xrightarrow{\text{a.s.}} C^{-1}_{11},
$$

and from the (alternative) proof of Lemma A.6, we could deploy Slutsky's Theorem to obtain

$$
\left(C^w_{n(11)}\right)^{-1}\frac{X'_{(1)}(D_n-I_n)\boldsymbol{\epsilon}}{\sqrt{n}} \xrightarrow{\text{c.d.}} N_q\left(\mathbf{0}\ ,\ \sigma^2_W\sigma^2_\epsilon C^{-1}_{11}\right) \quad a.s.\ P_D.
$$

Meanwhile, we deploy the matrix inversion identity (A.10) by taking $A = C_{n(11)}$ and

$$
B = \frac{1}{n}X'_{(1)}(D_n-I_n)X_{(1)}
$$

to obtain

$$
\begin{aligned}
\left(C^w_{n(11)}\right)^{-1} &= \left[C_{n(11)} + \frac{1}{n}X'_{(1)}(D_n-I_n)X_{(1)}\right]^{-1} \\
&= A^{-1} - A^{-1}BA^{-1}\left(I_q + BA^{-1}\right)^{-1}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\left[C^{-1}_{n(11)} - \left(C^w_{n(11)}\right)^{-1}\right]\frac{X'_{(1)}\boldsymbol{\epsilon}}{\sqrt{n}} \\
&= C^{-1}_{n(11)}\left[\frac{X'_{(1)}(D_n-I_n)X_{(1)}}{n}\right]C^{-1}_{n(11)}\left[I_q + \left(\frac{X'_{(1)}(D_n-I_n)X_{(1)}}{n}\right)C^{-1}_{n(11)}\right]^{-1}\frac{X'_{(1)}\boldsymbol{\epsilon}}{\sqrt{n}} \\
&= C^{-1}_{n(11)}\left[\frac{X'_{(1)}(D_n-I_n)X_{(1)}}{n^{1-c}}\right]C^{-1}_{n(11)}\left[I_q + \left(\frac{X'_{(1)}(D_n-I_n)X_{(1)}}{n}\right)C^{-1}_{n(11)}\right]^{-1}\frac{X'_{(1)}\boldsymbol{\epsilon}}{n^{\frac{1}{2}+c}},
\end{aligned}
$$

where Lemma A.1 and assumption (2.2) ensure that for any $0 < c < \frac{1}{2}$,

$$
\frac{1}{n^{1-c}}X'_{(1)}(D_n-I_n)X_{(1)} \xrightarrow{\text{a.s.}} \mathbf{0}
$$

and

$$\frac{X'_{(1)}\boldsymbol{\epsilon}}{n^{\frac{1}{2}+c}} \to \mathbf{0} \quad a.s.\ P_D.$$

Since $C_{n(11)}$ is invertible for all n, we have

$$C^{-1}_{n(11)} \to C^{-1}_{11},$$

and

$$\begin{aligned}\left[I_q + \left(\frac{X'_{(1)}(D_n - I_n)X_{(1)}}{n}\right) C^{-1}_{n(11)}\right]^{-1} &= C_{n(11)} \left(C^w_{n(11)}\right)^{-1} \\ &\xrightarrow{\text{a.s.}} C_{11}C^{-1}_{11} \\ &= I_q.\end{aligned}$$

Hence,

$$\left[C^{-1}_{n(11)} - \left(C^w_{n(11)}\right)^{-1}\right] \frac{X'_{(1)}\boldsymbol{\epsilon}}{\sqrt{n}} \xrightarrow{\text{c.p.}} \mathbf{0} \quad a.s.\ P_D.$$

Consequently, conditional on $\left\{\widehat{S}^w_n = S_0\right\}$ and $\left\{\widehat{S}_n = S_0\right\}$, Slutsky's Theorem ensures that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)} - \widehat{\boldsymbol{\beta}}^{LAS+LS}_{n(1)}\right) \xrightarrow{\text{c.d.}} N_q\left(\mathbf{0}\ ,\ \sigma^2_W\sigma^2_\epsilon C^{-1}_{11}\right) \quad a.s.\ P_D.$$

Finally, for any $t \in \mathbb{R}$,

$$\begin{aligned}&P\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)} - \widehat{\boldsymbol{\beta}}^{LAS+LS}_{n(1)}\right) \le t \middle| \mathcal{F}_n\right) \\ \le &P\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)} - \widehat{\boldsymbol{\beta}}^{LAS+LS}_{n(1)}\right) \le t\,,\ \left\{\widehat{S}^w_n = S_0, \widehat{S}_n = S_0\right\} \middle| \mathcal{F}_n\right) \\ &+ P\left(\widehat{S}^w_n \ne S_0 \middle| \mathcal{F}_n\right) + P\left(\widehat{S}_n \ne S_0 \middle| \mathcal{F}_n\right)\end{aligned}$$

where

$$P\left(\widehat{S}^w_n \ne S_0 \middle| \mathcal{F}_n\right) \to 0\ \ a.s.\ P_D \quad \text{and} \quad P\left(\widehat{S}_n \ne S_0 \middle| \mathcal{F}_n\right) \to 0\ \ a.s.\ P_D,$$

and

$$P\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_{n(1)} - \widehat{\boldsymbol{\beta}}^{LAS+LS}_{n(1)}\right) \le t\,,\ \left\{\widehat{S}^w_n = S_0, \widehat{S}_n = S_0\right\} \middle| \mathcal{F}_n\right) \to P(Z \le t)$$

almost surely under P_D for $Z \sim N_q\left(\mathbf{0}\ ,\ \sigma^2_W\sigma^2_\epsilon C^{-1}_{11}\right)$. □

