Robust Inference for High-Dimensional Linear Models via Residual Randomization

Y. Samuel Wang [*1] **Si Kai Lee** [*1] **Panos Toulis** [1] **Mladen Kolar** [1]

Abstract

We propose a residual randomization procedure designed for robust Lasso-based inference in the high-dimensional setting. Compared to earlier work that focuses on sub-Gaussian errors, the proposed procedure is designed to work robustly in settings that also include heavy-tailed covariates and errors. Moreover, our procedure can be valid under clustered errors, which is important in practice, but has been largely overlooked by earlier work. Through extensive simulations, we illustrate our method's wider range of applicability as suggested by theory. In particular, we show that our method outperforms state-of-art methods in challenging, yet more realistic, settings where the distribution of covariates is heavy-tailed or the sample size is small, while it remains competitive in standard, "well behaved" settings previously studied in the literature.

1. Introduction

The Lasso (Tibshirani, 1996) and its variants are typically used to estimate the coefficients of a linear model in the high-dimensional setting where the number of covariates, p, is larger than the number of samples, n. The Lasso has been shown to possess many desirable theoretical properties and has proven fruitful in applications across nearly all scientific domains (Bühlmann & Van De Geer, 2011). This widespread use has recently generated much interest in procedures for performing inference using Lasso estimates.

However, for parameters which are zero or nearly zero, the Lasso point estimates may have an irregular distribution, and naïvely constructing confidence intervals typically results in invalid inference. To overcome these difficulties, various procedures have been proposed. Wasserman & Roeder (2009) and Meinshausen et al. (2009) both used sample splitting procedures to form valid p-values. Zhang & Zhang (2014), Van de Geer et al. (2014), and Javanmard & Montanari (2014) proposed the desparsified and debiased Lasso, which adds a one step correction to the Lasso estimate. The resulting estimate is not sparse, but under certain conditions, it has asymptotically negligible bias. The debiased/desparsified estimate is asymptotically normal and can be used for inference.

An alternative approach for inference in high-dimensional linear models is the bootstrap. Chatterjee & Lahiri (2011) proposed a residual bootstrap procedure for the adaptive Lasso (Zou, 2006). One crucial requirement with this approach is the "beta-min" condition, which requires the non-zero parameters to be large enough in absolute value (i.e., much larger than $n^{-1/2}$). This condition can be overly restrictive in cases where the primary aim is to test null hypotheses on the significance of regression coefficients of the form $H_0 : \beta_j = 0$. To circumvent this problem, Dezeure et al. (2017) and Zhang & Cheng (2017) both proposed bootstrap procedures based on the desparsified Lasso (Van de Geer et al., 2014), which are capable of performing simultaneous inference over a set of parameters in the linear model; i.e., $H_0 : \beta_j = 0$ for all $j \in J \subseteq [p] = \{1, \dots, p\}$. Zhang & Cheng (2017) proposed bootstrapping the linearized part of the desparsified Lasso with a Gaussian multiplier bootstrap, while Dezeure et al. (2017) proposed using either a wild bootstrap or a residual bootstrap procedure for the entire estimator.

The above procedures are typically obtained under strong regularity conditions on covariates and require i.i.d. sub-Gaussian errors. Therefore, they may perform poorly in more realistic settings where n is relatively small, the covariates and errors are non-Gaussian and/or heavy-tailed, or have complex structures, such as heterogeneity or clustering.

In contrast, randomization methods (Fisher et al., 1935; Pitman, 1937; Kempthorne, 1952) are non-parametric and typically exact in finite samples, which makes them robust (Lehmann & Romano, 2006, Chapter 15). Randomization procedures leverage structure in the data for testing and inference —e.g., permutation tests exploit exchangeability through permuting the data—and rely less on analytical assumptions or asymptotic arguments.

[*]Equal contribution [1]Booth School of Business, University of Chicago, Chicago, USA. Correspondence to: Y. Samuel Wang <swang24@uchicago.edu>, Si Kai Lee <sikai@uchicago.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

We show that a robust alternative for inference in high-dimensional linear models is possible through the use of *residual randomization*, which was first proposed by Freedman & Lane (1983a;b) as an extension of Fisher's randomization test. Residual randomization builds a test that is exact in the idealized case where the true errors are known, and remains asymptotically valid when using regression residuals in lieu of the true errors. Recently, Toulis (2019) extended the original residual randomization procedure to include complex error structures (e.g., clustered errors), and showed that the procedure is asymptotically valid and empirically effective in the low-dimensional setting where p is fixed and n grows, including settings with heterogeneity and clustering. Toulis (2019) also proposed an extension to the high-dimensional setting (different than the one we propose), but did not give any theoretical guarantees.

1.1. Contributions

In this paper, we propose a novel residual randomization procedure for conducting hypothesis tests and computing confidence intervals for parameters in a high-dimensional linear model. In Section 2.3, we define a class of oracle tests that are exact for finite samples. However, these oracle tests are infeasible, and so we develop an approximate implementation for each test in the class. Broadly speaking, our procedure selects a specific test from the class for which the approximate but feasible test is close to the oracle. Thus, we explicitly prioritize controlling the empirical size of the test. Our procedure is specifically designed to be robust to small sample sizes and non-Gaussianity in both covariates and errors. This is in contrast to previous procedures that prioritize testing power and shorter confidence intervals.

We show theoretically that our procedure is valid even when the covariates and errors are sub-Weibull as opposed to the sub-Gaussian condition previously required in the literature. We also show that our procedure is sound when the errors have clustered dependence. Indeed, we see empirically that our residual randomization method is comparable to state-of-art methods in "well behaved" high-dimensional benchmarks and is superior in more complex settings, e.g., when n is small, the covariates are non-Gaussian, or the errors are heavy-tailed or heterogeneous.

2. Methodology

2.1. Setup

Throughout we let $|\cdot|_q$ and $\|\cdot\|_q$ denote the vector q-norm and matrix q-norm, respectively. For any positive integer d, we let $[d] = \{1, \dots, d\}$. For $X \in \mathbb{R}^{n\times p}$, let $X_{i,:}$ denote the ith row and $X_{:,v}$ denote the vth column. We let e_j denote the jth standard basis; i.e., a vector whose jth element is 1 and all other elements are 0.

We assume that the data $Y \in \mathbb{R}^n$ are generated from the linear model

$$Y = X\beta + \varepsilon, \tag{1}$$

where $\beta \in \mathbb{R}^p$ are the linear coefficients, $X \in \mathbb{R}^{n\times p}$ are the covariates, and $\varepsilon \in \mathbb{R}^n$ is the vector of (unobservable) errors. Thus, $X_{i,:}$ corresponds to the covariates for the ith observation and $X_{:,v}$ corresponds to the vth covariate. Let s denote the sparsity of β such that $|\beta|_0 \le s$.

