Imperial College London

Department of Mathematics

Comparative Evaluation of Post-Selection Inference Methods with LASSO

Sian Chee Tong

CID: 02085533

Supervised by Alastair Young

September 2, 2022

Submitted in partial fulfilment of the requirements for the MSc in Statistics of Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Sian Chee Tong Date: September 2, 2022

Abstract

We present the problem of classical inference after model selection, and discuss post-selection inference as a valid framework applicable to modern-day statistics. We focus our discussion over four novel methods in post-selection inference: Data splitting Cox [1975], Selective inference after LASSO selection Lee et al. [2016], Selective inference with randomized responses Tian and Taylor [2018], and (U, V) decomposition Rasines and Young [2021]. We focus on providing comparison studies on these methods with LASSO as the model selection procedure and subsequent inference for selected regression coefficients.

Valid inference after (U, V) decomposition and powerful dominance over data splitting are not thoroughly investigated when the sample size is small and under violation of normality assumptions. In the first part of the paper, we consider empirically evaluating the performance of (U, V) decomposition under these conditions. We show that (U, V) decomposition remains more powerful than data splitting in terms of selective coverage levels and confidence interval lengths. Moreover, we find the presence of oracle properties in (U, V) decomposition and therefore asymptotic requirement may be relaxed.

In the second part, we propose a conditional Monte-Carlo (MC) rejection sampling method for approximating Lee et al. [2016] on conditioning the model selected by LASSO which is known to be analytically intractable. We extend our method by providing randomization to the responses. We provide an evaluation of our method at the inference stage, comparing with other methods. We show that the dominance in inferential power against other methods can be potentially substantial at the expense of high computational cost.

Acknowledgements

I would like to thank Professor Alastair Young for his intellectual support towards the completion of my work. The conditional on selection Monte-Carlo scheme is necessarily crucial for providing better inferential power in post-selection inference in the future.

Furthermore, I would like to express my gratitude to the Imperial College Department of Mathematics for always be there for my queries.

I am immensely thankful to my parents and family for providing unconditional support towards my academic aspirations. I am truly blessed to have them in my life.

Contents

1. Introduction

Suppose we are interested in doing a (frequentist) statistical inference for a given dataset. We usually have a pre-determined probability model for the unknown data generating mechanism on data, Y and the hypothesis to be tested based on the data. For example, a common modelling structure is normal linear regression, and we perform hypothesis on a regression coefficient whether it is significantly different from zero. In the midst of exploratory data analysis, a variable selection method is determined to choose significant variables for further investigation. At the inference stage, we may opt to apply the conventional inferential procedure accordingly using our pre-determined model. In the normal linear regression setting, this simply mean computing confidence intervals and p-values using appropriate test statistics. As we have "cherry-picked" the variables, it is not surprising at all to conclude many significant inferential results, allowing us to report over-optimistic discoveries.

Unfortunately, this misleading approach is statistically wrong, and the fallacy lies from the fact that data was used twice, once for model selection and again for inference. This is also known as data snooping White [2000]. Thus, selection bias led towards the significance in our discovery (see Yong [2012], Begley and Ioannidis [2015]). Worse, the inferential properties no longer hold, including wrong distributional test statistics Lockhart et al. [2014] and inflated type-I errors Fithian et al. [2014].

The post-selection inference addresses the problem by compensating for the fact that inferences are made upon adaptive selection through variable selection methods. There are currently three prominent approaches in post-selection inference. The simplest one most Statisticians would agree on is data splitting (Cox [1975],Wasserman and Roeder [2009]). Secondly, the simultaneous inference approach (Benjamini and Yekutieli [2005], Berk et al. [2013]). This approach guarantees validity by controlling the false discovery rate (FDR) by consideration of simultaneous coverages across all variable selection procedures. Thirdly, the conditional inference approach, which considers inference based on the conditional distribution of the data given the selected model or more precisely, selection events (see Fithian et al. [2014], Lee et al. [2016]). Tian and Taylor [2018] establishes the randomized model selection framework, which simply injects randomized noise to the data, resulting in more powerful selective tests. This is adapted by Rasines and Young [2021] to develop a new strategy, namely (U, V) decomposition to improve over data splitting.

