A New Perspective on High Dimensional Confidence Intervals

Logan Harris
Department of Biostatistics
University of Iowa

Patrick Breheny
Department of Biostatistics
University of Iowa

July 15, 2025

Abstract

Classically, confidence intervals are required to have consistent coverage across all values of the parameter. However, this will inevitably break down if the underlying estimation procedure is biased. For this reason, many efforts have focused on debiased versions of the lasso for interval construction. In the process of debiasing, however, the connection to the original estimates are often obscured. In this work, we offer a different perspective focused on average coverage in contrast to individual coverage. This perspective results in confidence intervals that better reflect the original assumptions, as opposed to debiased intervals, which often do not even contain the original lasso estimates. To this end we propose a method based on the Relaxed Lasso that gives approximately correct average coverage and compare this to debiased methods which attempt to produce correct individual coverage. With this new definition of coverage we also briefly revisit the bootstrap, which Chatterjee and Lahiri (2010) showed was inconsistent for lasso, but find that it fails even under this alternative coverage definition.

1 Introduction

The objective function for lasso-penalized linear regression (Tibshirani, 1996) is

$$Q(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}, \lambda) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\mathbf{y}$ is a length n vector of independent outcomes, $\mathbf{X}$ is an $n \times p$ matrix of features, $\boldsymbol{\beta}$ is a length p vector of regression coefficients, and λ is a regularization parameter controlling the amount of penalization. Note that the objective function involves the addition of the L_1 penalty, $\lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{j=1}^{p}|\beta_j|$, to the squared error loss. This typically results in sparse estimates for some of the regression coefficients (i.e., $\widehat{\beta}_j = 0$) depending on the choice of the regularization parameter λ. Its ability to carry out both variable selection and estimation is particularly attractive, especially in scenarios where both predictive accuracy and interpretability are important. The lasso performs particularly well in cases where the number of features is large and the underlying model is sparse (Hastie et al., 2009), but has become popular in a wide variety of settings.

Nevertheless, inference for the lasso has proven challenging. By introducing both sparsity and shrinkage, the L_1 penalty greatly complicates the sampling distribution of the estimators. This complexity has given rise to a wide variety of inferential approaches. The majority of these approaches have focused on controlling the false discovery rate (FDR) of the selected features. Examples include the Covariance test (Lockhart et al., 2014), the Knockoff Filter (Barber and Candès, 2015; Candès et al., 2018), the marginal FDR (Breheny, 2019), and the Gaussian mirror (Xing et al., 2023).

There have also been various proposals for constructing confidence intervals (CIs), although the shrinkage/bias introduced by the L_1 penalty poses a number of challenges here. Several methods (Zhang and Zhang, 2014; Javanmard and Montanari, 2014) focus on "debiasing" the original point estimates from a lasso fit to facilitate more traditional forms of inference. An alternative approach, which accounts for the uncertainty in model selection by conditioning on the selected model is known as Selective Inference (Lee et al., 2016), although it is worth noting that this approach only produces intervals for variables that were selected.

In this manuscript, we offer a different perspective that allows for biased intervals and focuses on correct average coverage instead of correct individual coverage. This perspective results in confidence intervals that better reflect the original assumptions for obtaining the lasso estimates (i.e. sparsity), as opposed to debiased intervals, which often do not even contain the original lasso estimates. There is no objectively correct coverage definition of coverage, but we discuss the differences between them and hope the reader finds the debate illuminating.

The bootstrap is often a natural choice for handling complex sampling distributions. However, Chatterjee and Lahiri (2010) demonstrated that when applied to lasso estimators, the bootstrap is inconsistent – even if the lasso itself is $\sqrt{n}$-consistent with respect to estimating $\boldsymbol{\beta}$. For this reason, efforts to bootstrap the lasso have focused instead on bootstrapping de-biased (or de-sparsified) versions of the lasso (Dezeure et al., 2017). In this work, we revisit bootstrapping lasso, but find that even under a relaxed definition of average coverage, the bootstrap fails as it introduces "extra bias". However, this exercise points to a promising alternative and we show this alternative, based on the Relaxed Lasso, does produce approximately correct average coverage.

Section 2 examines the underlying concept of average coverage in more detail and shows that the bootstrap does not even produce average coverage. Section 3 introduces a method based on the Relaxed Lasso which does have approximately correct average coverage. Then Section 4 examines the performance of the proposed method across a number of simulations and includes a comparison to Selective Inference and the de-sparsified lasso. Lastly, in Section 5, we show the application of the proposed method to two data sets, one for acute respiratory illness and the other for gene expression data in mammalian eyes. For the sake of simplicity, we focus on lasso-penalized linear regression, but most of the discussion is relevant to any sparse penalty and loss.

2 A new coverage definition for penalized regression confidence intervals

When using penalized regression, we are introducing bias into the estimators by design. This has direct implications for confidence intervals constructed around these biased estimates. All methods we are aware of propose debiasing as a way to counteract the bias introduced in attempts to obtain traditional frequentist coverage properties. In Section 2.1, we instead propose an alternate perspective that focuses on targeting average coverage inspired by the connection between the penalties in penalized regression and Bayesian priors. Given the connection between the bootstrap and Bayesian posteriors, it might appear that bootstrap confidence intervals also meet this alternate definition for coverage. However, we show in Section 2.2 that the bootstrap also fails to meet average coverage targets.

2.1 Individual vs average coverage

Classical frequentist coverage is concerned with achieving proper coverage for each parameter individually. In penalized regression, bias is explicitly being introduced in the estimation procedure which poses a problem when targeting nominal interval coverage for individual parameters. Here we propose shifting focus to average coverage across all p confidence intervals. In penalized regression, these two definitions of coverage can be quite different and we refer to this notion as Individual vs Average Coverage.

Letting $\mathcal{A}(\mathbf{y})$ denote a process that produces an interval based on data $\mathbf{y}$, the coverage probability for the process is defined as $\text{Cover}(\theta) = \mathbb{P}\{\theta \in \mathcal{A}(\mathbf{y})\}$. Classical frequentist inference requires valid intervals to satisfy $\text{Cover}(\theta) = 1 - \alpha$ for all values of θ (or potentially $\geq 1 - \alpha$). This is, however, incompatible with Bayesian inference. Bayesian credible intervals cannot, in general, have the same coverage for each θ. What they satisfy instead is maintaining the expected coverage with respect to the prior distribution of θ: $\int \text{Cover}(\theta)p(\theta)\,d\theta = 1 - \alpha$ (this is not the definition of credibility, but it is a consequence, as we show later in this section). Unless the prior is uniform, the coverage of a Bayesian credible interval will be greater than $1 - \alpha$ for some θ and less than $1 - \alpha$ for other values of θ.

For example, consider the credible intervals for θ in a $\text{N}(\theta, \sigma^2)$ model with prior $\theta \sim \text{N}(0, \tau^2)$ with $\sigma = \tau = 1$. The left side of Figure 1 illustrates the coverage probability for the 80% Bayesian credible interval over a range of θ values. Where the prior density for θ is highest, the coverage is above 80%, whereas regions where the prior density is low have coverage below 80%. The expected coverage, however, is exactly 80% when integrated with respect to the prior. This is fundamentally true of any Bayesian model with an influential prior: $\text{Cover}(\theta) = 1 - \alpha$ for all values of θ will never be satisfied. The right side, which illustrates that a similar phenomenon happens for the method we propose, will be discussed in Section 4.1.

In high dimensional problems, there is yet another quantity we can consider: the average coverage. Rather than integrating over a hypothetical distribution of θ values, we can average over the distribution of parameters present. In other words, we might choose to require that our intervals satisfy $\frac{1}{p}\sum_{j=1}^{p} \text{Cover}(\theta_j) = 1 - \alpha$. This criteria is more closely aligned with the Bayesian perspective than a classical frequentist perspective, although it does not specifically require or involve a prior.

Our goal in this paper is not to argue that one of these perspectives is correct and the other is wrong, but rather that the average coverage perspective is reasonable and worthy of consideration. It should not be taken for granted that classical ideas developed for single parameter inference are the best way to approach simultaneous inference for large numbers of parameters. Furthermore, the Bayesian perspective seems to make sense in the context of penalized regression, since penalized regression is intentionally imposing shrinkage towards a prior notion of which parameter values are more likely.

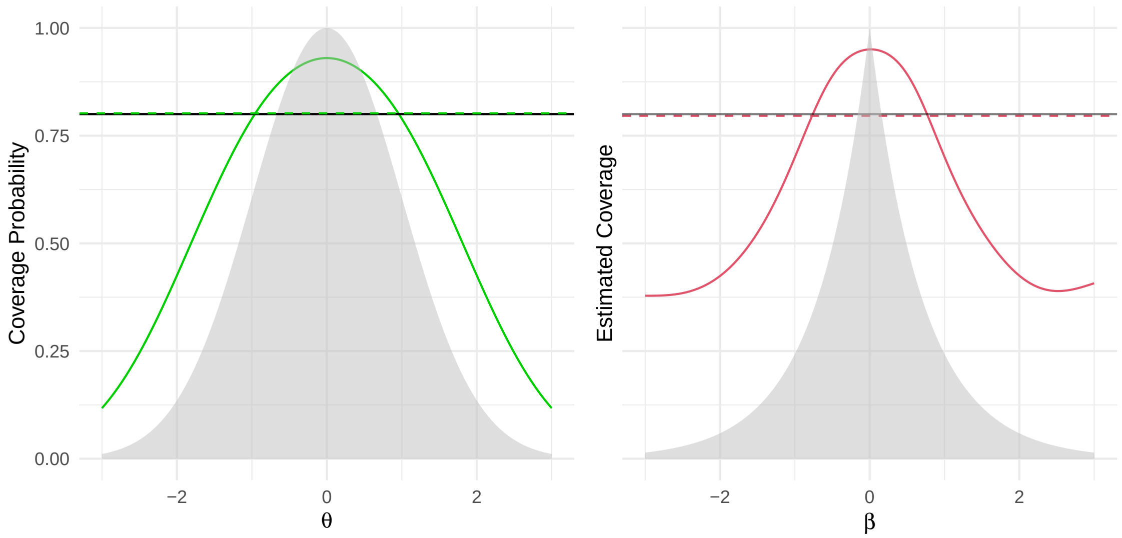


Figure 1: The left side of the figure provides coverage probabilities of Ridge CIs for a range of θ values as described in Section 2. The right side of the figure display results from the simulation described in Section 4.1. For the right side of the figure, the fitted curve is from a Binomial GAM fit with coverage being modeled as a smooth function of β. The dashed line represent the average for the RL-P across all 1000 independently generated datasets and the solid black line indicates the nominal coverage rate. The shaded distribution in the background depicts the Laplace distribution the βs were drawn from. The analogous lines in the left side of the figure are exact calculations with the shaded normal in the background representing the prior.

In Section 3, we propose a new method and in Section 4 we see that the resulting intervals, while they do not satisfy classical coverage requirements, perform quite well with respect to the average coverage criterion. We end this section with a short theorem making the explicit connection between Bayesian credible intervals and average coverage.

Theorem 1. *If the likelihood is correctly specified according to the true data generating mechanism, $p(\mathbf{y}|\boldsymbol{\theta})$, then a $1-\alpha$ credible set for any parameter θ_j will satisfy $\int Cover(\theta_j)p(\boldsymbol{\theta})d\boldsymbol{\theta} = 1-\alpha$.*

Proof. By definition, a $100(1-\alpha)\%$ credible region for θ_j is any set $\mathcal{A}_j(\mathbf{y})$ such that $\int I\{\theta_j \in \mathcal{A}_j(\mathbf{y})\}p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta} = 1-\alpha$. The coverage probability, meanwhile, is defined as $\int I\{\theta_j \in \mathcal{A}_j(\mathbf{y})\}p(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y}$. The average coverage, integrated with respect to the prior $p(\boldsymbol{\theta})$, is therefore

$$\begin{aligned}\int\int I\{\theta_j \in \mathcal{A}_j(\mathbf{y})\}p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\mathbf{y}d\boldsymbol{\theta} &= \int\int I\{\theta_j \in \mathcal{A}_j(\mathbf{y})\}p(\boldsymbol{\theta}|\mathbf{y})p(\mathbf{y})d\mathbf{y}d\boldsymbol{\theta} \\ &= \int\int I\{\theta_j \in \mathcal{A}_j(\mathbf{y})\}p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}p(\mathbf{y})d\mathbf{y} \\ &= \int(1-\alpha)p(\mathbf{y})d\mathbf{y} \\ &= 1-\alpha,\end{aligned}$$

and as a result the average coverage is equal to the nominal coverage. □

The above theorem concerns a single parameter of interest θ_j and its credible interval. An immediate corollary of this result is that the average coverage of p such intervals will also be equal to $1-\alpha$.