In contrast to previous work which requires sub-Gaussian covariates and errors, we allow $X_{i,:}$ and ε_i to follow sub-Weibull(α) distributions. Sub-Weibull random variables are a class of distributions with tails of the form $\exp(-|x|^\alpha)$ (Kuchibhotla & Chakrabortty, 2018; Vladimirova et al., 2020). The class of sub-Gaussian and sub-exponential distributions can be obtained as a subclass by setting α equal to 2 and 1, respectively. However, when $\alpha < 1$, the tails can be heavier than sub-exponential. Finally, we also assume that $X_{i,:}$ are drawn i.i.d., and $X_{i,:}$ and ε_i are uncorrelated so that $\mathbb{E}(X_{i,:}\varepsilon_i) = 0$.

We propose a procedure to test linear hypotheses of the form

$$H_0 : a^\top \beta = a_0, \tag{2}$$

for some $a \in \mathbb{R}^p$ and $a_0 \in \mathbb{R}$; we then invert the test to form confidence intervals. The form in (2) includes many hypotheses of interest; e.g., setting $a = e_j$ and $a_0 = 0$ implies $H_0 : \beta_j = 0$, setting $a = e_j - e_k$ and $a_0 = 0$ implies $H_0 : \beta_j = \beta_k$. In Remark 2 we briefly discuss how the procedure can be generalized to tests of the form $H_0 : A^\top\beta = a_0$ where $A \in \mathbb{R}^{p\times d}$ and $a_0 \in \mathbb{R}^d$. This would allow for simultaneous inference as in Zhang & Cheng (2017) and Dezeure et al. (2017).

2.2. Residual Randomization

Following the framework of Toulis (2019), we require two key constructs: (1) A set $\mathcal{G}$ of linear maps $G : \mathbb{R}^n \mapsto \mathbb{R}^n$ such that $G\varepsilon \overset{d}{=} \varepsilon$, conditional on X; and (2) An invariant $t_n : \mathbb{R}^n \to \mathbb{R}$, where $t_n(\varepsilon) \overset{d}{=} t_n(G\varepsilon)$ for all $G \in \mathcal{G}$ and any finite n. We emphasize that all the test we propose is conditional on X.

Given these two definitions, if we can find a test statistic, $T_n(Y, X)$, such that $T_n(Y, X) = t_n(\varepsilon)$ under H_0, then we can compare the observed value of T_n with $\{t_n(G\varepsilon) : G \in \mathcal{G}\}$ to test H_0. Indeed, conditioned on X and ε (or alternatively X and Y), the random variable

$$\pi = \sum_{G\in\mathcal{G}} \mathbb{1}\{t_n(G\varepsilon) \ge T_n\}/|\mathcal{G}|, \tag{3}$$

is uniform over the set $\{0, \dots, |\mathcal{G}|\}/|\mathcal{G}|$. Thus, π can be used as a p-value and rejecting the null hypothesis when

$\pi \leq \pi_0$ yields a hypothesis test with exact size π_0[1] (save for the discreteness in π which can be easily remedied by adding uniform noise to π or increasing $|\mathcal{G}|$). We consider the following three invariances for the distribution of ε.

Exchangeability: If all elements of ε are exchangeable, then $\mathcal{G}$ could be all $n \times n$ permutation matrices. This includes the standard setting with i.i.d. errors.

Sign Symmetry: If $\varepsilon_i \overset{d}{=} -\varepsilon_i$ for all $i \in [n]$, then $\mathcal{G}$ could be all $n \times n$ diagonal matrices with ± 1 on the diagonal. This allows for heteroskedastic errors where ε_i may depend on $X_{i,:}$, but is symmetric around 0 conditional on $X_{i,:}$.

Clustered Exchangeability: In many cases, the data can be partitioned into disjoint clusters, such that exchangeability is only reasonable within a cluster; e.g., to model country-specific effects in users of an online platform; $\mathcal{G}$ should then be the set of within-cluster permutations. This generalizes exchangeability, but we will keep the two settings distinct.

In the low-dimensional setting, where $n \gg p$, Toulis (2019) considered tests of the form $a^\top \beta = a_0$ and used the test statistic $T_n = \sqrt{n}(a^\top \hat{\beta} - a_0)$, where $\hat{\beta}$ is the least squares estimate of β, and set the invariant $t_n(u) = \sqrt{n} a^\top (X^\top X)^{-1} X^\top u$. Since $\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \varepsilon$, under the null hypothesis that $a^\top \beta = a_0$,

$$T_n = \sqrt{n}(a^\top \hat{\beta} - a_0) = \sqrt{n} a^\top (\hat{\beta} - \beta) = t_n(\varepsilon). \tag{4}$$

Thus, given the true errors ε, we could form an exact p-value as in Eq. (3). In practice, ε is unknown so one would instead use the residuals, $\hat{\varepsilon} = Y - X\hat{\beta}$, or the restricted residuals, $y - X\hat{\beta}_{\mathrm{r}}$, where $\hat{\beta}_{\mathrm{r}}$ is the restricted OLS estimates under H_0. In this case, the test is approximate, but as shown in Toulis (2019), it can attain the correct size asymptotically.

Remark 1 *The robustness properties of residual randomization can be deduced from (3). Specifically, the performance of the test does not depend on the distribution of the errors. Similarly, regularity conditions on X are not needed because the test is conditioned on X. Finally, the decision of the test remains invariant to monotone transformations of t_n, and so the test remains robust to rescaling of the data. In our experiments, we demonstrate that the robustness properties of the exact test (using the true errors) are also inherited by the approximate procedure using the residuals.*

2.3. High-Dimensional Residual Randomization

In the high-dimensional setting, where $p > n$, the OLS estimate $\hat{\beta}$ is ill-defined and the low-dimensional residual randomization method cannot be directly applied. In this section, we propose adjustments appropriate for the high-dimensional setting and show that this test can be further optimized for robustness.

[1]We describe a one sided test, but a two-sided test could be similarly defined.

Instead of the OLS estimator, we use a version of the debiased Lasso (Javanmard & Montanari, 2014) which corrects for the regularization bias by adding a term proportional to the subgradient of the objective at the Lasso solution $\hat{\beta}^l(\lambda_1)$, where λ_1 is the penalty parameter. When it is obvious, we will simply write $\hat{\beta}^l$ instead of $\hat{\beta}^l(\lambda_1)$. Appropriate values of λ_1 are problem dependent, and we give theoretically sufficient choices in Section 3.

Specifically, for some $M \in \mathbb{R}^{p \times p}$, we use the estimator

$$\hat{\beta}^{d,M} = \hat{\beta}^l + \frac{1}{n} M X^\top (Y - X\hat{\beta}^l). \tag{5}$$

To form the high-dimensional test statistic, we replace $\hat{\beta}$ in (4) with $\hat{\beta}^{d,M}$:

$$T_n^{(M)} = \sqrt{n}(a^\top \hat{\beta}^{d,M} - a_0). \tag{6}$$

Setting $S = \frac{1}{n} X^\top X$, this yields

$$\begin{aligned} T_n^{(M)} = \sqrt{n} a^\top (I - MS)(\hat{\beta}^l - \beta) + \frac{1}{\sqrt{n}} a^\top M X^\top \varepsilon \\ + \sqrt{n}(a^\top \beta - a_0). \end{aligned} \tag{7}$$

For some fixed $\hat{\beta}^l$ and any $M \in \mathbb{R}^{p \times p}$, comparing $T_n^{(M)}$ to

$$t^{(M)}(G\varepsilon) = \sqrt{n} a^\top (I - MS)(\hat{\beta}^l - \beta) + \frac{1}{\sqrt{n}} a^\top M X^\top G\varepsilon \tag{8}$$

for all $G \in \mathcal{G}$ would yield an exact test as described in (3) under the null, since $a^\top \beta - a_0 = 0$. When the null hypothesis does not hold such that $a^\top \beta = a_1 \neq a_0$, under weak conditions, $\max_G \frac{1}{\sqrt{n}} a^\top M X^\top G\varepsilon$ will be bounded at a rate of $\log(pn)$. However, $T_n^{(M)}$ contains an additional $\sqrt{n}(a^\top \beta - a_0) = \sqrt{n}(a_1 - a_0)$ term which will grow as $O(\sqrt{n})$ leading to rejection of the null hypothesis.