Proof of Theorem 3.5. Since $Y = X_{(1)}\boldsymbol{\beta}_{0(1)} + \boldsymbol{\epsilon}$, by conditioning on $\left\{\widehat{S}^w_n = S_0\right\}$, we have $\widehat{\boldsymbol{\beta}}^w_{n(2)} = \boldsymbol{\beta}_{0(2)} = \mathbf{0}$, and

$$\widehat{\boldsymbol{\beta}}^w_{n(1)} - \boldsymbol{\beta}_{0(1)} = \left(X'_{(1)}D_nX_{(1)}\right)^{-1} X'_{(1)}D_nY - \boldsymbol{\beta}_{0(1)}$$

$$
\begin{aligned}
&= \left(X'_{(1)} D_n X_{(1)}\right)^{-1} X'_{(1)} D_n \boldsymbol{\epsilon} \\
&= \left(C^w_{n(11)}\right)^{-1} \frac{X'_{(1)} D_n \boldsymbol{\epsilon}}{n} \\
&\xrightarrow{\text{c.p.}} \mathbf{0} \quad a.s.\ P_D
\end{aligned}
$$

by Lemmas A.4 and A.6. Finally, for any $\xi > 0$,

$$
\begin{aligned}
&P\left(\left\|\widehat{\boldsymbol{\beta}}^w_n - \boldsymbol{\beta}_0\right\|_2 > \xi \middle| \mathcal{F}_n\right) \\
&\leq P\left(\left\|\widehat{\boldsymbol{\beta}}^w_n - \boldsymbol{\beta}_0\right\|_2 > \xi\,,\ \widehat{S}^w_n = S_0 \middle| \mathcal{F}_n\right) + P\left(\widehat{S}^w_n \neq S_0 | \mathcal{F}_n\right) \\
&\to 0 \quad a.s.\ P_D.
\end{aligned}
$$

□

Remark A.1. *Consider Theorem 3.3 with centering on* $\boldsymbol{\beta}_0$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_n - \boldsymbol{\beta}_0\right).$$

Using the same technique in the proof of Theorem 3.3, we work with

$$V_n(\boldsymbol{u}) := Q_n\left(\boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) - Q_n\left(\boldsymbol{\beta}_0\right)$$

which can be simplified into

$$\boldsymbol{u}'\left(\frac{X' D_n X}{n}\right)\boldsymbol{u} - 2\boldsymbol{u}'\left(\frac{X' D_n \boldsymbol{\epsilon}}{\sqrt{n}}\right) + \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^{p} W_{0,j}\left(\left|\sqrt{n}\beta_{0,j} + u_j\right| - \left|\sqrt{n}\beta_{0,j}\right|\right).$$

Again, assumption 2.4 ensures convergence of the first term, whereas argument for the penalty term in the proof of Theorem 3.3 still applies to the third term. However, the second term has

$$\frac{X' D_n \boldsymbol{\epsilon}}{\sqrt{n}} = \frac{1}{\sqrt{n}} X'\left(D_n - \mu_W I_n\right)\boldsymbol{\epsilon} + \frac{1}{\sqrt{n}} X'\boldsymbol{\epsilon},$$

where

$$\frac{1}{\sqrt{n}} X'\left(D_n - \mu_W I_n\right)\boldsymbol{\epsilon} = \mathcal{O}_p(1) \quad a.s.\ P_D,$$

but $(X'\boldsymbol{\epsilon})/(\sqrt{n})$ *is asymptotically normal under* P_D *(Knight and Fu, 2000). Thus, conditional on* $\mathcal{F}_n$, $(X' D_n \boldsymbol{\epsilon})/(\sqrt{n})$ *depends on the sample path of realized data* $\{y_1, y_2, \cdots\}$, *thus causing* $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^w_n - \boldsymbol{\beta}_0\right)$ *to be unable to achieve convergence in conditional distribution almost surely under* P_D.