The more interesting question, however, is what happens when we average with respect to the distribution of θ values present in a high-dimensional problem rather than integrating with respect to the prior. In other words, is it true that

$$\tfrac{1}{p}\sum_{j=1}^{p}\text{Cover}(\theta_j) \approx \int \text{Cover}(\theta)p(\theta)\,d\theta?$$

Note that the left-hand side does not require a Bayesian perspective as no probability distributions of parameters are involved, only the empirical distribution of different values for different parameters. Arguably, this has a simpler interpretation than the conventional frequentist confidence interval. Rather than appealing to hypothetical intervals for hypothetical alternative data sets, we are making a more concrete statement here about the p intervals that have just been constructed.

Intuitively, it would seem that the equation above should be true if this empirical distribution of θ values resembles the prior implied by the penalty and as long as the intervals constructed arise from a distribution resembling a Bayesian posterior. Given the connection between the Bayesian posterior and the bootstrap first pointed out by Rubin (1981), this would seem to suggest that bootstrap based intervals should satisfy the above equation, but, in Section 2.2, we explain why this is not the case. However, in Section 3 we propose an alternative CI construction method inspired by the Bayesian posterior and in Section 4.2, we find that this relationship generally holds for this approach even if the distribution of θ values is quite different from the prior implied by the lasso penalty.

2.2 Does the bootstrap give average coverage?

The connection between the bootstrap and a Bayesian posterior was first drawn by Rubin (1981) and further explored by Efron (1982) and Lo (1987). Its well known that the frequentist properties of the bootstrap break down in high dimensions, however, the connection between the bootstrap and a Bayesian posterior along with Theorem 1 would suggest that perhaps the bootstrap would give correct average coverage. That being said the previously mentioned works which drew the connection between the bootstrap and the Bayesian posterior focused on the low dimensional setting. Here, we provide an example showing that the connection between the bootstrap and the Bayesian posterior fundamentally breaks down for penalized regression, especially when the dimensionality of the problem increases, meaning that the bootstrap does in fact not even give good average coverage.

To provide an example, we return to ridge regression for two reasons. First, we do not face the complication of having estimates shrunk all the way to zero and second, posterior credible intervals can be computed in closed form for Ridge. The simulation is set up to isolate the effect of increasing dimensionality by increasing p $(20, 100, 200)$ but holding $n = 200$ and $\lambda = 0.4$. λ is the ratio of the prior precision $(1/\tau^2)$ to the information (n/σ^2) and so for the empirical distribution of $\boldsymbol{\beta}$ to be equivalent to the prior (the ideal scenario as indicated by Theorem 1) the prior variance (τ^2) must be set to $\sigma^2 \ / \ n\lambda$. In this simulation, $\sigma^2 = 100$, so $\tau^2 = 1.25$ and $\boldsymbol{\beta}$ was set to the $\frac{1:p}{p+1}$ quantiles of a $\text{N}(0, \tau^2 = 1.25)$ distribution. The elements of $\mathbf{X}$ were generated independently from a $\text{N}(0, 1)$ and then $\mathbf{y}$ was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\varepsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$. 1000 data sets were generated for each p and intervals were constructed using both a pairs bootstrap and a Bayesian posterior. Results are provided in Figure 2, the dashed lines give the average coverages and the solid lines are the estimated coverages as functions of β.

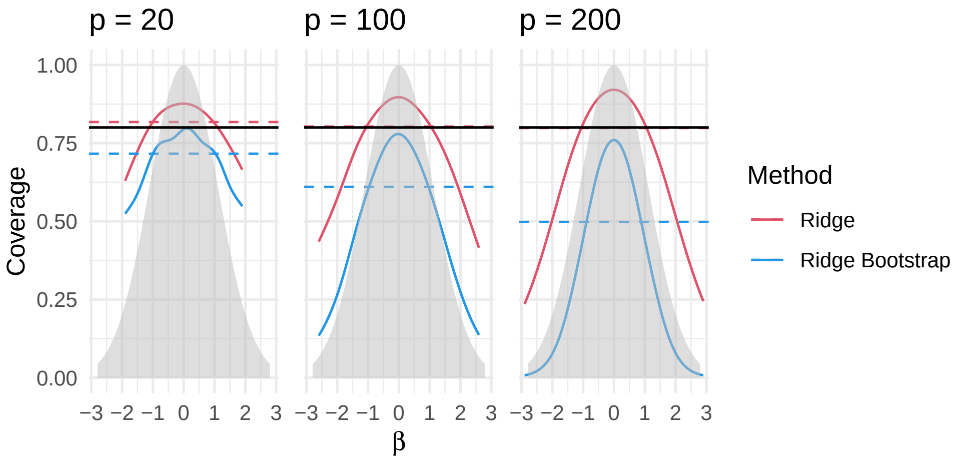


Figure 2: Average coverages across p parameters (dashed lines) and estimated coverages as functions of β (solid curves) for intervals constructed using a pairs bootstrap (Ridge Bootstrap) and a Bayesian posterior (Ridge). Full details of the simulation set up can be found in Section 2.2.

Even when dimensionality is low, the bootstrap introduces additional bias, an issue that grows with the dimensionality causing a further departure from the coverage behavior of the Bayesian posterior. While this is a single example using Ridge regression, we find that the extra bias introduced by bootstrapping is also fundamentally an issue for the lasso. In fact, because of the sparsity introduced by the lasso penalty, the issues are even more noticeable as seen in Supplement 7.1. For further explanations for the breakdown, we refer the reader to Supplement 7.6. Supplement 7.6 starts with a simple proof in the 1 and 2 predictor setting for both ridge and lasso showing that while this issue increases with dimensionality, that it is present even in low dimensions. Additionally, these proofs indicate that this bias is heavily dependent on the size of the penalty (λ). This is followed by a simulation that decomposes the source of the bootstrap bias in a high dimensional setting for lasso.

3 Relaxed Lasso Posterior confidence intervals

While the bootstrap is not a viable option as outlined in Section 2.2, this section and the discourse around Theorem 1 suggest that if intervals are constructed from a distribution resembling a Bayesian posterior that they should have correct average coverage. So, here we propose the **Relaxed Lasso Posterior** (RL-P), which constructs intervals from the distribution of β_j conditional on the selected features, viewed as a Bayesian posterior. The remainder of this section presents its specific application to lasso-penalized linear regression. Specifically, we define and derive the conditional distributions needed for the interval construction. In this section, we provide a high level derivation of the conditional posterior distributions for lasso-penalized regression. Complete details, including how to perform sampling, are provided in Supplemental Materials.

As with other penalized regression approaches, the lasso can be formulated as a Bayesian regression model by setting an appropriate prior. This was initially noted by Tibshirani (1996) and explored more extensively by Park and Casella (2008). For Ridge regression, the prior is a Normal distribution which leads to conjugacy allowing for interval construction to be straightforward. Figure 2 provides the coverage behavior of these intervals showing that they do achieve approximately correct average coverage in ideal settings. Here, for lasso, we derive the conditional distribution of $\widehat{\beta}_j(\lambda)$ in attempts to provide intervals analogous to those produced by Ridge.

For the lasso, the corresponding prior is a Laplace distribution, also referred to as the double-exponential distribution:

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\gamma}{2} \exp(-\gamma\,|\beta_j|), \gamma > 0.$$

Let $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$. Then, $\hat{S}_j = \hat{S}$ if $j \notin \hat{S}$ and $\hat{S}_j = \hat{S} - \{j\}$ if $j \in \hat{S}$. The conditional posterior for β_j is defined as the distribution for β_j conditional on $\hat{S}_j$. Define $\mathbf{Q}_{\hat{S}_j}$ as $\mathbf{I} - \mathbf{X}_{\hat{S}_j}(\mathbf{X}_{\hat{S}_j}^T\mathbf{X}_{\hat{S}_j})^{-1}\mathbf{X}_{\hat{S}_j}^T$, the projection matrix onto the features selected by the lasso. Then, we find the likelihood for β_j conditional on the selected features is:

$$L(\beta_j|\hat{S}_j) \propto \exp(-\frac{\mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{x}_j}{2\sigma^2}(\beta_j - \tilde{\beta}_j)^2)$$

where $\tilde{\beta}_j = (\mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{x}_j)^{-1}\mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{y}$. This can be seen as a mild extension of the relaxed lasso. It is equivalent to the relaxed lasso for features in $\hat{S}$ but also is capable of providing intervals for features in $\hat{S}^C$.

A normal likelihood and Laplace prior are not conjugate. However, the conditional posterior can be shown to be a composition of right and left truncated normals where the truncation occurs at zero for right and left tails respectively. In this manuscript, we assume that $\mathbf{X}$ has been standardized s.t. $\mathbf{x}_j^T\mathbf{x}_j = n$. Then for β_j (see appended materials for details),

$$p(\beta_j|\hat{S}_j) \propto \begin{cases} C_- \exp\{-\frac{\tilde{n}}{2\sigma^2}(\beta_j - (\tilde{\beta}_j + \lambda))^2\}, \text{ if } \beta_j < 0, \\ C_+ \exp\{-\frac{\tilde{n}}{2\sigma^2}(\beta_j - (\tilde{\beta}_j - \lambda))^2\}, \text{ if } \beta_j \geq 0 \end{cases} \tag{1}$$

where $\tilde{n} = \mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{x}_j$, $C_- = \exp(\tilde{\beta}_j\lambda\tilde{n}/\sigma^2)$ and $C_+ = \exp(-\tilde{\beta}_j\lambda\tilde{n}/\sigma^2)$.

This formulation is attractive because it allows efficient computation of quantiles. This consists of first determining which normal distribution (left or right tail) the probability corresponds to, then calculating the quantile from the corresponding normal distribution. Again, full details are provided in Supplemental Materials.

This solution corresponds to a particular value of λ and $\hat{\sigma}^2$. Throughout, we use cross validation (CV) to select λ and estimate σ^2, then produce confidence intervals corresponding to these values. Specifically, we use the value of λ which minimizes the cross validation error (CVE) and use the estimate for σ^2 recommended by (Reid et al., 2016):

$$\hat{\sigma}^2 = \frac{1}{n - |\hat{S}_{\lambda_{CV}}|}||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda_{CV})||_2^2,$$

where $|\hat{S}_{\lambda_{CV}}| = \sum\Big(\widehat{\boldsymbol{\beta}}(\lambda_{CV}) \neq 0\Big)$.

The relaxed lasso posterior intervals are available through the `pipe_ncvreg` function in the current version of the R package `ncvreg` (3.16.0).

4 Results

We begin Section 4.1 by examining the coverage of the Relaxed Lasso Posterior intervals in what might be considered the "ideal" scenario, where the values of θ are distributed according to the prior distribution implied by the lasso as discussed in Section 2. We then examine how this coverage is affected by various changes to data generating mechanism to assess the robustness of the proposed method (Section 4.2). Finally, we compare the proposed confidence interval method to other confidence interval approaches for penalized regression that have been proposed in the literature (Sections 4.3 and 4.4), which reveals a number of interesting contrasts between methods that attempt to debias the intervals and those that do not.

Unless otherwise noted, the nominal coverage rate in all of these experiments is 80%.

4.1 Coverage

Given the connection between average coverage and Bayesian credible intervals made by Theorem 1 and the surrounding discussion, this would suggest that the RL-P method should have approximately correct average coverage when the empirical distribution of $\boldsymbol{\theta}$ matches the prior implied by the lasso penalty, a Laplace (double exponential) distribution.