This procedure could be applied for any M.[2] Thus, in contrast to the low-dimensional setting, in the high-dimensional setting, we have not just a single test, but a class of tests indexed by M, all of which are exact under the null hypothesis. For any M, we call this the *oracle randomization distribution*. Since all oracle tests are exact, a good rule of thumb would be to select the matrix M, which gives the test with the most power, or alternatively, which—when inverted—yields the shortest confidence intervals. In some sense, this is the motivation behind Zhang & Cheng (2017) and Dezeure et al. (2017) setting M as an estimate of Σ^{-1}, albeit for a bootstrap procedure which is not exact. As suggested by the Gauss-Markov theorem, this should asymptotically give the shortest confidence intervals.

[2]If $a^\top M X^\top = 0$, then the distribution over G would be a point mass. Nonetheless some randomization procedure for breaking ties could be used to maintain exact size. In (13), letting $\lambda < 1$ will prevent such an M from being feasible.

However, since we do not have access to $\hat{\beta}^l - \beta$ or ε, we cannot directly use any of these oracle tests. Thus, we use the following invariant with $\hat{\varepsilon} = Y - X\hat{\beta}^l = \varepsilon + X(\beta - \hat{\beta}^l)$ being the residuals from the Lasso regression:

$$\hat{t}^{(M)}(G\hat{\varepsilon}) = \frac{a^\top M X^\top G\hat{\varepsilon}}{\sqrt{n}} = \frac{a^\top M X^\top G(\varepsilon + X(\beta - \hat{\beta}^l))}{\sqrt{n}} = \frac{1}{\sqrt{n}} a^\top M X^\top G X(\beta - \hat{\beta}^l) + \frac{1}{\sqrt{n}} a^\top M X^\top G\varepsilon. \tag{9}$$

For any M, we refer to the resulting distribution—when selecting G uniformly from $\mathcal{G}$—as the *attainable randomization distribution*, because it only involves quantities which we can directly access and compute. The oracle (8) and attainable (9) distributions share the same second term, but differ in their first terms $(1/\sqrt{n})a^\top M X^\top G X(\beta - \hat{\beta}^l)$ and $\sqrt{n}a^\top(I - MS)(\hat{\beta}^l - \beta)$. Thus, each attainable distribution no longer retains the exact size that its corresponding oracle test enjoys. In particular, selecting an attainable test based on minimizing its oracle's confidence interval length may result in poor finite sample performance.

Instead of prioritizing short confidence intervals, we prioritize the correct size of the test by selecting M to minimize the distance between the attainable test and its corresponding oracle test. Indeed, in Section 3 we show that the Wasserstein-1 distance between the oracle and attainable distributions is upper bounded by

$$\sqrt{n}|\hat{\beta}^l - \beta|_1 \quad |a^\top(I - MS)|_\infty + \frac{\sum_G \left| a^\top M X^\top G X/n \right|_\infty}{|\mathcal{G}|} \Bigg) \Bigg(\tag{10}$$

Roughly speaking, $|a^\top(I - MS)|_\infty$ regulates the difference between the means of the attainable and oracle distributions, while $\sum_G |a^\top M X^\top G X/n|_\infty / |\mathcal{G}|$ is the cost of using residuals instead of the true errors. Intuitively, one would expect (and we confirm empirically) that prioritizing the minimization of $|a^\top(I - MS)|_\infty$ over $\sum_G |a^\top M X^\top G X/n|_\infty / |\mathcal{G}|$ has a larger effect on the accuracy of the attainable test's p-value. Towards this end, we upweight the first term by setting $\delta \geq 1$ and select M which minimizes:

$$\delta |a^\top(I - MS)|_\infty + \frac{\sum_G \left| a^\top M X^\top G X/n \right|_\infty}{|\mathcal{G}|}. \tag{11}$$

While (11) can be solved using a linear program solver, for computational convenience, instead of directly optimizing (11) with respect to M, we instead solve

$$\min_{\lambda \in [0,1)} \delta |a^\top(I - M_\lambda S)|_\infty + \frac{\sum_G \left| a^\top M_\lambda \, {}_1 \, X^\top G X/n \right|_\infty}{|\mathcal{G}|}, \tag{12}$$

where

$$M_\lambda = \arg\min_M |a^\top M|_1 \quad \text{s.t.} \quad \left| a^\top(I - MS) \right|_\infty \leq \lambda. \tag{13}$$

Algorithm 1 Test $a^\top \beta = a_0$

Require: Y, X, $a^\top$, a_0, $\mathcal{G}$, λ_1, δ
Compute Lasso estimate $\hat{\beta}^l(\lambda_1)$ and $\hat{\varepsilon} = Y - X\hat{\beta}^l$
Compute $\lambda^\star$ and $M_{\lambda^\star}$ from (13) and (12)
Compute $\hat{\beta}^{d, M_{\lambda^\star}} = \hat{\beta}^l + \frac{1}{n} M_{\lambda^\star} X^\top (Y - X\hat{\beta}^l)$
Let $T_n^{(M_{\lambda^\star})} = \sqrt{n}(a^\top \hat{\beta}^{d, M_{\lambda^\star}} - a_0)$
for $G \in \mathcal{G}$ **do**
  Set $\hat{t}^{(M_{\lambda^\star})}(G\hat{\varepsilon}) = \frac{1}{\sqrt{n}} a^\top M_{\lambda^\star} X^\top G\hat{\varepsilon}$
end for
Return $\frac{\sum_G \mathbb{I}\{\hat{t}^{(M_{\lambda^\star})}(G\hat{\varepsilon}) > T_n^{(M_{\lambda^\star})}\}}{|\mathcal{G}|}$

The problem in (13) is the CLIME (Cai et al., 2011) problem, and we solve it using the `fastclime` package (Pang et al., 2014). In Section 3 we show that, for any $\delta \geq 1$, using $M_{\lambda^\star}$, where $\lambda^\star$ is a minimizer of (12), ensures that the selected attainable and oracle distributions converge.

To select M, Javanmard & Montanari (2014) solve a problem with the same constraint as (13), but instead minimize the $M_{i,:}^\top S M_{i,:}$ for all $i \in [p]$, which—similar to Zhang & Cheng (2017) and Dezeure et al. (2017)—prioritizes shorter confidence intervals. When it is sparse, the inverse covariance of $X_{i,:}$ can be consistently estimated by solving (13) (Cai et al., 2011). In that case the residual randomization procedure should still produce asymptotically efficient confidence intervals and would be asymptotically equivalent to the other procedures. However, in finite samples, we see empirical improvements in robustness. We detail our procedure in Algorithm 1, which produces a p-value for testing the null hypothesis that $a^\top \beta = a_0$.