Lemma A.11 (Rate of Convergence). *Adopt all assumptions in Theorem 3.4. If there exists* $0 < c_4 < \frac{1}{2}$ *such that*

$$0 \leq c_3 < \min\{2(c_2 - c_1)\,,\ 2c_1 - 1\} - c_4$$

under weighting schemes (1.4) and (1.5), or

$$0 \leq c_3 < \min\left\{\frac{2}{3}(c_2 - c_1) - \frac{c_4}{3}\,,\, 2c_1 - 1 - c_4\right\}$$

under weighting schemes (1.6), then

$$P\left(\widehat{S}_n^w \neq S_0 \middle| \mathcal{F}_n\right) = o\left(n^{-c_4}\right) \quad a.s.\ P_D.$$

Proof. The result is immediate by extracting the additional n^{-c_4} factor from the proofs of Lemmas A.3 and A.7 as well as Theorem 3.1. In particular, from the proofs of Lemma A.6 and Theorem 3.1, it is clear that the rate of convergence of $P[(A_n^w)^c|\mathcal{F}_n]$ is faster than that of $P[(B_n^w)^c|\mathcal{F}_n]$, whereas the conditions in Lemma A.11 ensure that $P[(B_n^w)^c|\mathcal{F}_n] = o\left(n^{-c_4}\right)$ a.s. P_D. Finally,

$$P\left(\widehat{S}_n^w \neq S_0 \middle| \mathcal{F}_n\right) \leq P[(A_n^w)^c|\mathcal{F}_n] + P[(B_n^w)^c|\mathcal{F}_n] = o\left(n^{-c_4}\right) \quad a.s.\ P_D. \qquad \square$$

Acknowledgments

The authors thank the associate editor and an anonymous referee for their valuable feedback and suggestions that lead to a substantially improved manuscript. Insights from Nick Polson and Steve Wright have also served as helpful guideposts in this effort.

References

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

Bissiri, P. G., Holmes, C. C. and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **78** 1103–1130. MR3557191

Camponovo, L. (2015). On the validity of the pairs bootstrap for lasso estimators. *Biometrika* **102** 981–987. MR3431568

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43** 1986–2018. MR3375874

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *The Annals of Statistics* **33** 414–436. MR2157808

Chatterjee, A. and Lahiri, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proceedings of the American Mathematical Society* **138** 4497–4509. MR2680074

Chatterjee, A. and Lahiri, S. N. (2011a). Strong consistency of lasso estimators. *Sankhya: The Indian Journal of Statistics, Series A* **73** 55–78. MR2887087

Chatterjee, A. and Lahiri, S. N. (2011b). Bootstrapping Lasso estimators. *Journal of the American Statistical Association* **106** 608–625. MR2847974

Das, D., Gregory, K. and Lahiri, S. N. (2019). Perturbation bootstrap in Adaptive Lasso. *The Annals of Statistics* **47** 2080–2116. MR3953445

Das, D. and Lahiri, S. N. (2019). Distributional consistency of the lasso by perturbation bootstrap. *Biometrika* **106** 957–964. MR4031208

Durrett, R. (2010). *Probability: Theory and Examples (Cambridge Series in Statistical and Probabilistic Mathematics)*, 4th ed. Cambridge: Cambridge University Press, New York, USA. MR2722836

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

Fong, E., Lyddon, S. and Holmes, C. C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

Geyer, C. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

Gramacy, R. B., Moler, C. and Turlach, B. A. (2019). monomvn: Estimation for MVN and Student-t Data with Monotone Missingness R package version 1.9-13.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

Henderson, H. V. and Searle, S. R. (1981). On Deriving the Inverse of a Sum of Matrices. *SIAM Review* **23** 53–60. MR0605440

Jin, Z., Ying, Z. and Wei, L.-J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88** 381–390. MR1844838

Johnson, V. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107** 649–660. MR2980074

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28** 1356–1378. MR1805787

Lai, T. L., Robbins, H. and Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. *Proceedings of National Academy of Sciences* **75** 3034–3036. MR0518953

Liu, H. and Yu, B. (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* **7** 3124–3169. MR3151764

Lyddon, S. P., Holmes, C. C. and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106** 465–478. MR3949315

Lyddon, S., Walker, S. and Holmes, C. (2018). Nonparametric Learning from Bayesian Models with Randomized Objective Functions. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems. NIPS'18* 2075–2085. Curran Associates Inc.

Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the

consistency of a general bootstrap. *The Annals of Statistics* **20** 1611–1624. MR1186268

Minnier, J., Tian, L. and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* **106** 1371–1382. MR2896842

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42** 789–817. MR3210987

Newton, M., Polson, N. G. and Xu, J. (2021). Weighted Bayesian Bootstrap for Scalable Posterior Distributions. *The Canadian Journal of Statistics*.

Ng, T. L. (2022). Random Weighting in LASSO Regression and in Discrete Mixture Models, PhD thesis, University of Wisconsin-Madison.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the LASSO and Its Dual. *Journal of Computational and Graphical Statistics* **9** 319–337. MR1822089

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103** 681–686. MR2524001

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199. MR1128411

Pompe, E. (2021). Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification. *arXiv: 2103.14445*.

Shao, J. (2003). *Mathematical Statistics (Springer Texts in Statistics)*, 2nd ed. Springer, New York, USA. MR2002723

R Core Team (2019). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **58** 267–288. MR1379242

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* **7** 1456–1490. MR3066375

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth ed. Springer, New York. ISBN 0-387-95457-0. MR1337030

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563. MR2274449

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469