We generated 1000 independent data sets; for each data set, RL-P intervals were constructed from the distributions described in Section 3. Each data set was simulated as follows. The elements of $\mathbf{X}$ were generated independently from a $\mathrm{N}(0,1)$ with $n = 100$, $p = 101$, and $\boldsymbol{\beta}$ was set to the $\frac{1:101}{102}$ quantiles of a Laplace distribution. The coefficients were then scaled so that $\boldsymbol{\beta}^T\boldsymbol{\beta} = \sigma^2$, with independent features this results in a signal-to-noise ratio (SNR) of 1. Finally, $\mathbf{y}$ was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The results are shown in the right-hand side of Figure 1, where the dotted line represents the average coverage across all coefficients, while the solid line is the smoothed estimate of coverage as a function of β. The black line indicates the nominal coverage rate, which is set to be 80%.

The RL-P method has average coverage nearly exactly equal to the nominal 80%. The RL-P has high coverage rates for values of β near zero and lower coverage rates for values of β larger in magnitude. This occurs because for values near zero, the lasso penalty shrinks estimates towards the truth. This leads to a coverage pattern similar to that of Bayesian credible intervals as depicted on the left hand side of Figure 1 and as described in Section 2.

We note that the low coverage for large β values arises from the bias in the lasso penalty and is not inherent to the proposed method. For example, the Minimax Concave Penalty (MCP) exhibits this behavior to a lesser degree (Supplementary Materials). Additionally, the extent of over-coverage for small β values and under-coverage for large β values diminishes as n increases (Supplementary Materials).

4.2 Robustness for Average Coverage

We will now shift our attention to the robustness of the RL-P method under alternative scenarios. We begin by assessing coverage when there is correlation among the predictors. Next, we consider how RL-P performs under various distributions of β. Finally, we look at how the coverage changes across the range of λ values.

4.2.1 Correlation

Figure 3 illustrates the coverage of the RL-P as the level of correlation $\rho = \mathrm{Cor}(\mathbf{x}_i, \mathbf{x}_j)$ for $|i - j| = 1$ increases. Otherwise, the simulation design is the same as in Section 4; in fact, the design is exactly the same for the left panel. The violin plots provide the coverages across 1000 simulated data sets for four values of n. The amount of correlation starts at $\rho = 0$ (no correlation) in the left plot and increases to 0.5 in the middle plot and 0.8 on the right. The coverage behavior remains largely intact for increasing levels of correlation, with the main effect being a slight shift upward in average coverage. The RL-P intervals tend to be over-conservative in the presence of correlation, however, as n increases, average coverage still tends around the nominal coverage rate set.

4.2.2 Distribution of Beta

Given the results in Section 4.1 that the coverage depends on the magnitude of β, one might expect that the average coverage is sensitive to the distribution of $\boldsymbol{\beta}$. Table 1 shows the results of $\boldsymbol{\beta}$ distributed as a Laplace as well as 7 alternative distributions. Otherwise, the setup is the same as described in Section 4. Results are shown for 4 sample sizes, $n = 50$, 100, 400, and 1000. As before, to maintain the specified SNR of 1, $\boldsymbol{\beta}$ is normalized. Prior to normalization, Sparse 1 had $\boldsymbol{\beta}_{1-10} = \pm(0.5, 0.5, 0.5, 1, 2)$ with the rest equal to zero, Sparse 2 had $\boldsymbol{\beta}_{1-31}$ set to 31 evenly distributed quantiles from

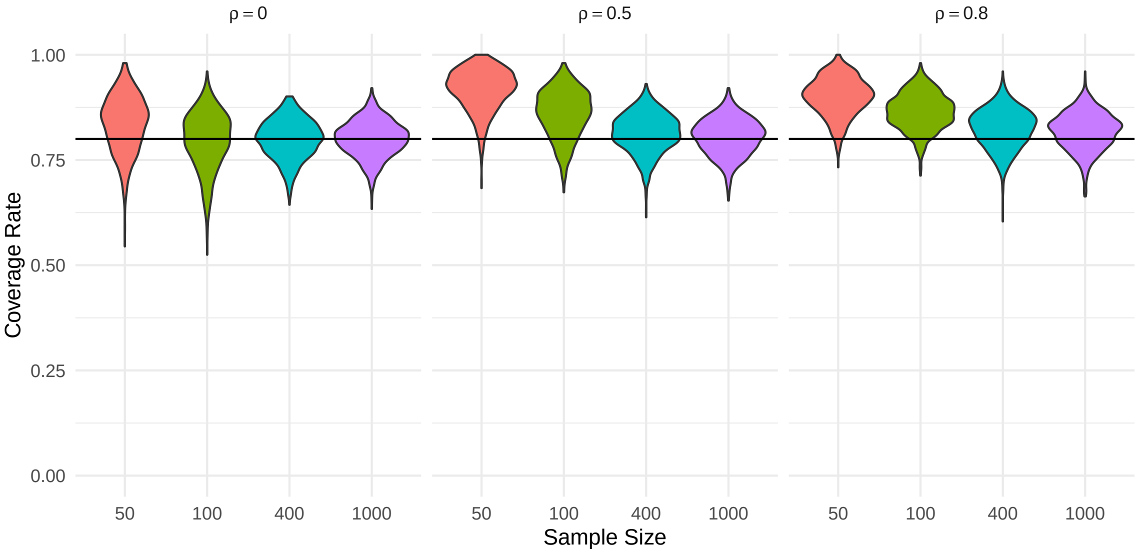


Figure 3: This figure presents results for the simulation described in Section 4.2.1. The violin plots are for the coverage rates across 1000 simulated datasets for the RL-P method and across three different levels of autoregressive correlation among the covariates, $\rho = 0$ (no correlation), $0.5, 0.8$. For this simulation, p = 100, and the results for each level of correlation are presented across four different sample sizes, n = $\frac{1}{2}$p, p, 4p, 10p. The horizontal black line provides reference for the 80% nominal coverage rate.

$N(0,1)$ with the rest equal to zero, and Sparse 3 had $\boldsymbol{\beta}_{1-51}$ set to 51 evenly distributed quantiles from $N(0,1)$ with the rest equal to zero. For the T distribution, df was set to 3 and the Beta distribution quantiles were computed from Beta(0.1, 0.1) - 0.5, prior to normalization. The first column of the table provides a visual depiction of these distributions.

The results shown in Table 1 align with what one might expect from Theorem 1. First, note that under $\boldsymbol{\beta}$ generated from a Laplace, the average coverage of the RL-P method is slightly conservative for $n = 50$ but converges to the nominal rate for the other sample sizes. Distributions that are similar to the Laplace follow this same pattern. For example, when $\boldsymbol{\beta}$ is generated from a T distribution, the coverage rates are nearly identical to the Laplace. When the density / mass is more concentrated near zero, such as with Sparse 1 and 2, the average coverage is above the nominal level. When there is more density away from zero, such as with the normal, the coverage is somewhat below nominal. The worst average coverage occurs when β is generated from a Beta(0.1, 0.1) - 0.5 distribution; this is not surprising since the lasso is a poor choice of penalty in this scenario. Even so, the coverage only drops to 69%, and still converges to the nominal rate as n increases.

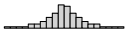
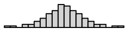
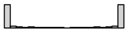
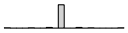
		Sample Size			
	Distribution	50	100	400	1000
	Laplace	83.6%	79.6%	80%	80%
	T	83.3%	78.9%	79.8%	80%
	Normal	81.6%	76.2%	79.7%	79.9%
	Uniform	79.8%	73.2%	79.5%	79.9%
	Beta	78.1%	68.8%	79.3%	79.9%
	Sparse 3	85%	82.6%	81.9%	81.9%
	Sparse 2	87.5%	85.6%	84.5%	84.4%
	Sparse 1	92.6%	91%	90.3%	90.4%

Table 1: Results are from the simulation described in Section 4.2.2. The nominal coverage rate is 80%.

4.2.3 Selection of λ

Throughout the manuscript, λ is set to the value that minimizes CV error; here, we examine how the choice of λ affects coverage. The design remains the same as in Section 4.1 except that for each data set generated, RL-P intervals are obtained for 25 different values of λ. Specifically, λ was evenly distributed on the $\log_{10}$ scale from λ_{max} to $\lambda_{\text{min}} = 0.05\lambda_{\text{max}}$. At each value, confidence intervals were obtained and coverage was recorded. This was repeated 1000 times, then a generalized additive model (GAM) was fit to provide a smooth estimate of the coverage as a function of λ and $|\beta|$. Relative coverage is defined here as the estimated coverage rate minus the nominal coverage rate (red values denote coverage less than nominal, blue values above nominal). The x-axis for λ is presented relative to λ_{max} and the solid black lines delineate the center 95% of λ_{CV} over the 1000 simulations. The dashed black line indicates the median λ_{CV} and the blue line represents the value of λ which provided coverage closest to that of nominal.

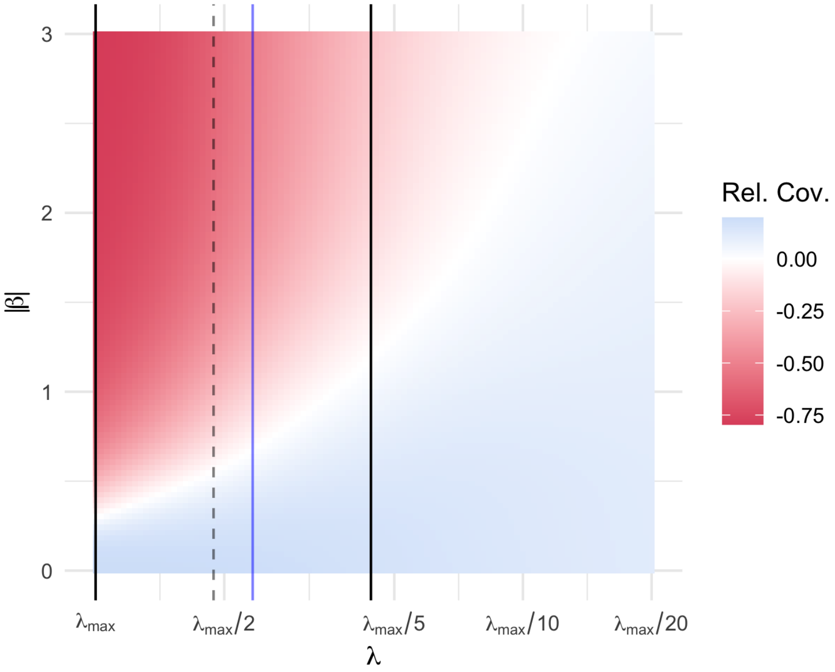


Figure 4: The heatmap displays relative coverage for RL-P across a range of λs per the simulation described in Section 4.2.3. A Binomial GAM was used to estimate coverage as a smooth function of the $|\beta|$ and λ. The x-axis for λ is presented relative to λ_{max} and the solid black lines indicate the center 95% of λ_{CV}s over the 1000 simulations. The dashed black line indicates the median λ_{CV} and the blue line represents the value of λ which provided coverage closest to that of nominal.

To obtain average coverage near nominal, λ should be chosen such that the over-coverage for small $|\beta|$ values is balanced by the under-coverage for large $|\beta|$ values. The blue line in Figure 4 represents the λ value for which this balance is best achieved. In this scenario, and in general, λ_{CV} does a reasonable job at achieving this balance: sometimes below the "perfect balance" line, sometimes above, but usually in reasonable agreement.

However, clearly the value of λ does matter, again supporting the idea presented in Theorem 1. This simulation is similar to when $\boldsymbol{\beta}$ is distributed as alternative distributions, but here, instead of altering the data generating mechanism, we are adjusting the prior implied by the lasso penalty. When this implied prior is reasonably close to the data generating mechanism, coverage is near nominal. When the implied prior is more concentrated at zero (e.g. λ near λ_{max}) or more diffuse (e.g. λ near λ_{min}), then coverage is below and above nominal, respectively.

4.3 Effect of Correlation on Individual Intervals

Figure 3 illustrates that the average coverage of the proposed RL-P method is robust to increasing correlation. However, this doesn't mean that the intervals themselves are unaffected by correlation. In this section, we illustrate the effect of correlation between features on the intervals themselves and contrast the intervals produced by lasso with those produced by ridge regression.