Remark 2 *Thus far, we have assumed a 1-dimensional hypothesis test. However, similar to Zhang & Cheng (2017) and Dezeure et al. (2017), this can generalized be to testing several null-hypotheses simultaneously. In particular, one might instead use*

$$T_n = |\sqrt{n}(A^\top \hat{\beta}^d - a_0)|_\infty \tag{14}$$

and the corresponding invariant

$$t_n(G\hat{\varepsilon}) = \left| \frac{1}{\sqrt{n}} A^\top M X^\top G\hat{\varepsilon} \right|_\infty. \tag{15}$$

We focus on the 1*-dimensional case for expositional clarity.*

2.4. Confidence Intervals

To form a univariate confidence interval for β_j, we invert the hypothesis test for $\beta_j = a_0$ (Rosenbaum, 2003). In particular, for $a = e_j$ we can compute the distribution of $t(G\hat{\varepsilon})$ and set $\tau_{\pi_0/2}$ and $\tau_{1-\pi_0/2}$ to the $\pi_0/2$ and $1 - \pi_0/2$ quantiles. Finally, to invert the level π_0 two-sided test we

select a_0 such that $\sqrt{n}(\hat{\beta}^d - a_0) \in [\tau_{\pi_0/2}, \tau_{1-\pi_0/2}]$. This is equivalent to the confidence interval for β:

$$\left(\hat{\beta}^d - \frac{\tau_{1-\pi_0/2}}{\sqrt{n}}, \hat{\beta}^d - \frac{\tau_{\pi_0/2}}{\sqrt{n}}\right) \tag{16}$$

Note that as opposed to the multiplier bootstrap confidence intervals proposed by Zhang & Cheng (2017), the confidence intervals in (16) may be asymmetric because they are produced from the attainable randomization distribution, which may be asymmetric. This is similar to the procedure proposed by Dezeure et al. (2017).

3. Main Results

Let $F_t(X, \varepsilon)$ denote the oracle randomization distribution of t_n in (8) conditional on X and ε when G is selected uniformly from $\mathcal{G}$. Let $F_{\hat{t}}(X, \varepsilon)$ denote the attainable distribution; i.e., the distribution of $\hat{t}_n$ in (9) when G is chosen uniformly from $\mathcal{G}$. As the notation implies, both $F_t(X, \varepsilon)$ and $F_{\hat{t}}(X, \varepsilon)$ depend on X, ε and are random distributions with respect to X, ε. We give a finite sample characterization of the Wasserstein-1 distance between these two random distributions under certain assumptions, and we show that the distance goes to 0 with probability going to 1. All proofs are given in the supplement.

We first state Lemma 1 which, as mentioned in Section 2.3 shows that the distance between the oracle and attainable distributions can be decomposed into the estimation error of $\hat{\beta}^l$ as well as two terms which, roughly speaking, regulate the difference in means and variance.

Lemma 1 *For any $M \in \mathbb{R}^{p \times p}$, let $d_1\left(F_t(X, \varepsilon), F_{\hat{t}}(X, \varepsilon)\right)$ denote the Wasserstein-1 distance between the oracle randomization distribution and attainable randomization distributions. Then,*

$$d_1\left(F_t(X, \varepsilon), F_{\hat{t}}(X, \varepsilon)\right) \leq \left\|\hat{\beta}^l - \beta\right\|_1 \times \left[\left\|\sqrt{n} a^\top (I - MS)\right\|_\infty + \left\|a^\top M\right\|_1 \mathbb{E}_Q\left(\left\|X^\top G X / \sqrt{n}\right\|_\infty\right)\right]. \tag{17}$$

where Q is the uniform distribution over G in $\mathcal{G}$.

We now provide conditions under which the two terms in (17) can be controlled for the $M_{\lambda^\star}$ selected by (13) and (12). Condition 1 requires that the tails of $X_{i,:}$ and ε_i be sub-Weibull and bounds certain moments of the observed covariates. In Condition 1, $\|\varepsilon_i\|_{\Psi_\alpha}$ denotes the Orlicz norm of ε_i with $\Psi_\alpha(x) = \exp(|x|^\alpha) - 1$ and $\|\tilde{X}_{i,:}\|_{J,\Psi_\alpha} = \sup_{|\theta|_2=1} \|\tilde{X}_{i,:}^\top \theta\|_{\Psi_\alpha}$ is the joint Orlicz norm.

Condition 1 (Covariates) *Suppose that $X_{i,:} \in \mathbb{R}^p$ are generated i.i.d. with mean 0 and covariance Σ. Let $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalues of Σ. Suppose each element of $X_{i,:}$ is sub-Weibull(α) and the de-correlated covariates $\tilde{X}_{i,:}^\top = \Sigma^{-1/2} X_{i,:}^\top$ are jointly sub-Weibull(α) with*

$$\max\left(\|\tilde{X}_{i,:}^\top\|_{J,\Psi_\alpha}, \max_v \|X_{i,v}\|_{\Psi_\alpha}\right) \leq \kappa. \tag{18}$$

We also define

$$\Gamma = \max\left\{\max_{u,v\in[p]^2} \mathbb{E}\left([X_{i,u}X_{i,v}]^2\right), \max_{v\in[p]} \mathbb{E}\left(\left[a^\top \Sigma^{-1} X_{i,:}^\top X_{i,v}\right]^2\right), \max_{v\in[p]} \mathbb{E}\left(\left[a^\top \Sigma^{-1} X_{i,:}^\top X_{j,v}\right]^2\right)\right\}. \tag{19}$$

We assume the high-dimensional regime where n and p both grow and p can be much larger than n. Condition 2 also implicitly restricts the 2-norm of Σ and Σ^{-1} through $\kappa^\star$. In particular, we require $n \log(n)^{-4/\alpha} \log(pn)^{1-4/\alpha}$ to scale linearly with the condition number of Σ. We will also require $8\sqrt{\Gamma(\log(pn) + 2\log(p))/n} < 1$ so that Σ^{-1} is in the feasible set of (13) for some $\lambda < 1$ with probability tending to 1.

Condition 2 (Sample Size) *Suppose*

$$\kappa^\star = \kappa^2 \max\left(|a|_2 \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}}, 1\right) \tag{20}$$

and

$$n > \max\left\{\frac{4C_\alpha^2 (\kappa^\star)^2 \left[\log(2n)\right]^{4/\alpha} \left[3\log(pn)\right]^{4/\alpha - 1}}{\Gamma}, 64\Gamma(\log(pn) + 2\log(p))\right\} \tag{21}$$

for some constant C_α which only depends on α.

We now give conditions on the set of group actions, $\mathcal{G}$.