In this example, we have $n = p = 100$. However, only one β_j is non-zero: $\beta_A = 1$ and $\beta_B, \beta_{N1}, \ldots, \beta_{N98} = 0$. Additionally, the data are simulated such that $\text{Cor}(\mathbf{x}_A, \mathbf{x}_B) = .99$ but all of the N (noise) variables are uncorrelated with A, B, and each other. The distribution of $\mathbf{X}$ and $\mathbf{y}$ is unchanged from Section 4.1, although here $\sigma^2 = 1$.

Figure 5 depicts the results from $1,000$ simulated data sets. On the left, 1000 CIs are shown for for 3 features: A, B, and $N1$; the CIs are colored black if they contain the true parameter value and red if they do not. On the right, confidence intervals for the first 20 variables for a randomly selected example data set are displayed.

Although A is the only feature with a true signal, its high correlation with B produces a large amount of uncertainty about which feature contains the signal, or whether both A and B contain signal. Ridge regression makes a fairly strong assumption here that the signal being divided equally between A and B is much more likely than either A or B having all the signal. This results in intervals that are very similar for A and B. As a result, the correlation between A and B does not introduce much uncertainty – the confidence intervals for β_A and β_B are no wider than that of the noise features. With the Lasso on the other hand, we apply an equal penalty to A having all the signal, B having all the signal, and the signal being shared between A and B. As a result, Lasso estimates are very sensitive to correlation and, accordingly, the RL-P CIs for A and B are often much wider than those for the noise features. Noticeably, the RL-P intervals for A tend to be larger than those for B, indicating that even with very high correlation, the Lasso typically attributes more of the signal to the causal feature A; this is not the case with the Ridge penalty. It is important to note that the width of the RL-P intervals tend to either be very narrow or very wide. Looking back at the construction explains why, the variance of the conditional distribution is largely determined by how much information in $\mathbf{x}_j$ is orthogonal to $\mathbf{X}_{\hat{S}_j}$. When variable A is selected, the interval for variable B will be wide and vice versa.

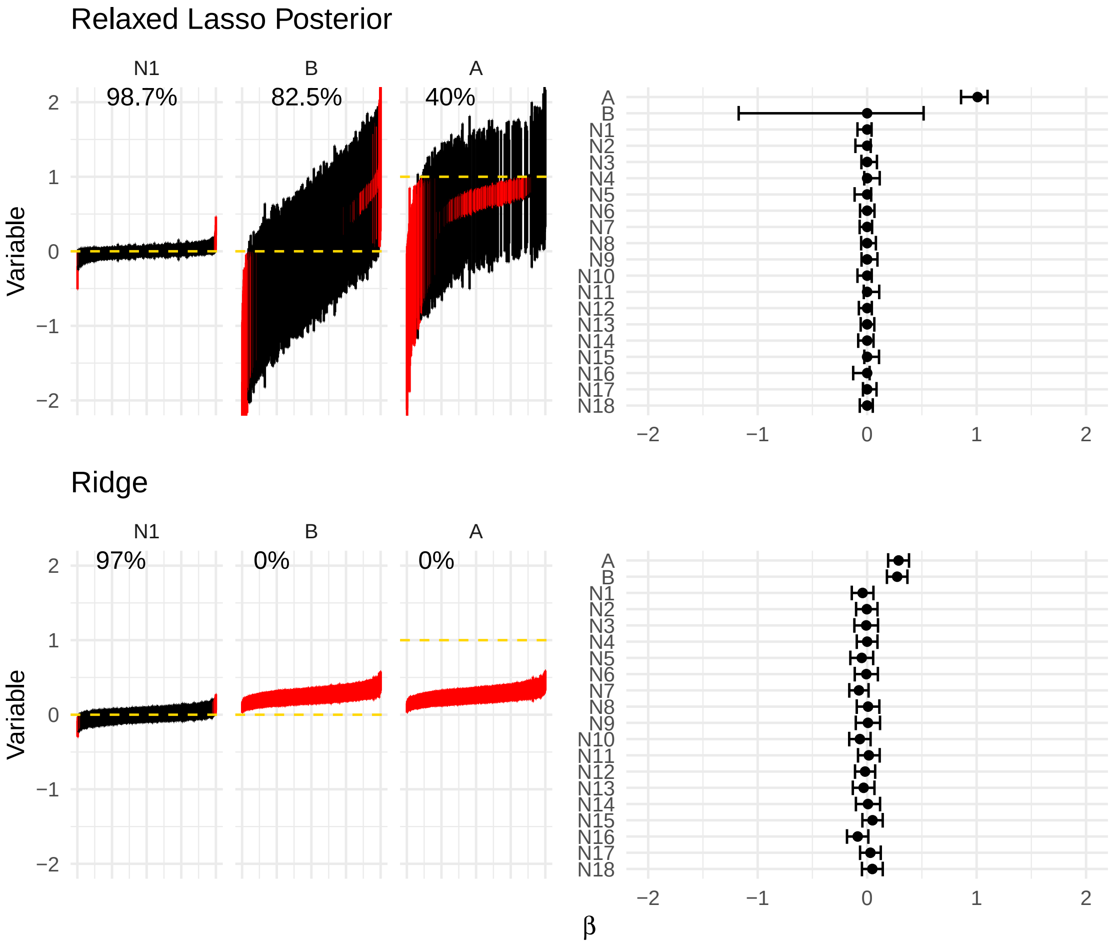


Figure 5: Provides results for simulation described in Section 4.3. The right plots show a single example of a intervals produced by Ridge (top) and RL-P (bottom) from one (randomly selected) of the 1000 datasets for the first 20 variables. The left plot summarizes the resulting CIs for the variables A, B, and $N1$ across the 1000 simulations. All 1000 CIs are plotted, sorted by their midpoint, with those colored red that did not contain contain the true coefficient value (indicated by the horizontal dashed gold line).

4.4 Comparison to other methods

As noted in the introduction, there are few methods for obtaining intervals for the lasso that have been developed and implemented with available software. Two that we were able to identify were Selective Inference (implemented in the `selectiveInference` R package) and the de-sparsified lasso (implemented in the `hdi` R package).

Selective Inference, de-sparsified lasso, and RL-P are based on different principles and operate in fundamentally distinct ways. To review, the de-sparsified lasso (Zhang and Zhang, 2014), as the name suggests, provides a method to debias the original point estimates from a lasso fit to facilitate classical approaches to inference. Note that this process of debiasing changes the underlying model, a point we return to in Section 6. Alternatively, Selective Inference (Lee et al., 2016; Tibshirani et al., 2016) aims to account for the uncertainty in model selection by conditioning on the selected model. This conditioning is also a form of bias correction, although it accomplishes this in a less direct way than de-sparsified lasso, which adds a term to explicitly counteract bias. Note that through conditioning on the selected model, Selective Inference only directly provides intervals for the covariates that were selected.

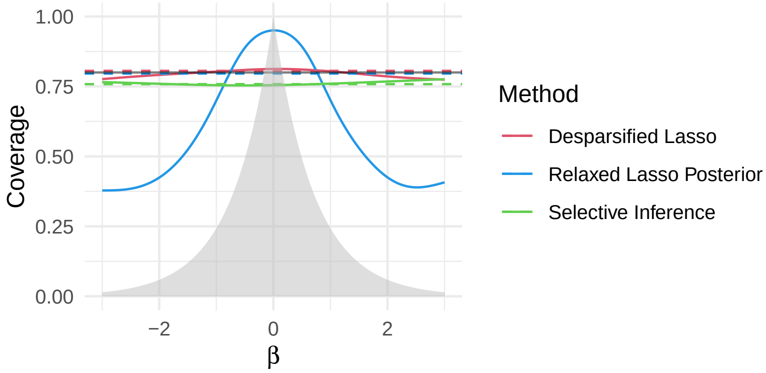


Figure 6: Results are from the simulation described in Section 4.4 and identical in setup to that of Section 4.1. The fitted curves are from Binomial GAMs fit with coverage being modeled as a smooth function of β. The dashed lines represent the average coverage for each method across all 1000 independently generated datasets and the solid black line indicates the nominal coverage rate. The shaded distribution in the background depicts the Laplace distribution the βs were drawn from.

We conducted a simulation study to compare these three methods; the setup is identical to that described in Section 4.1. For each software package, their default options were used. Notably, this means that for de-sparsified lasso's implementation in the `hdi` package, λ is set using the 1SE rule from cross-validation, whereas for Selective Inference and RL-P, λ was set at the value which minimizes CV error.

Selective Inference and de-sparsified lasso adopt a more classical perspective than RL-P, which is evident in Figure 6. While all three methods have reasonable average coverage, they achieve this in different ways. As demonstrated in Section 2.1, methods for constructing intervals can either achieve consistent coverage across all values of the target parameter, or they can reflect the shrinkage imposed by the penalty – they cannot achieve both. Intervals which reflect the shrinkage imposed by the penalty result in uneven coverage across β. As shown in Figure 6, Selective Inference and de-sparsified lasso provide the first kind of interval. Either directly or indirectly, the shrinkage imposed by the lasso has been undone by intervals they provide and the result is flat coverage across values of β. This is unlike RL-P, which reflects the shrinkage of the lasso and results in higher coverage where the prior density (implied by the penalty) is higher.

Figure 7 illustrates how the coverage, interval width, and computational burden of these methods compares. The top panel shows the distribution of average coverage across all 1000 simulations. The de-sparsified lasso has coverage centered around the nominal 80% coverage. Meanwhile, the distribution of average coverage for Selective Inference is very wide, spanning 0% to 100%. For $n = 50$ and $n = 100$, although centered around nominal coverage, average coverage was often well above or well below the nominal rate (and indeed was sometimes near zero). The average coverage is less variable at $n = 400$, although it is consistently above the nominal rate but still with a noticeable tail trailing down to 0%. The behavior for RL-P has been covered previously, specifically that it is slightly over-conservative when $n < p$ but converges to nominal coverage as n increases above p.

The middle plot provides the median CI width across all covariates from all 1000 simulations. The de-sparsified lasso does tend to produce wider intervals, especially when $p \leq n$ relative to RL-P. Selective Inference, on the other hand, produces even wider intervals. In fact, the vertical limits of the panel had to be truncated – to capture the full bar for Selective Inference when n = 50, the limits would need to be expanded to 150.

Selective Inference differs from de-sparsified lasso and RL-P in that it does not provide intervals for all parameters, only

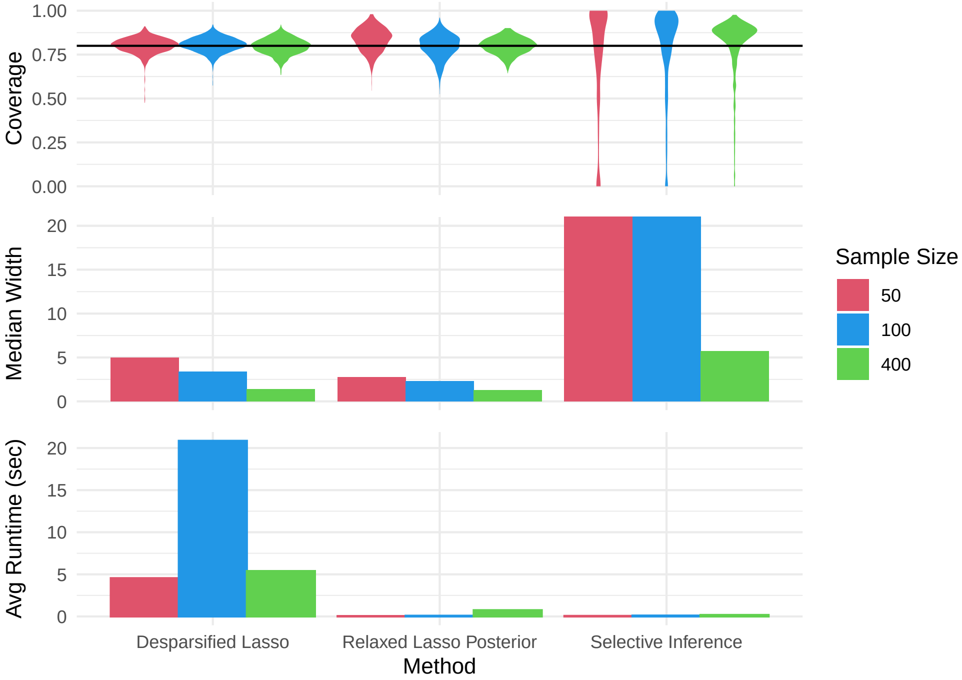


Figure 7: Results are from the simulation described in Section 4.4. Each plot provides corresponding results for each of de-sparsified lasso, RL-P, and Selective Inference for three different sample sizes. The top provides violin plots of average coverages, the middle is a bar plot of the the median CI widths, and the bottom is a bar plot of the average run times, across all 1000 simulated datasets. The y limits have been truncated for the median width from 150 to 20.

the subset of parameters with nonzero coefficients. And even when Selective Inference does construct intervals, they are often infinitely wide (we will see this again for the real data in Section 5). More information on how often these two issues arise in this simulation is found in Supplementary Materials. Additionally, Kivaranovic and Leeb (2021) provide an in-depth discussion of the widths of CIs produced by methods like Selective Inference that use the polyhedral approach and show that the expected value of interval width is infinite.

The bottom panel of Figure 7 provides the average run times for each of the methods. The runtime varied considerably between the methods, with Selective Inference the fastest and de-sparsified lasso by far the slowest. The only noticeable difference in speeds between RL-P and Selective Inference was that Selective Inference scales better with n. That said, in our testing, speed was not a concern for Selective Inference or RL-P, but was prohibitive for de-sparsified lasso. Although not shown in the figure, de-sparsified lasso also scales quite poorly with p, as we will see in Section 5.2.

5 Real Data Analysis

In this section, we apply the RL-P method to two real datasets: a study of acute respiratory illness conducted by the World Health Organization (WHO-ARI) and a study of gene expression in the mammalian eye (Scheetz et al., 2006). These two datasets sit on opposite ends of the spectrum in terms of dimensionality. WHO-ARI contains 816 observations and 66 features while Scheetz2006 contains just 120 observations but with 18975 features.

In this section, we also consider the intervals produced by de-sparsified lasso and Selective Inference, comparing the intervals both to each other and to the point estimates provided by the lasso at λ_{CV}.

5.1 World Health Organization study on acute respiratory illnesses (WHO-ARI)

The WHO-ARI study considered a few acute illness in young infants across several countries, and the dataset used here is a subset of 816 infants who presented with pneumonia in the country Ethiopia, which represents the main cause of morbidity and mortality for infants under 3 months of age (Harrell et al., 1998). Our goal here is to identify risk factors for increased severity among infants presenting with serious infections. The outcome is ordinal (taking on a number from 1 - 5); however, for simplicity we treat the outcome as following a Gaussian distribution. The variables collected contain information on vital signs, family history, and clinical observations and represent a range of data types from binary to ordinal to continuous. With $N \approx 10p$, this dataset is not necessarily high dimensional. However, sparsity is beneficial both for interpretation and for the practical implementation of using the lasso estimates in clinical practice.

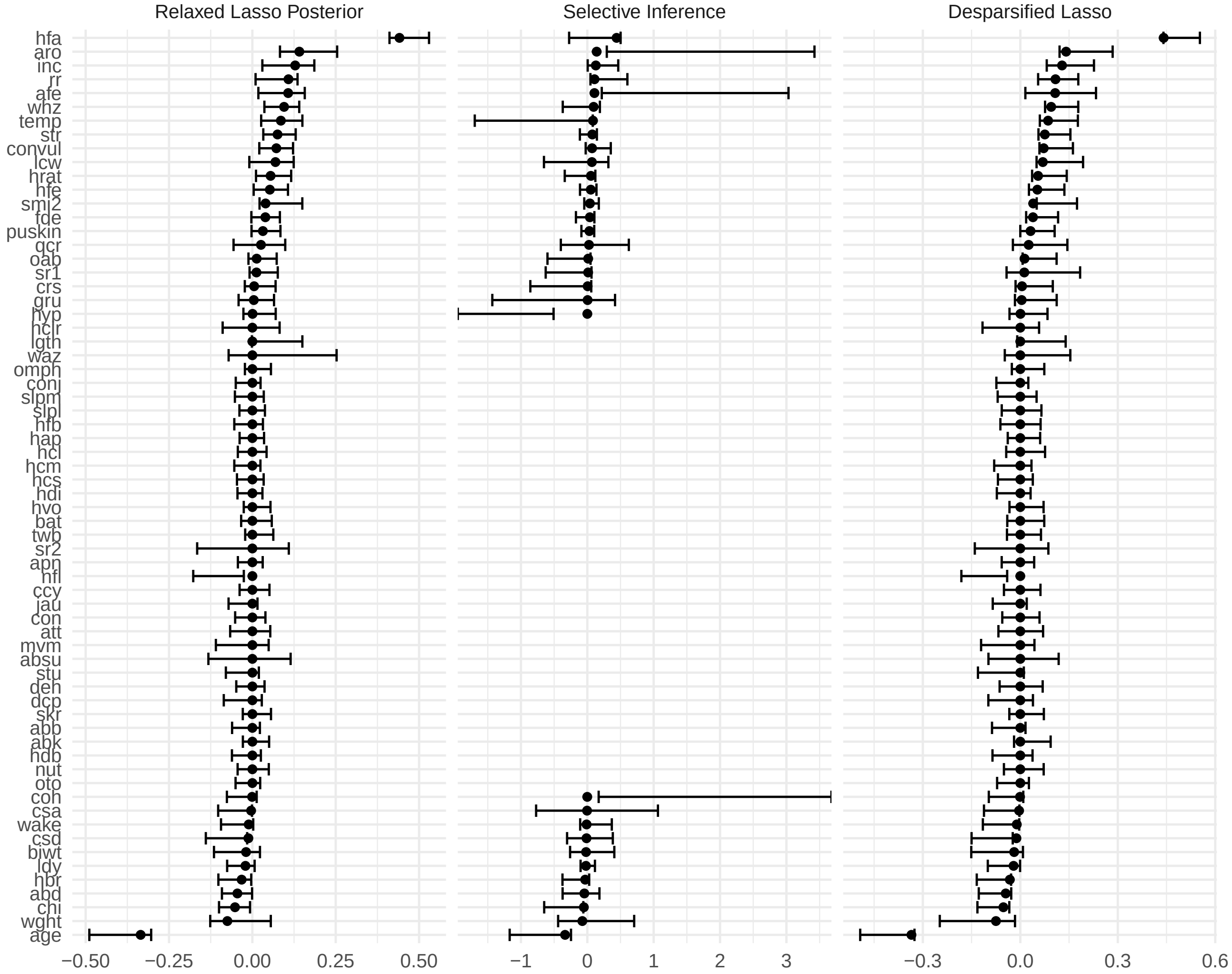


Figure 8: Confidence intervals produced by three different methods for all 66 variables in the WHO-ARI dataset described in Section 5.1.

Figure 8 provides the confidence intervals from each of the three methods along with corresponding point estimates from the lasso. The intervals are provided on the standardized scale to aid in visualization. It is important to emphasize that the range of the x-axis is different for each of the plots corresponding to the three methods. RL-P and de-sparsified lasso share similar patterns, although the intervals from RL-P are generally narrower. Additionally, while RL-P intervals, relative to the point estimates, tend to be more symmetric, de-sparsified lasso's intervals are more often skewed away from zero as a result of debiasing. As mentioned earlier, Selective Inference does not produce an interval for every parameter, only for the 32 (out of 66) features that were selected. Furthermore, of these 32, two intervals are infinitely wide and several others are much wider than any intervals produced by either de-sparsified lasso or RL-P. Altogether, de-sparsified lasso produces 25 intervals that do not contain zero, RL-P produces 19 that do not contain zero, while only 8 of the 32 Selective Inference intervals do not contain zero.

5.2 Gene expression in the mammalian eye (Scheetz2006)

Scheetz et al. (2006) measured the RNA levels from the eyes of 120 rats. Of 31000 different probes used, 18976 were detected at a sufficient level to be considered "expressed." For this analysis we treat one of the genes, Trim32, as the outcome since it is known to be linked to the genetic disorder Bardet-Biedl Syndrome (BBS). The remaining 18975 genes are used as covariates with the goal of determining other genes whose expression is associated with Trim32 and thus may also contribute to BBS.

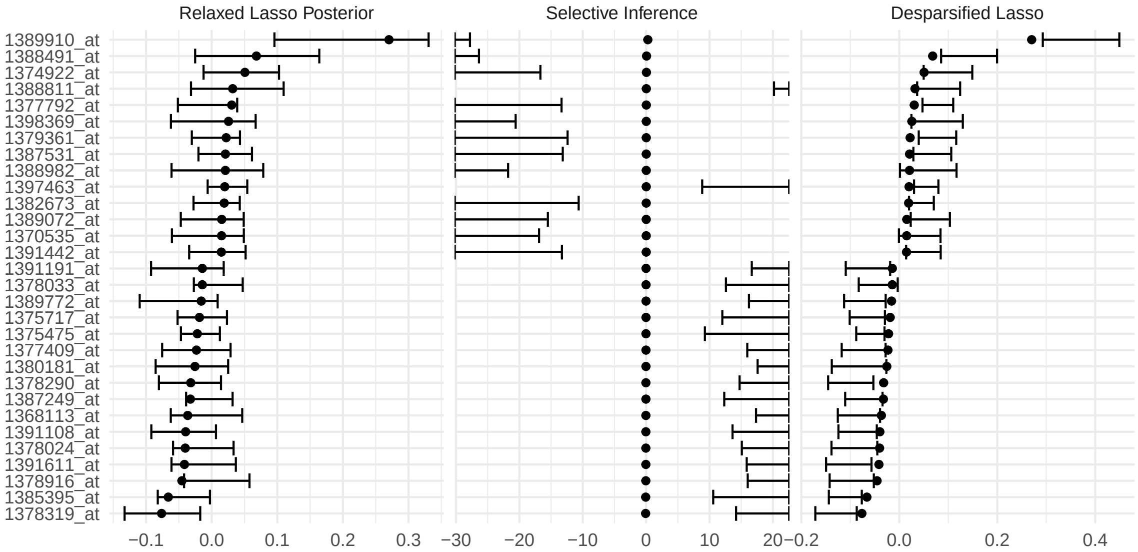


Figure 9: Confidence intervals produced by three different methods for the 30 variables with the largest absolute point estimates in the Scheetz2006 dataset described in Section 5.2.

Compared to the WHO-ARI data, the increased dimensionality here leads to more pronounced differences between RL-P and de-sparsified lasso (Figure 9). The intervals of RL-P are more shrunken towards zero and the intervals of de-sparsified lasso are more pushed away from zero. Additionally, while 71 of the RL-P intervals do not contain their respective point estimates, this occurs for 981 intervals produced by the de-sparsified lasso. In addition, there is a large discrepancy for the number of significant intervals (intervals not containing zero) between the two methods. De-sparsified lasso produces 989 intervals which exclude zero, while RL-P produces 77. Selective Inference provides intervals for 66 of the 18975 features. For this high-dimensional data ($p > 100n$), however, every single one of them has a lower or upper bound that is infinite. Additionally, none of the Selective Inference intervals contain zero, and in fact, have no overlap with any of the de-sparsified lasso or RL-P intervals (note again that the x-axis is different for each of the three methods). Even more peculiar, of the 66 intervals created by Selective Inference, 62 of them were completely of the opposite sign as the corresponding lasso estimate.

Lastly, it is important to note that de-sparsified lasso took over 6 hours to produce confidence intervals for the Sheetz2006 dataset on a MacBook Pro with 16 GB of RAM and an Apple M1 Pro chip. This is a result of de-sparsified lasso's computational cost growing with p. In comparison, RL-P took 1.2 seconds while Selective Inference took about three tenths of a second. With respect to computational burden, de-sparsified lasso is feasible for small to moderately sized datasets, but does not scale for large p.

6 Discussion

Should intervals be biased? Over the past several decades, statisticians have grown more comfortable with the idea of biased estimators. Nevertheless, the statistics community still seems uncomfortable with biased intervals. However, if you have chosen to use a biased estimation method, it would seem reasonable that the resulting intervals should reflect that bias. This is in conflict with classical frequentist ideas of confidence intervals, but as we have shown, agrees with Bayesian posterior intervals.