Condition 3 (Exchangeability) *Let $\mathcal{G} \subset \mathcal{G}_p$ where $\mathcal{G}_p$ is the set of all matrices corresponding to a permutation g of $[n]$ such that (i) $[n] = N_1 \cup N_2$ for some N_1 and N_2 equal-sized disjoint sets, and (ii) for all $j \in N_1$, $g(j) \in N_2$ and for all $j \in N_2$, $g(j) \in N_1$.*

Condition 4 (Sign Symmetry) *Let $\mathcal{G} \subset \mathcal{G}_s$ where $\mathcal{G}_s$ is the set of all diagonal matrices containing only ± 1 such that there is an equal number of positive and negative 1's.*

Condition 5 (Cluster Exchangeability) *Suppose there exist n_c disjoint sets L_k with $[n] = \bigcup_k^{n_c} L_k$ and $|L_k| = n/n_c = J$ such that $\{\varepsilon_i\}_{i \in L_k}$ are exchangeable, but may otherwise be dependent. That is, $\mathcal{G} \subset \mathcal{G}_c$, where $\mathcal{G}_c$ is the set of all block diagonal matrices where the G_{L_k, L_k} block is a permutation matrix satisfying Condition 3.*

In Conditions 3, 4, and 5 we implicitly assume that n is even to simplify the analysis; when n is odd, the last observation

may be discarded. A more subtle requirement is that we do not assume the support of the randomization distribution is over all possible maps in either $\mathcal{G}_p$, $\mathcal{G}_c$, or $\mathcal{G}_s$. These sets grow exponentially in n. Instead, we consider the more practical scenario where $|\mathcal{G}|$ is fixed with respect to n. This is similar to using the bootstrap or a Monte Carlo procedure with a number of draws that may be increased when a better approximation is desired, but in general stays fixed with n. Condition 5 also implicitly requires that $|L_k| > 2$ since $|\mathcal{G}_c| = 1$ if all clusters have size 2 because there would be only 1 sub-matrix which satisfies Condition 3 for each L_k.

Given the conditions on the covariates and group actions, we now state Lemma 2 and 3, which show that the two terms in Lemma 1 can be controlled.

Lemma 2 *Under Conditions 1 and 2 and either Condition 3, 4, or 5, we have*

$$P\left(\frac{\sum_G |X^\top G X|_\infty}{|\mathcal{G}|} \geq 8\sqrt{\frac{2\Gamma(\log(pn) + 2\log(p))}{n}}\right) \leq 6|\mathcal{G}|(np)^{-1}. \tag{22}$$

Lemma 3 *Under the Conditions 1 and 2, we have*

$$P\left(|a^\top (I - \Sigma^{-1}S|_\infty \geq 8\sqrt{\frac{\Gamma(\log(pn) + 2\log(p))}{n}}\right) \leq 3(np)^{-1}. \tag{23}$$

Thus, with probability at least $1 - 3(np)^{-1}$ the feasible set of (13) is non-empty with $\lambda = 8\sqrt{\frac{\Gamma(\log(pn)+2\log(p))}{n}}$ and

$$|a^\top M_\lambda|_1 \leq |a^\top \Sigma^{-1}|_1. \tag{24}$$

Corollary 1 *Assume the conditions of Lemma 3 and Lemma 2. Then with probability greater than $1-3(np)^{-1}-6|\mathcal{G}|(np)^{-1}$ using $M^\star$ selected from (13) and (11) yields*

$$d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq \left|\hat{\beta}^l - \beta\right|_1 \times \left[8\left(\delta + \left|a^\top \Sigma^{-1}\right|_1\right)\sqrt{2\Gamma(\log(pn) + 2\log(p))}\right] \tag{25}$$

Corollary 1 implies that using any procedure that can produce an estimate $\hat{\beta}^l$ such that $|\hat{\beta}^l - \beta|_1 = O_p((\log p)^{-1/2})$ is sufficient for showing that the oracle and attainable distributions converge in Wasserstein distance. For concreteness, we consider two settings and apply existing results on Lasso estimation. However, other estimators (i.e., SCAD or best subset selection) can also be used as long as $|\hat{\beta} - \beta|_1$ attains the correct rate. First, we consider Lasso estimates when ε_i is sub-Weibull(α) as in Kuchibhotla & Chakrabortty (2018) who require some additional assumptions summarized in Condition 6 that follows. We also consider the setting of Belloni et al. (2016) who require sub-Gaussian covariates and errors, but allow for clustered error dependence.

Condition 6 (Lasso with sub-Weibull errors) *Suppose ε_i is sub-Weibull(α) with $\|\varepsilon_i\|_{\Psi_\alpha} \leq \kappa$. Suppose that*

$$\lambda_{\min} \geq 54 \min_{1\leq h\leq p}\left\{\Xi_{n,h} + \frac{32k\Xi_{n,h}}{h}\right\}\left(\tag{26}$$

where

$$\begin{aligned}\Xi_{n,h} &= 14\sqrt{2}\sqrt{\frac{\Upsilon_{n,h} h\log(36np/h)}{n}} \\ &\quad + \frac{C_\alpha \kappa^2 h(\log(2n))^{2/\alpha}(h\log(36np/h))^{2/\alpha}}{n} \\ \Theta_h &= \{\theta \in \mathbb{R}^p : |\theta|_0 \leq h, |\theta|_2 \leq 1\} \\ \Upsilon_{n,h} &= \sup_{\theta\in\Theta_h} \operatorname{var}\left[\left(X_{i,:}^\top \theta\right)^2\right].\end{aligned} \tag{27}$$

Furthermore, suppose that the Lasso penalty term λ_1 is set such that

$$\lambda_1 = 14\sqrt{2}\sigma\sqrt{\frac{\log(np)}{n}} + \frac{C_{\alpha/2}\kappa^2(\log(2n))^{2/\alpha}(2\log(np))^{2/\alpha}}{n}, \tag{28}$$

and in addition to Condition 2

$$n > \frac{C_{\alpha/2}^2 \kappa^4 (\log(pn))^{8/\alpha - 1}}{\sigma^2}, \tag{29}$$

where $\sigma = \max_{v\in[p]} \operatorname{var}(X_{i,v}\varepsilon_v)$ and $C_{\alpha/2}$ is a constant only depending on α.

Theorem 1 (Sub-Weibull Errors and Covariates) *Suppose Conditions 1, 2, and 6 hold. Under either Condition 3 or 4, with probability not less than $1 - \frac{6|\mathcal{G}|+3}{np} - \frac{3}{np} - \frac{3}{n}$,*

$$d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right) \leq \frac{1360\left(\delta + \left|a^\top\Sigma^{-1}\right|_1\right)\sigma s\sqrt{\Gamma}}{\lambda_{\min}}\frac{\log(np)}{\sqrt{n}}. \tag{30}$$

Recall in Condition 5 with clustered errors, n_c denotes the number of clusters, where each cluster has size J. Then, Theorem 2, presented below, combines Corollary 1 with the results of Belloni et al. (2016) on Lasso estimates under clustered errors. In particular, they propose the Cluster-Lasso procedure and show that its performance depends on a term which measures the within-cluster dependence:

$$\imath_J = J \min_{1\leq v\leq p} \mathbb{E}\left(\frac{1}{J}\sum_{j=1}^J \ddot{X}_{ijv}^2 \ddot{\varepsilon}_{ij}^2\right) \Big/ \mathbb{E}\left(\frac{1}{J}\left[\sum_{j=1}^J \ddot{X}_{ijv}^2 \ddot{\varepsilon}_{ij}^2\right]^2\right), \tag{31}$$

where $\ddot{X}_{ijv}$ and $\ddot{\varepsilon}_{ij}$ denotes the vth covariate and jth observation in the ith cluster which have been adjusted by their respective cluster means. Under complete independence $\imath_J = J$, and the rate in Theorem 2 recovers the rate under independent errors. In the worst case, however, $\imath_J = 1$, so that each cluster is essentially one observation. It is worth noting that Belloni et al. (2016) allow cluster dependence in both covariates and errors; however, we allow for dependent errors but still require the covariates to be i.i.d. For completeness, we include the assumptions of Belloni et al. (2016) in the supplement.