One objection to having biased intervals is that it results in under-coverage for large values of β – the parameters that are typically of greatest interest. The same objection, however, applies to the lasso estimates themselves. There are many alternatives to the lasso, including the adaptive lasso, MCP, and SCAD , which reduce the bias imposed by the lasso for large values of β (Zou, 2006; Zhang, 2010; Fan and Li, 2001). Using RL-P with any of these alternative approaches results in less biased intervals (Supplementary Materials).

In contrast, most of the literature on high-dimensional intervals focuses on constructions that are debiased in some way. It is debatable, however, whether these intervals still reflect the assumptions that went into the original lasso estimates. This does not mean that the intervals produced by approaches such as de-sparsified lasso and Selective Inference are incorrect. However, neither de-sparsified lasso nor Selective Inference preserves the assumptions of the original lasso estimates; this is seen most clearly in Section 5.2.

At a fundamental level, it is not possible to shrink point estimates towards zero while not also shrinking intervals towards zero. Attempting to accomplish both will lead to inconsistencies. For an analyst attempting to "pair" the lasso with desparsified lasso or Selective Inference, it is critical to recognize that the underlying assumptions for the point estimates and for the intervals are not the same. Presenting intervals and point estimates that do not agree is unsatisfying.

Rather than try to debias or otherwise correct for the lasso penalty when constructing intervals, we develop here the Relaxed Lasso Posterior and show that it offers a more coherent approach where the intervals reflect the lasso point estimates. Adopting these intervals requires a change in perspective, where the emphasis of the intervals is average coverage across the set of parameters as opposed to individual parameter coverage. If nothing else, we hope that this article raises interesting questions about which perspective is preferable and the extent to which single-parameter inferential ideas are appropriate for high-dimensional inference.

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, **43** 2055–2085.

Breheny, P. J. (2019). Marginal false discovery rates for penalized regression models. *Biostatistics*, **20** 299–314.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 551–577.

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

Clarté, L., Vandenbroucque, A., Dalle, G., Loureiro, B., Krzakala, F. and Zdeborová, L. (2024). Analysis of bootstrap and subsampling in high-dimensional regularized regression.

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST*, **26** 685–719.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, vol. 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

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.

Harrell, F., Margolis, P., Gove, S., Mason, K., Mulholland, E., Lehmann, D., Muhe, L., Gatchalian, S. and Eichenwald, H. (1998). Development of a clinical prediction model for an ordinal outcome: the world health organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants. *Statistics in Medicine*, **17** 909–944.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics, Springer.

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

Karoui, N. E. and Purdom, E. (2016). Can we trust the bootstrap in high-dimension?

Kivaranovic, D. and Leeb, H. (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, **116** 845–857.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, **44** 907–927.

Lo, A. Y. (1987). A large sample study of the bayesian bootstrap. *The Annals of Statistics*, **15** 360–375.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, **42** 413–468.

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

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

Reid, S., Tibshirani, R. and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, **26** 35–67.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, **9** 130 – 134.

Scheetz, T., Kim, K.-Y., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., Sheffield, V. and Stone, E. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, **103** 14429–14434.

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

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, **111** 600–620.

Xing, X., Zhao, Z. and Liu, J. S. (2023). Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, **118** 222–241.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38** 894–942.

Zhang, C. H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76** 217–242.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.

7 Appendices: additional results

7.1 Traditional Bootstrap Example

Figure 10 shows the coverage as a function of the value of β (solid line) and the average coverage (dashed line) using a traditional pairs bootstrapping approach on the simulation setup described in Section 4.1. Compare to the right side of Figure 1.

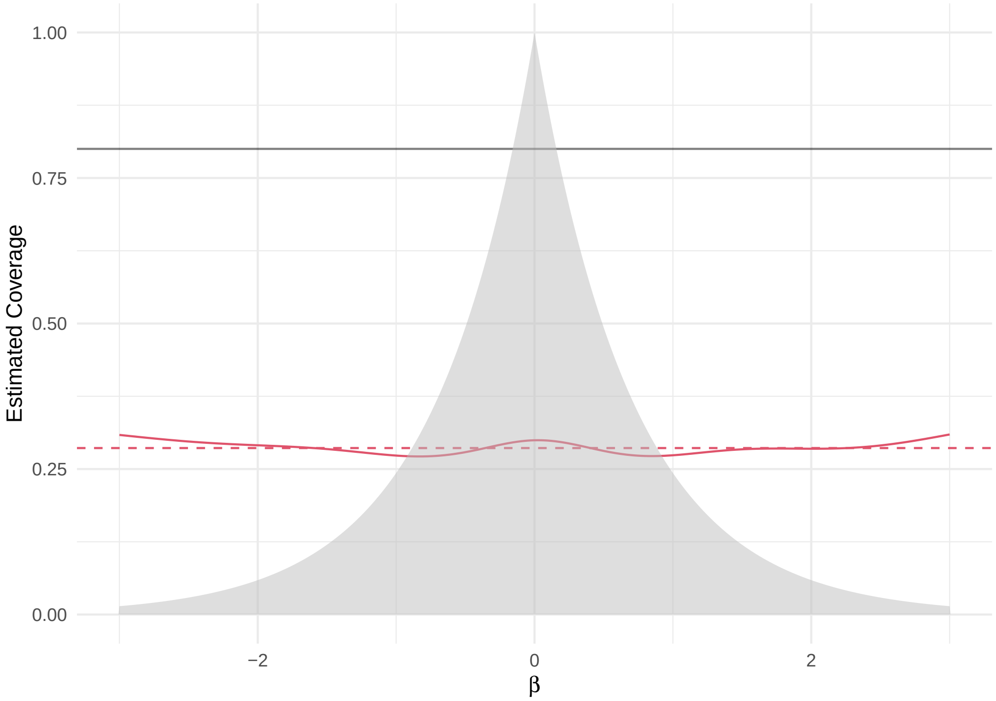


Figure 10: Corresponds to the setup used in the right side of Figure 1, but using a traditional bootstrapping approach.

7.2 Sampling from the Full Conditional Posterior

Here we assume $\mathbf{X}$ has been standardized s.t. $\mathbf{x}_j^T\mathbf{x}_j = n$. Define $\mathbf{Q}_{\hat{S}_j}$ as $\mathbf{I} - \mathbf{X}_{\hat{S}_j}(\mathbf{X}_{\hat{S}_j}^T\mathbf{X}_{\hat{S}_j})^{-1}\mathbf{X}_{\hat{S}_j}^T$, the projection matrix onto the features selected by the lasso. For β_j conditional on the selected features

$$L(\beta_j|\hat{S}_j) \propto \exp(-\frac{\tilde{n}}{2\sigma^2}(\beta_j^2 - 2\tilde{\beta}_j\beta_j)),$$

where $\tilde{\beta}_j = (\mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{x}_j)^{-1}\mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{y}$ and $\tilde{n} = \mathbf{x}_j^T\mathbf{Q}_{\hat{S}_j}\mathbf{x}_j$.

The lasso can be formulated as a Bayesian regression model with a laplace (double exponential) prior. In this case, the prior for β_j is proportional to $\exp(-\frac{\tilde{n}\lambda}{\sigma^2}|\beta_j|)$. This prior ensures that the meaning of λ is maintained.

With this the form of the full conditional posterior can be worked out as follows:

$$\begin{aligned}
p(\beta_j|\hat{S}_j) &\propto \exp(-\frac{\tilde{n}}{2\sigma^2}(\beta_j^2 - 2\tilde{\beta}_j\beta_j))\exp(-\frac{\tilde{n}\lambda}{\sigma^2}|\beta_j|) \\
&= \exp(-\frac{\tilde{n}}{2\sigma^2}(\beta_j^2 - 2\tilde{\beta}_j\beta_j + 2\lambda|\beta_j|)) \\
&= \exp(-\frac{\tilde{n}}{2\sigma^2}(\beta_j^2 - 2(\tilde{\beta}_j\beta_j - \lambda|\beta_j|))) \\
&= \begin{cases} \exp(-\frac{\tilde{n}}{2\sigma^2}(\beta_j^2 - 2(\tilde{\beta}_j + \lambda)\beta_j)), & \text{if } \beta_j < 0, \\ \exp(-\frac{\tilde{n}}{2\sigma^2}(\beta_j^2 - 2(\tilde{\beta}_j - \lambda)\beta_j)), & \text{if } \beta_j \geq 0 \end{cases} \\
&\propto \begin{cases} C_-\exp\{-\frac{\tilde{n}}{2\sigma^2}(\beta_j - (\tilde{\beta}_j + \lambda))^2\}, & \text{if } \beta_j < 0, \\ C_+\exp\{-\frac{\tilde{n}}{2\sigma^2}(\beta_j - (\tilde{\beta}_j - \lambda))^2\}, & \text{if } \beta_j \geq 0 \end{cases}
\end{aligned}$$

where $C_- = \exp(\tilde{\beta}_j\lambda\tilde{n}/\sigma^2)$ and $C_+ = \exp(-\tilde{\beta}_j\lambda\tilde{n}/\sigma^2)$.

Note the piecewise defined posterior is made up of a kernel of two normal distributions. This can be leveraged and draws can be efficiently obtained through a mapping onto the respective normal distributions. To define this mapping, it helps to introduce a concept and some notation. First, the use of "tails" here refers to the entirety of a distribution between 0 and $\pm\infty$. That is, the lower tail is any part of the distribution below 0 and the upper tail is any part greater than 0, therefore $P(X \in \textit{lower} \cup X \in \textit{upper}) = 1$. Accordingly, we will let the tail probabilities in each of the two normals to transformed on to be denoted Pr_- and Pr_+ respectively and the probability in each of the tails of the posterior, denoted $Post_-$ and $Post_+$ respectively. $Pr_\pm$ is trivial to compute with any statistical software. $Post_\pm$ is conceptually simple, although care must be taken to avoid numerical instability. With this notation in place, note that,

$$p(\beta_j|\hat{S}_j) \propto \begin{cases} C_-Pr_-, & \text{if } \beta_j < 0, \\ C_+Pr_+, & \text{if } \beta_j \geq 0 \end{cases}$$

which implies that $Post_- = \frac{C_-Pr_-}{C_-Pr_- + C_+Pr_+}$ and similarly for $Post_+$. To avoid numerical instability, or rather to handle it properly when it is unavoidable, we will work on the log scale. This works well for most of the problem, but computation of $Post_-$ and $Post_+$ need something a bit more since, for example, $\log(Post_-) = \log(C_-Pr_-) - \log(C_-Pr_- + C_+Pr_+)$. That is, the denominator still must be computed then the log taken which does not allow operating on the log scale to fully address potential numerical instability. Instead, let $\ell_- := \log C_- + \log Pr_-$, $\ell_+ := \log C_+ + \log Pr_+$, and $\Delta := \ell_- - \ell_+$, then $\log(Post_-)$ can be computed with $\Delta - \log\big(1 + \exp(\Delta)\big)$. This still doesn't completely address the issue, however, if $\exp(\Delta)$ is infinite then $C_-Pr_- >> C_+Pr_+$ and $\log(Post_-) \approx 0$ which means $Post_- \approx 1$ (equivalently $Post_+ \approx 0$).

With these values, we can compute the quantile by mapping the corresponding probability p for the posterior onto the probability p^* for the corresponding normals. Which normal the quantile of interest ultimately comes from is determined based on $Post_\pm$. If $p \leq Post_-$, then p would be mapped onto the negative normal. If $p > Post_-$, then p would be mapped onto the positive normal. For example, if $Post_+ = 0.98$ and $p = 0.1$ then p would be mapped onto the positive normal. The transformation to map a given probability from the posterior depends on which tail the quantile resides in on the posterior (equivalently which normal it is being mapped to, the positive or negative). This map is simply:

$$p^* = p \times (Pr_\pm/Post_\pm)$$

Once the respective probability is mapped, one can simply use the inverses of the normal CDF that the probability was mapped to. That being said, there is a nuance worth pointing out. When transforming the probabilities, the step to determine which tail the respective quantile comes from occurs first. With this, the probability should be adjusted so that it refers to the probability between the quantile of interest and the respective tail. After this, then the transformation can be applied. These steps are summarized in Algorithm 1.

Algorithm 1 Compute Quantile for RL-P Laplace-Normal Distribution

```
Require: $\tilde{\beta}_j,\ \sigma^2,\ \tilde{n};\lambda,\ p$    ▷ mean, variance, sample size, penalty, target significance level

// 1. Compute prior mass for negative and positive regions (on log-scale)
$Pr_- \leftarrow \Phi\big(0;\ \tilde{\beta}_j + \lambda,\ \frac{\sigma^2}{\tilde{n}}\big)$
$Pr_+ \leftarrow 1 - \Phi\big(0;\ \tilde{\beta}_j - \lambda,\ \frac{\sigma^2}{\tilde{n}}\big)$

// 2. Compute posterior weights $Post_-$, $Post_+$ with log-scale stabilization
$\ell_- \leftarrow \log C_- + \log Pr_-$
$\ell_+ \leftarrow \log C_+ + \log Pr_+$
$\Delta \leftarrow \ell_- - \ell_+$
if $\exp(\Delta) = \infty$ then
    $\log Post_- \leftarrow 0$    ▷ since $C_- Pr_- \gg C_+ Pr_+$
else
    $\log Post_- \leftarrow \Delta - \log\big(1 + \exp(\Delta)\big)$
end if
$Post_- \leftarrow \exp(\log Post_-)$
$Post_+ \leftarrow 1 - Post_-$

// 3. Invert CDF on the appropriate component
if $p \leq Post_-$ then
    $w \leftarrow \dfrac{Pr_-}{Post_-}$
    $q \leftarrow \Phi^{-1}\big(p\,w;\ \tilde{\beta}_j + \lambda,\ \frac{\sigma^2}{n}\big)$
else
    $w \leftarrow \dfrac{Pr_+}{Post_+}$
    $q \leftarrow \Phi^{-1}\Big(1 - (1-p)\,w;\ \tilde{\beta}_j - \lambda,\ \frac{\sigma^2}{n}\Big)$
end if
return $q$
```

7.3 Coverage Behavior Under Alternative Sample Sizes

Figure 11 displays coverage estimates as a smooth function of β for three values of n: 50, 100, and 400 but otherwise uses the same setup as the simulation described in Section 4.1.

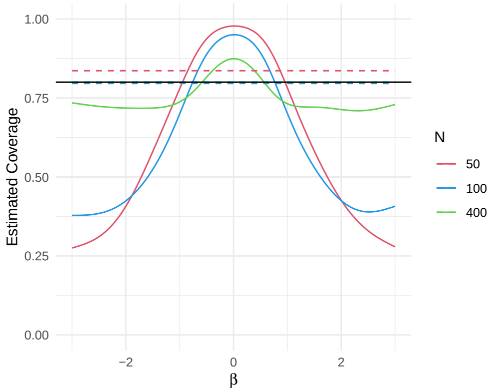


Figure 11: The results displayed are from a simulation with the same set up as in Section 4.1 but with n set to three different values: 50, 100, 400. The fitted curves are from Binomial GAMs fit with coverage being modeled as a smooth function of β. The dashed lines represent the average coverages across all 1000 independently generated datasets and the solid black line indicates the 80% nominal coverage rate.

For $n = 50$, coverage is overconservative with the characteristic high coverage for values near zero and low coverage for larger values of β. However, As n is increased to 100 then 400 the characteristic over coverage for small β values and under coverage for large β values lessens, largely attributable to λ_{CV} being smaller.

7.4 Selective Inference Intervals

n	# Simulations Null Selected	Average # Parameters	# Simulations Inf Median Width	# Simulations Any Inf Width
50	239	13.0	127	349
100	69	28.5	125	571
400	0	73.4	39	626

Table 2: Additional information on the results for Selective Inference in the simulation described in Section 4.4.

Because Selective Inference only provides intervals for the subset of parameters with nonzero coefficients, the results in Figures 6 and 7 are just for this subset. This is in contrast to the results for de-sparsified lasso and RL-P which are over all parameters. The first two columns of Table 2 provide more details on the size of the lasso models selected by cross validation, and hence, on the number of intervals constructed. The first column indicates how many simulations (out of 1000) the intercept-only model was selected, in which case no intervals are produced. The second column gives the average number of selected parameters (inclusive of when none were selected). It is the subset of parameters represented by column 2 that are the results in Figures 6 and 7 are from.

For the intervals that were constructed, the third and fourth columns provide the number of simulations that had a infinite median width or any interval with an infinite width, respectively, features that were not evident in Figure 7. Note that simulations where the null model was chosen by definition can not have intervals of infinite width.

7.5 MCP Example

	Coverage (%)	
$\lvert\beta\rvert$	Relaxed Lasso Posterior	Relaxed MCP Posterior
0.00	96.25%	98.21%
1.47	48.92%	61.20%
2.95	36.20%	48.60%
5.90	32.55%	68.70%

Table 3: Coverage rates by magnitude of β for RL-P using both the lasso and the MCP penalty approximation applied to the Sparse 1 scenario described in Section 4.2.2. The nominal coverage rate is 80%.

The results here provide an example of a version of the RL-P method but with an approximation to the MCP penalty to obtain intervals. The MCP penalty closely resembles that of the lasso near zero but eventually levels out to a constant penalty for larger values of β unlike the lasso which applies a penalty proportional to the magnitude of β. Note that although βs with intermediate magnitudes are still under covered (albeit with coverage notably higher than RL-P lasso), that the largest βs have coverage at nearly 70%, over doubling the coverage of the RL-P lasso.

7.6 Additional Details on Bootstrap Bias

We start with a simple proof of the additional bias introduced by bootstrapping in one and two dimensional settings starting first with Ridge regression and then the lasso. After these simple proofs, we provide a simulation that helps generalize this issue to high dimensional settings. Note that here the goal is to provide an intuitive understanding of the issue. Others have already proved this issue more generally (Karoui and Purdom, 2016; Clarté et al., 2024), however, we feel that the lack of straight forward examples may be in part why the bootstrap is still a tool used for penalized regression in general.

Let $\frac{1}{n}x_1^T x_1 = s_{11}$, and s_{11}^* be the bootstrapped version of s_{11}. Furthermore more that $E_{boot}[s_{11}^*] = s_{11}$.

For Ridge, it can be shown that $E(\widehat{\beta}_1) = \frac{s_{11}}{s_{11}+\lambda}\beta_1$, letting $g(s_{11}) = \frac{s_{11}}{s_{11}+\lambda}$ note that:

$$\begin{aligned} g'(s_{11}) &= \lambda(s_{11}+\lambda)^{-2} \\ g''(s_{11}) &= -2\lambda(s_{11}+\lambda)^{-3} < 0 \end{aligned}$$

Thus by Jensen's inequality we have,

$$E_{boot}[g(s_{11}^*)] \leq g(E_{boooot}[s_{11}^*]) = g(s_1 1).$$

Multiplying by β_1 then gives,

$$E_{boot}[\widehat{\beta}_1^*] = E_{boot}[g(s_{11}^*)]\beta_1 \leq g(s_{11})\beta_1 = E[\widehat{\beta}_1].$$

That is, even in a single parameter setting, just due to the bootstrap variability of s_{11} alone, we would expect bias in the bootstrapped estimate of β_1.

Now consider a two parameter setting where we assume $\beta_2 = 0$. Let $V = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} s_{11} & s \\ s & s_{22} \end{pmatrix}$.

Similarly here, one can find that

$$\begin{aligned} E[\hat{\beta}_1 | V, \lambda] &= \frac{s_{11}(s_{22}+\lambda) - s^2}{(s_{11}+\lambda)(s_{22}+\lambda) - s^2}\beta_1 \\ &= g(s_{11}, s_{22}, s)\beta_1 \end{aligned}$$

Furthermore, with some patience, it is possible to work out the Hessian of g:

$$H = \nabla^2 g = \frac{\lambda}{D^3}\begin{pmatrix} -2(s_{22}+\lambda)^3 & -2(s_{22}+\lambda)s^2 & 4(s_{22}+\lambda)^2 s \\ -2(s_{22}+\lambda)s^2 & -2s^2(s_{11}+\lambda) & 2(s_{11}+\lambda)(D+2s^2)s \\ 4(s_{22}+\lambda)^2 s & 2(s_{11}+\lambda)(D+2s^2)s & -2(s_{22}+\lambda)(A+3s^2) \end{pmatrix}.$$

Because $\lambda/D^3 > 0$, negative semidefiniteness of H is equivalent to that of the scaled matrix $\tilde{H} = D^3 H/\lambda$.

$$\tilde{H}_{11} = -2(s_{22}+\lambda)^3 < 0$$

$\det(\tilde{H}_{[1:2,1:2]})$:

$$\begin{aligned} \det(\tilde{H}_{[1:2,1:2]}) &= \det\begin{pmatrix} -2(s_{22}+\lambda)^3 & -2(s_{22}+\lambda)s^2 \\ -2(s_{22}+\lambda)s^2 & -2s^2(s_{11}+\lambda) \end{pmatrix} \\ &= 4(s_{22}+\lambda)^2 s^2 D \geq 0 \end{aligned}$$

and

$$\det(\tilde{H}) = -\,8\,\lambda\,(s_{22}+\lambda)^3 s^2\,(s_{11}+\lambda)\,D \leq 0.$$

As such, H is negative-semidefinite and consequently $g(s_{11}, s_{22}, s)$ is globally concave in all three arguments. Now, let V^* be the bootstrapped version of V and note that $E_{boot}[(s_{11}^*, s_{22}^*, s^*)] = (s_{11}, s_{22}, s)$. By Jensen's inequality:

$$E_{boot}[g(S^*)] \leq g(E_{boot}[S^*]) = g(S).$$

Multiplying by β_1 then gives:

$$E_{boot}[\hat{\beta}_1^*] = E_{boot}[g(S^*)]\beta_1 < g(S)\beta_1 = E[\hat{\beta}_1].$$

A similar argument can be made with the lasso, however, the mathematical details become more involved. Whereas with ridge we started off assuming the true values, with lasso it is easier to work with assumed conditions on the lasso estimates themselves as they directly affect the KKT conditions. Starting with a single parameter set up, assume that $\widehat{\beta}_1 > 0$. Then,

$$\begin{aligned}\frac{1}{n}\mathbf{x}_1^T(\mathbf{y} - \mathbf{x}_1\widehat{\beta}_1) = \lambda\\ \widehat{\beta}_1 &= \frac{1}{ns_{11}}\mathbf{x}_1^T\mathbf{y} - \frac{\lambda}{s_{11}}\\ E[\widehat{\beta}_1] &= \beta_1 - \frac{\lambda}{s_{11}}\end{aligned}$$

Let $h(s_{11}) = -\frac{\lambda}{s_{11}}$, then

$$\begin{aligned}h'(s_{11}) &= \frac{\lambda}{s_{11}^2},\\ h''(s_{11}) &= \frac{-2\lambda}{s_{11}^3} < 0\end{aligned}$$

and $h(s_{11})$ is therefore concave. So by Jensen's inequality

$$E_{boot}[h(s_{11}^*)] \leq h(E_{booot}[s_{11}^*]) = h(s_11).$$

Adding β_1 then gives,

$$E_{boot}[\widehat{\beta}_1^*] = \beta_1 + E_{boot}[h(s_{11}^*)] \leq \beta_1 + g(s_{11}) = E[\widehat{\beta}_1].$$

Now consider a two parameter setting. If we assume $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 = 0$, bias can only be guaranteed when $\beta_2 = 0$. This boils down to the same details as the single parameter case we just considered. If both $\hat{\beta}_1$ and $\hat{\beta}_2$ are assumed positive, we arrive at the following KKT conditions:

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{X}^T y - \lambda 1_2$$

Again, let $\boldsymbol{V} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} s_{11} & s \\ s & s_{22}\end{pmatrix}$, then

$$\hat{\beta}_1 = \frac{s_{22}(x_1^T y - n\lambda) - s(x_2^T y - n\lambda)}{n(s_{11}s_{22} - s^2)}$$

and

$$\begin{aligned}E(\hat{\beta}|X) &= \beta_1 - \lambda\frac{s_{22} - s}{s_{11}s_{22} - s^2}\\ &= \beta_1 + h(s_{11}, s_{22}, s)\end{aligned}$$