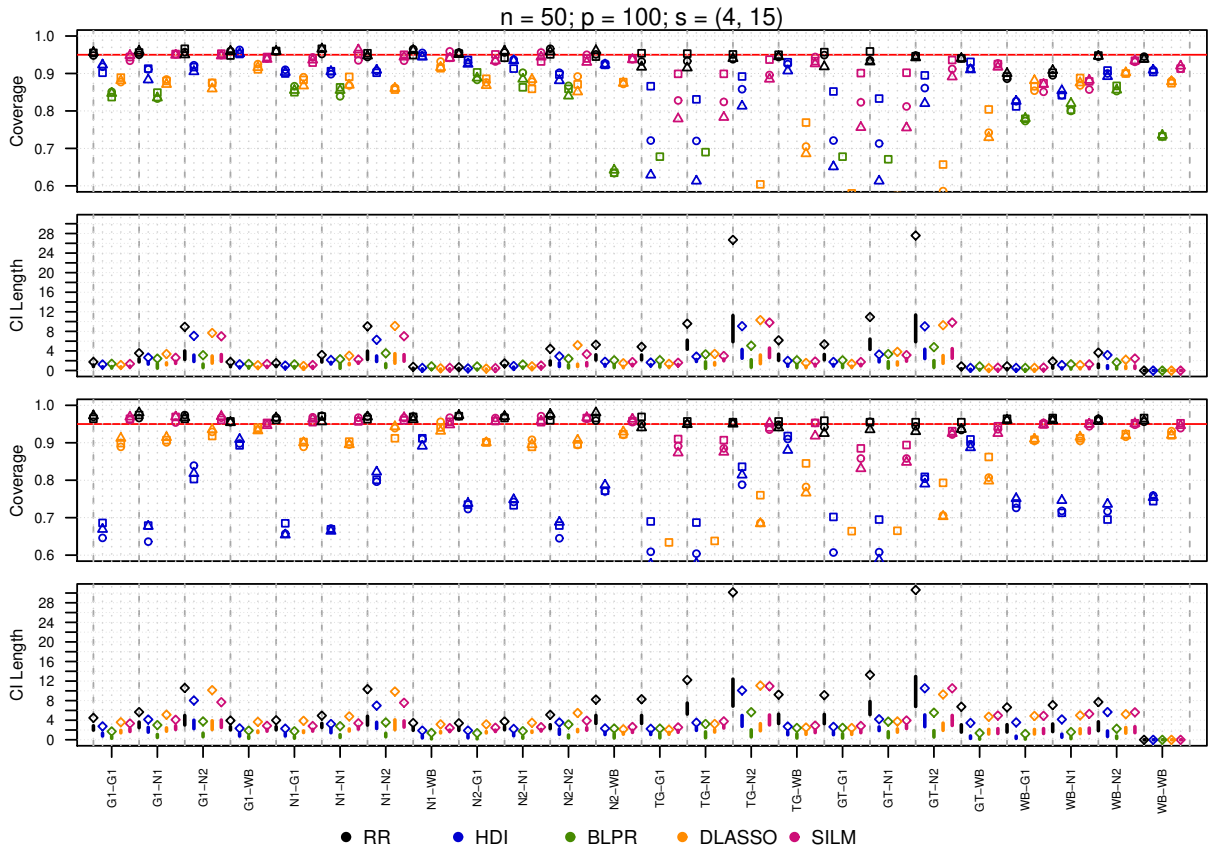

Figure 1. Empirical coverage and confidence interval length for the *active variables* with $n = 50$, $p = 100$, 1000 replications and *exchangeable errors*. The top two panels are for $s = 4$ and the bottom two are for $s = 15$. The first and third panels show empirical coverage rates for each procedure; the sandwich coordinate is denoted by Δ, isolated is $\square$, and adjacent is $\circ$. In the bottom panel, the line segment spans the .25 quantile and .75 quantile of the confidence interval lengths and the single point indicates the .99 quantile. Instead of showing the quantiles for each coordinate, we instead plot the maximum .25 (or .75, .99) quantile across the sandwich, isolated, and adjacent coordinates. The labels on the horizontal axis indicate a different simulation setting and are coded as "Covariate - Errors" where the different covariate and error settings are detailed in the main text. For some settings and procedures, the empirical coverage drops below .6 and is not shown.

Theorem 2 (Clustered Errors) *Suppose Conditions 1, 2, and 5 hold and $|\mathcal{G}| = O(1)$. Further assume the conditions of Theorem 1 of Belloni et al. (2016) and let $\hat{\beta}^l$ be the Cluster-Lasso. Then $d_1\left(F_t(X,\varepsilon), F_{\hat{t}}(X,\varepsilon)\right)$ is*

$$O_p\left(\frac{s\left(\delta + \left\|a^\top \Sigma^{-1}\right\|_1\right)\sqrt{\Gamma_p(\log(pn)+\log(p))}}{\sqrt{n\,\iota_J}}\right) \quad (32)$$

In Theorems 1 and 2, we explicitly include the term $|a^\top\Sigma^{-1}|_1$, and for convergence, we require $|a^\top\Sigma^{-1}|_1 = O(\sqrt{n}/(s\log(np)))$.

Remark 3 *Convergence in Wasserstein-1 implies weak convergence, as well as L_1 convergence of the quantile and distribution functions. Similar to Bickel & Freedman (1981), Bickel & Freedman (1983), and Lopes (2014), we use this to justify the procedure's use for hypothesis tests and confidence intervals.*

4. Numerical Experiments

We compare nominal 95% confidence intervals (CIs) over 1000 trials of `BLPR` (Liu et al., 2017), `HDI` (Dezeure et al., 2017), `DLASSO` (Javanmard & Montanari, 2014)[3], `SILM` (Zhang & Cheng, 2017) and residual randomization (`RR`) with $(n = 50, p = 100)$ and $(n = 100, p = 300)$. We slightly modified `SILM` to output marginal confidence intervals and ensure that the modified variance estimator does divide by 0 if the support of the estimated β is n. Additional details are in the supplement, and the code is available at: `https://github.com/atechnicolorskye/rrHDI`.