Differentiation gives

$$\nabla^2 h \;=\; \frac{\lambda}{D^3}\begin{pmatrix} -2s_{22}^2(s_{22} - s) & -2s_{22}s^2 & 4s_{22}^2 s \\ -2s_{22}s^2 & -2s_{11}s^2 & 2\,s_{11}s\big(D + 2s^2\big) \\ 4\,s_{22}^2 s & 2\,s_{11}s\big(D + 2s^2\big) & -2s_{22}\big(s_{11}s_{22} - s^2 + 3s^2\big)\end{pmatrix}.$$

Leading (1×1) minor:

$$H_{11} = -\frac{2\lambda s_{22}^2(s_{22} - s)}{D^3} < 0$$

Leading (2×2) minor:

$$\det\big(H_{[1:2,1:2]}\big) = \frac{4\lambda^2 s_{22}^2 s^2 D}{D^6} \geq 0.$$

Full determinant:

$$\det(H) = -\frac{8\lambda^3\, s_{22}^3 s^2 s_{11}\, s_{11} D}{D^9} \leq 0.$$

The Hessian is negative-semidefinite. Therefore the map $g(s_{11}, s_{22}, s)$ is jointly concave, and multivariate Jensen's inequality implies

$$E_{\text{boot}}\big[\hat{\beta}_1^*\big] \;\le\; E\big[\hat{\beta}_1 \mid X,\, z_1 = z_2 = +1\big],$$

so the bootstrap mean of $\hat{\beta}_1$ sits strictly closer to zero than the original estimate whenever both fitted coefficients are positive. Note that in each of these settings a common thread is that the curvature, which affects the amount of bias introduced by the bootstrap is heavily dependent on λ. Thus, the larger λ is the more the bootstrap will be biased. We also see in the two parameter setting that the correlation between features has a considerable role in the curvature. How much each of these contribute to the bootstrap bias is problem dependent, however, our exploration, such as the example we provide next, suggests that λ is likely the larger influencer in general. That said, full exploration is outside the scope of this work, but could be explored by considering the eigenvalues for the hessians above under various scenarios.

How does this extend to high dimensions? To answer this question, we consider the following simulation study.

In order to increase interpretability, we consider a simplified scenario here. Consider a set up with $n = p = 100$ where there is 1 true non-null variable, A s.t. $\beta_A = 2$ that is correlated with a null variable B with $\rho = 0.5$. All other variables are generated independent of A and each other. For this set up, A is always selected to be in the model, both with the original data and for all bootstrap replications at λ_{CV}. This is important as decomposing the bias is more complicated for variables that are not selected to be in the model.

Note, the bias for $\widehat{\beta}_j \neq 0$ is $\frac{1}{n}\mathbf{x}_j^T\epsilon + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\boldsymbol{\beta}_{-j}^* - \widehat{\boldsymbol{\beta}}_{-j}) + \lambda$. We can break this down further to apply to the scenario outlined above. $Bias_A = \frac{1}{n}\mathbf{x}_A^T\epsilon + \frac{1}{n}\mathbf{x}_A^T\mathbf{x}_B(\beta_B^* - \widehat{\beta}_B) + \frac{1}{n}\mathbf{x}_A^T\mathbf{X}_N(\boldsymbol{\beta}_N^* - \widehat{\boldsymbol{\beta}}_N) + \lambda$. That is, we can decompose the bias into four parts. The first is the irreducible bias which comes from the chance correlation between $\mathbf{x}_A$ and the errors. The last is the bias directly introduced by the lasso penalty. The other two components attribute bias from the single B variable and all 98 N variables respectively. By taking the simulation set up in the previous paragraph and repeating it 1000 times and each time saving the bias attributable to each of the components, we can get an idea of the distribution of the bias components. More specifically, for each generated dataset, we can decompose the bias for the estimates from the original data as well as from the 1000 bootstrap replications. For the bootstrap replications, we can then save the mean bias. Figure 12 shows a summary from doing just that. The top panel gives the densities of the mean bootstrap biases across the 1000 repetitions, the middle gives the densities for the biases on the original dataset across the 1000 repetitions, and the bottom give the densities of their paired differences. In this depiction, the contribution of λ is excluded. Additionally note here that a positive bias is used to indicated bias *towards* zero.

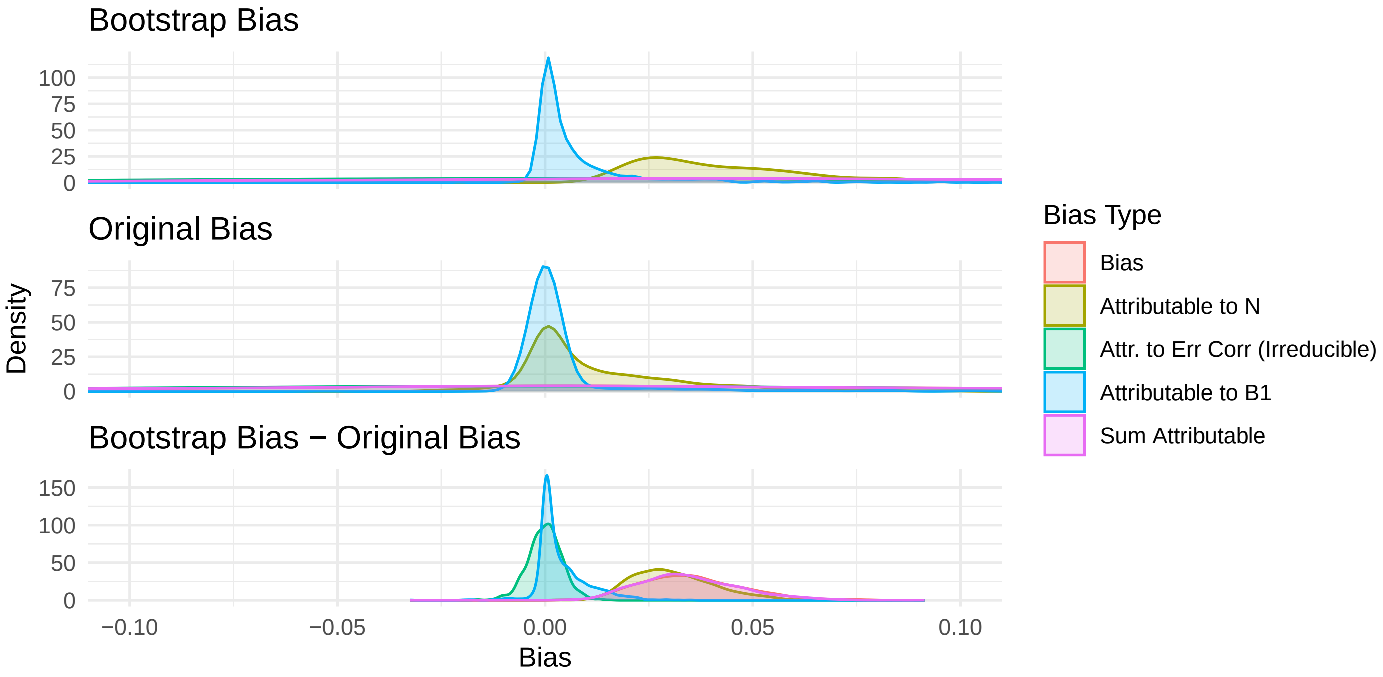


Figure 12: n = p = 100, $\beta_A = 2$, $\beta_B = 0$, $\rho_{A,B} = 0.5$. All other βs = 0 and generated under independence

While the derivation above suggests that increasing correlation is the main contributor to increased bootstrap bias, we see that it is actually the cumulative effect of the 98 N variables that drives the bias rather than the correlation with variable B. This has important implications in that even when overall correlations are low, that the sparsity in high dimensional settings is going to lead to a significant bootstrap bias. It is important to differentiate this bias and the bias inherent in the motivation for Section 2.1. The framework put forth in Section 2.1 allows for the bias introduced by penalization which we are arguing is permissible in this newly suggested coverage framework, however, the additional bootstrap bias here is what inherently leads to the breakdown of the bootstrap even for average coverage.

7.7 A note about stability selection

It is no secret that bootstrapping lasso has fundamental issues. The related work in this manuscript is simply meant to show easy to comprehend details to this end. With these issues in mind, it is not common to see a traditional bootstrapping approach applied to the lasso for the purposes of interval construction. Rather, the popular use of resampling techniques for the lasso is in stability selection introduced by Meinshausen and Bühlmann (2010). However, here, we show that stability selection is still affected by bootstrap bias.

Consider the following set up. $n = 50$, $p = 500$, 4 βs contain signal: $\beta_{1-4} = (0.25, 0.5, 1, 2)$ and the rest are zero. $\mathbf{X}$ were generated independently from a N$(0, 1)$. Finally, $\mathbf{y}$ was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$. Using this data setup, stability selection was performed as outlined in Algorithm 2.

Algorithm 2 Bootstrap Stability Selection at a *single* CV-chosen λ

```
Require: $R$ replicated data sets, $B$ bootstraps per data set, $p$ predictors
  Let: $\mathbf{A} \in \{0,1\}^{R\times p}$, $\mathbf{A}^* \in [0,1]^{R\times p}$
for $i = 1, \dots, R$
    Generate $(\mathbf{X}, \mathbf{y})$ from the data-generating process
    $\lambda_{\text{CV}} \leftarrow \arg\min_\lambda \text{CV-Error}(\lambda)$        ▷ 10-fold CV on $(\mathbf{X}, \mathbf{y})$
    Obtain lasso estimates at $\lambda_{\text{CV}}$ and save $\mathbf{A}_{i,\cdot} \leftarrow \left[\hat{\beta}(\lambda_{\text{CV}}) \neq 0\right]$
    Initialise $\mathbf{B}^* \leftarrow \mathbf{0}_{B\times p}$
    for $b = 1, \dots, B$        ▷ pairs bootstraps
        Draw indices $\mathcal{I}_b$ with replacement from $\{1, \dots, n\}$
        $(\mathbf{X}^b, \mathbf{y}^b) \leftarrow (\mathbf{X}_{\mathcal{I}_b,\cdot}, \mathbf{y}_{\mathcal{I}_b})$
        Fit lasso on $(\mathbf{X}^b, \mathbf{y}^b)$ at $\lambda_{\text{CV}}$
        $\mathbf{B}^*_{b,\cdot} \leftarrow \left[\hat{\beta}^*(\lambda_{\text{CV}}) \neq 0\right]$
    $\mathbf{A}^*_{i,\cdot} \leftarrow \frac{1}{B}\sum_{b=1}^{B} \mathbf{B}^*_{b,\cdot}$
```

After we computed $\bar{\mathbf{A}} \leftarrow \frac{1}{R}\sum_{i=1}^{R} \mathbf{A}_{i,\cdot}$, $\bar{\mathbf{A}}^* \leftarrow \frac{1}{R}\sum_{i=1}^{R} \mathbf{A}^*_{i,\cdot}$ and then compare $\bar{\mathbf{A}}$ (prob. of being selected in the original fit) with $\bar{\mathbf{A}}^*$ (expected bootstrap stability) to evaluate how well the inclusion probabilities mirror single-λ selection behavior. The results are presented in Table 4.

	Inclusion probability	
Predictor	Original selection (%)	Bootstrap stability (%)
β_1	19.1%	9.8%
β_2	64.0%	34.5%
β_3	99.2%	87.9%
β_4	100.0%	100.0%

Table 4: Results for simulation described in Section 7.7 showing that stability selection also suffers from bootstrap bias. Original selection(%) provides the empirical selection probabilities for the 4 parameters while Bootstrap stability gives the bootstrap estimate for selection.

Clearly, there is a fundamental bias for stability selection. That said, further exploration is necessary to understand if this contradicts the claims of Meinshausen and Bühlmann (2010) as they primarily focus on FDR control for finite sample problems like the one considered here.

Note, this implementation deviates from the proposed stability selection algorithm originally introduced by Meinshausen and Bühlmann (2010) in order to retain a connection to the bootstrap implementation in this manuscript, however, even when considering the entire lasso path fit on sub-bagged samples the bias remained largely the same.