In each setting, we sample random $X \in \mathbb{R}^{n\times p}$ with rows drawn i.i.d. from either (**N1**) $N(0, I)$; (**G1**) Gamma$(1,1) - 1$; (**N2**) $N(\mu, 1)$ with $P(\mu = -2) = P(\mu = 2) = 0.5$;

[3]https://web.stanford.edu/ montanar/sslasso/code.html

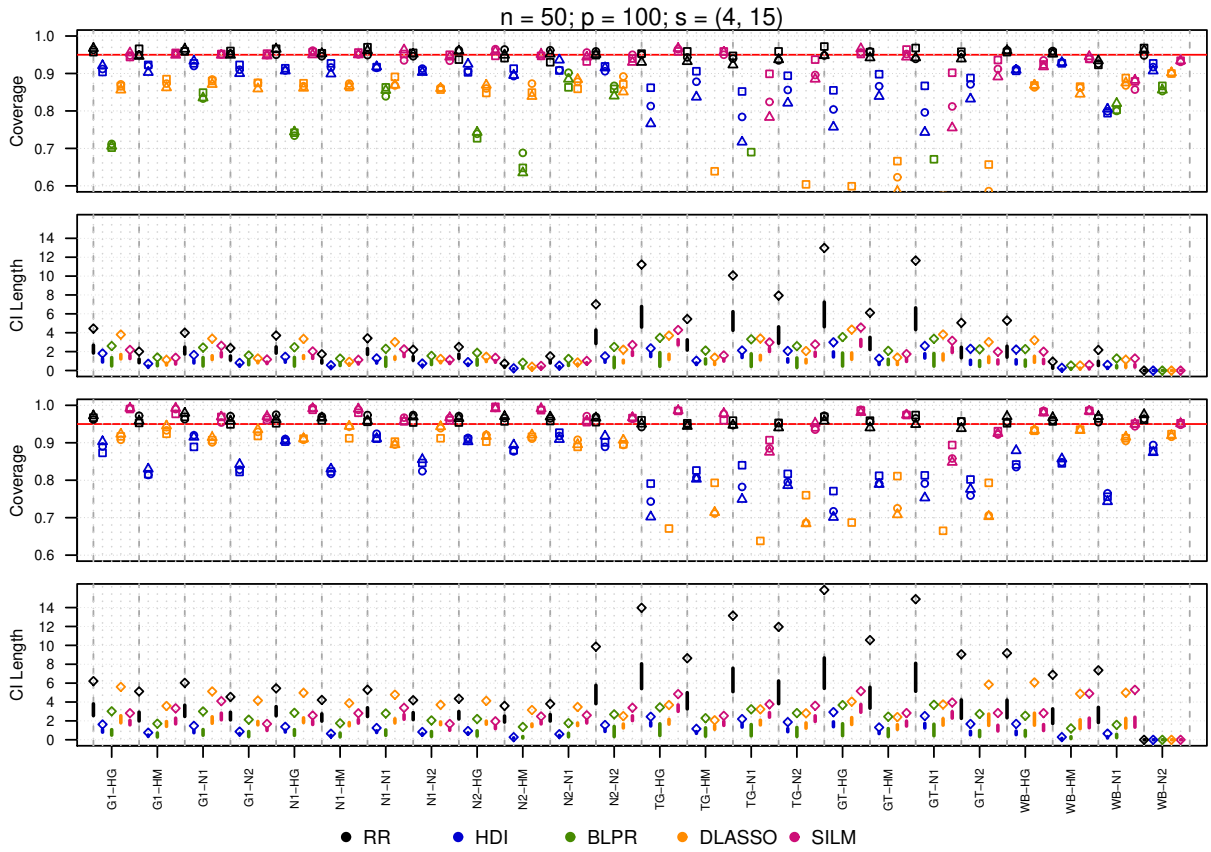

Figure 2. Empirical coverage and confidence interval length for the *active variables* with $n = 50$, $p = 100$, 1000 replications and *sign symmetric errors*. The top two panels are for $s = 4$, the bottom two panels are for $s = 15$. All other elements remain the same as Figure 1.

(**NT**) $N(0, \Sigma)$ for $\Sigma_{ij} = .8^{|i-j|}$; (**GT**) Gamma$(\Sigma) - 1$ for $\Sigma_{ij} = .8^{|i-j|}$; or (**WB**) each element is a centered Weibull with scale $= 1$ and shape $= 1/2$.

We sample the errors $\varepsilon \in \mathbb{R}^n$ from (**N1**) $N(0, 1)$; (**G1**) Gamma$(1, 1) - 1$; (**N2**) $N(\mu, 1)$ with $P(\mu = -2) = P(\mu = 2) = 0.5$; (**WB**) a centered Weibull with scale $= 1$ and shape $= 1/2$; (**HN**) normal with $N(0, 2\|X_i\|_2^2/p)$; or (**HM**) $N(\mu, 2\|X_i\|_2^2/p)$ with $P(\mu = -2) = P(\mu = 2) = 0.5$. The exchangeable settings exclude the heteroskedastic cases of **HN** and **HM**; the symmetric settings exclude **G1** and **WB**.

For a fair comparison, we have `HDI` use wild bootstrap for the symmetric settings. Also, apart from `SILM`, empirical performance is generally not affected by the scale of the covariates; however, in certain settings `SILM` performed very poorly when the covariates were not standardized. Thus, we standardize the covariates to benefit `SILM`.

For each setting, we draw $\beta \in \mathbb{R}^p$ with $s = 4$ or 15 *active* (i.e., non-zero) coordinates drawn from the Rademacher distribution and set the remaining $p - s$ *inactive* coordinates to 0. We arrange entries in β in such that there is one active entry between two inactive entries (*isolated*), one active between an active entry and an inactive entry (*adjacent*), and one active entry between two other active entries (*sandwiched*). We also use the same scheme for the inactive variables. We then set $Y = X\beta + \varepsilon$.

To obtain $M_{\lambda^\star}$, we solve (13) to up to 500 iterations using `fastclime` (Pang et al., 2014) which starts with $\lambda = 1$ and uses warm starts to progressively shrink λ. We further select $\lambda^\star$ via (12) by using a grid search over the λ values used by `fastclime` with $\delta = 10,000$. Empirically, a larger value of δ generally results in better coverage, but comes at the expense of confidence interval length; broadly speaking though, we see that for $\delta \geq 1,000$, the performance of the proposed procedure is fairly insensitive to the value of δ.

Since in practice we do not know the appropriate Lasso tuning parameter λ_1 a priori, for the residual randomization procedure we employ the Square-Root Lasso (Belloni et al., 2011) implemented in `RPtests` (Shah & Buhlmann, 2017) to obtain estimates for $\hat{\beta}^l$. We follow (Zhang & Cheng, 2017) and rescale $\hat{\varepsilon}$ by $\sqrt{n/(n - |\hat{\beta}^l|_0)}$ as a finite-sample correction. For all settings, we use $1,000$ group actions/bootstrap resamples.

In Figures 1 and 2 we show the empirical coverage and the confidence interval length for the active variables when $(n = 50, p = 100)$ for exchangeable and symmetric errors respectively. In the supplement, we provide corresponding figures for the inactive variables as well as results for $(n = 100, p = 300)$.

Among the competing methods, SILM performs the best, and has generally satisfactory performance across all settings except the Toeplitz case and the setting with Weibull covariates for $s = 4$. For $s = 4$, HDI generally outperforms DLASSO, but when $s = 15$, HDI's performance decreases drastically. Generally, BLPR performs the worst and when $s = 15$ has coverage less than .6. All the competing methods perform poorly when the covariates have Toeplitz covariance. The sandwich coordinates typically have the lowest coverage and the isolated coordinates typically have empirical coverage closest to the nominal rate.

In contrast, we see that RR nearly obtains the nominal 95% coverage regardless of exchangeability or sign symmetry, and across all experimental configurations. This remarkable stability can be explained by our selection procedure for M and the general properties of randomization tests (see Introduction and Remark 1). At the same time, RR typically yields larger interval lengths. While the interval length from RR is generally longer than the competing methods, this is especially true when the covariates have Toeplitz covariance. We posit that this is because solving (13) for poorly conditioned sample covariances can yield large $|a^\top M|_1$ and thus results in larger confidence intervals.

In the supplement, we show the empirical coverage for the inactive coordinates with $(n = 50, p = 100)$ as well as all coordinates when $(n = 100, p = 300)$. For $(n = 100, p = 300)$, the results are qualitatively similar to the results for $(n = 50, p = 100)$; RR almost always has the best empirical coverage followed by SILM, HDI, DLASSO, and then BLPR. However, all methods (including RR) generally perform less well; in particular, in contrast to the $(n = 50, p = 100)$ case, there are some settings where RR does not attain nominal coverage. For the inactive variables when $(n = 50, p = 100)$ and $(n = 100, p = 300)$, all procedures except BLPR typically achieve (or exceed) nominal coverage.

Table 1 gives the average computation time (in seconds) required for each of the procedures with $n = 100, p = 300$. We form a confidence interval for each of the 300 coordinates using $1,000$ group actions/bootstrap resamples. Unsurprisingly, RR requires more computational effort than competing procedures that appeal to asymptotic limiting distributions (BLPR and DLASSO). However, the computational effort is comparable to the resampling-based methods in our experiments (HDI and SILM).

Table 1. Mean Computation Time in Seconds.

Method	BLPR	HDI	DLASSO	SILM	RR
Time (s)	18.9	374.2	11.2	61.1	135.1

5. Discussion

The theoretical guarantees coupled with the excellent empirical performance suggest that residual randomization is an appealing alternative for robust inference in high-dimensional linear models. Across a wide range of settings, it attains nominal coverage even when n is small or the errors are heavy tailed. Of course, this does not come for free, as we observe that the confidence intervals produced are generally larger than those produced by competing methods—especially when there is strong cross-correlation in covariates. Nonetheless, we believe that in most practical applications slightly larger confidence intervals is a small price to pay in exchange for better, more robust coverage.

While the procedure is asymptotically valid for any fixed $\delta \geq 1$, one practical concern is specifying a tuning parameter δ when selecting $M^\star$. If δ is too small, this may result in degraded empirical performance; however, the simulations show that there is a threshold of δ for which our procedure performs well across all settings. This threshold seems to coincide with a point at which the length of our confidence intervals become insensitive to increasing δ. Thus, this threshold can be well approximated—at least via heuristics.

Nonetheless, in Section 3 of the supplement, we describe an alternative procedure which also performs well empirically and is also asymptotically valid when an upper bound on Γ in (19) is known.

Several questions may be fruitful to pursue in the future. First, to form confidence intervals, we require control of $|\hat{\beta}^l - \beta|_1$. However, it would be interesting to investigate residual randomization procedures which only require small in-sample prediction error; i.e., $\frac{1}{n}|X(\hat{\beta}^l - \beta)|$. This would allow testing individual coordinates of β without assuming stringent restricted eigenvalue or irrepresentability conditions. There are additional advantages of the residual randomization framework which could be further exploited methodologically. For example, we primarily focused on selecting M, but one could also use the observed covariates X to select specific invariances $G \in \mathcal{G}$ which optimize certain test properties.

Acknowledgments

This work was completed in part with resources provided by the University of Chicago Research Computing Center.

References

Belloni, A., Chernozhukov, V., and Wang, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr043. URL https://doi.org/10.1093/biomet/asr043.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605, 2016.

Bickel, P. J. and Freedman, D. A. Some asymptotic theory for the bootstrap. *The annals of statistics*, pp. 1196–1217, 1981.

Bickel, P. J. and Freedman, D. A. Bootstrapping regression models with many parameters. *Festschrift for Erich L. Lehmann*, pp. 28–48, 1983.

Bühlmann, P. and Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Cai, T., Liu, W., and Luo, X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm10155. URL https://doi.org/10.1198/jasa.2011.tm10155.

Chatterjee, A. and Lahiri, S. N. Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.*, 106(494):608–625, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm10159. URL https://doi.org/10.1198/jasa.2011.tm10159.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, 2017. ISSN 1133-0686. doi: 10.1007/s11749-017-0554-2. URL https://doi.org/10.1007/s11749-017-0554-2.

Fisher, R. A. et al. The design of experiments. *The design of experiments.*, (2nd Ed), 1935.

Freedman, D. and Lane, D. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983a.

Freedman, D. A. and Lane, D. Significance testing in a nonstochastic setting. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pp. 185–208. Wadsworth, Belmont, CA, 1983b.

Javanmard, A. and Montanari, A. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Kempthorne, O. The design and analysis of experiments. 1952.

Kuchibhotla, A. K. and Chakrabortty, A. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Liu, H., Xu, X., and Li, J. J. A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *arXiv preprint arXiv:1706.02150*, 2017.

Lopes, M. E. A residual bootstrap for high-dimensional regression with near low-rank designs. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3239–3247, Cambridge, MA, USA, 2014. MIT Press.

Meinshausen, N., Meier, L., and Bühlmann, P. p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.tm08647. URL https://doi.org/10.1198/jasa.2009.tm08647.

Pang, H., Liu, H., and Vanderbei, R. The fastclime package for linear programming and large-scale precision matrix estimation in r. *Journal of Machine Learning Research*, 15(14):489–493, 2014. URL http://jmlr.org/papers/v15/pang14a.html.

Pitman, E. J. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

Rosenbaum, P. R. Exact confidence intervals for nonconstant effects by inverting the signed rank test. *The American Statistician*, 57(2):132–138, 2003.

Shah, R. and Buhlmann, P. *RPtests: Goodness of Fit Tests for High-Dimensional Linear Regression Models*, 2017. URL https://CRAN.R-project.org/package=RPtests. R package version 0.1.4.

Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN.

Toulis, P. Life after bootstrap: Residual randomization inference in regression models. *arXiv preprint arXiv:1908.04218*, 2019.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

Wasserman, L. and Roeder, K. High-dimensional variable selection. *Ann. Statist.*, 37(5A):2178–2201, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS646. URL https://doi.org/10.1214/08-AOS646.

Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1): 217–242, 2014. ISSN 1369-7412. doi: 10.1111/rssb.12026. URL https://doi.org/10.1111/rssb.12026.

Zhang, X. and Cheng, G. Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.*, 112 (518):757–768, 2017. ISSN 0162-1459. doi: 10.1080/01621459.2016.1166114. URL https://doi.org/10.1080/01621459.2016.1166114.

Zou, H. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006. ISSN 0162-1459. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.