In this report, we provide evaluation studies across the methods in the conditional inference framework. In particular, we place attention on extending the evaluation of (U, V) decomposition under non-asymptotic settings. We will also demonstrate a conditional MC rejection sampling based approach as an approximating extension to Lee et al. [2016].

The structure of the report is organized as follows. In Section 2, we provide a brief background on the post-selection inference procedure for the LASSO. Our chosen variable selection procedure is LASSO mainly because we wish to align with the framework by Lee et al. [2016]. In between, we make sense of the inferential objectives in post-selection inference. We provide a detailed account on data splitting, the polyhedral method paired with LASSO by Lee et al. [2016], and (U, V) decomposition by Rasines and Young [2021], as we will be utilizing these methods throughout our work. We emphasize the benefits and limitation of all methods, especially for Lee et al. [2016] in terms of frequency properties, which leads us to propose Monte-Carlo (MC) approximations to the polyhedral method in Section 4 before extending the approximation to randomized noise schemes.

Prior to Section 4, we provide an empirical review of (U, V) decomposition. We conduct a sensitivity analyses to changes in assumptions made by Rasines and Young [2021]. This allows us to examine the robustness of the approach and sets the stage for comparing against other methods, particularly the randomized MC method proposed in Section 4. We consider the case when there are non-normal errors and altering the variance structure when it is in fact known. Our results are compared against data splitting as in the original paper by Rasines and Young [2021] but we focus on the inferential stage mainly because our purpose is to examine the validity of frequency properties. We find that (U, V) decomposition remains relatively insensitive towards assumption violations, when error distributions are skewed and heavy-tailed respectively. We also see that having fixed, known variance indeed contributes to smaller deviation in interval lengths when compared to estimated cases, but not by much. In all cases, (U, V) decomposition remains dominant over data splitting with regards to inferential performance. We conclude that even without being in an asymptotic regime, (U, V) decomposition remains relatively powerful, at least against data splitting.

We show that while (U, V) decomposition sets out remarkably simple formulation and analysis, the loss in inferential power in terms of interval lengths is noticeable when comparing against the MC approximation method in Section 4. This is because (U, V) decomposition discards the selection stage information required to maximize inferential power. Further, we emphasize that the MC approach opens up opportunity for further enhancement and thus, potentially providing better inferential power than other methods considered so far. In Section 5, we conclude our findings with discussions of potential future works.

2. Preliminaries

In this section, we briefly describe for post-selection inference in a linear regression model with emphasis on inferential specifications. We provide review of the key selective inference methods that we will be using in our work: (1) Data splitting, (2) Selective inference for the LASSO with the Polyhedral Lemma, and (3) (U, V) decomposition.

We use the statistical convention to represent a random vector or matrix as a capital letter, and an observed vector as a lowercase letter.

2.1. Post-selection inference in linear regression with LASSO

When the goal is to study the effects on the responses from the exploratory variables, the standard approach is to model the data with a linear regression model. We require that our response vector: $Y \in \mathbb{R}^n$ following the form:

$$Y = \mu + \mathcal{E}, \tag{2.1}$$

where $\mu = X\beta$ is a linear function of p known and specified variables, which constitute the design matrix, $X = (X_1, \dots, X_p) \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ is a vector of unknown model parameters. We assume that the errors, $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n) \in \mathbb{R}^n$ are independent and identically distributed (i.i.d), following an unknown distribution with mean zero and variance, σ^2. When $\mathcal{E} \sim \mathcal{N}(0_n, \sigma^2 I_n)$, then (2.1) is the normal linear regression model. From now on, we assume the normal linear regression model, unless specified concretely.

Having explored the data, we wish to identify potentially interesting variables before proceeding to the inference stage. This is often guided by variable selection (also known as model selection) procedures to choose a subset of variables, $M \subseteq \{1, \dots, p\}$. If the variable selection procedure is the LASSO, then, for a given regularization parameter λ, LASSO chooses the "interesting" variables for us, upon solving the convex optimization problem:

$$\hat{\beta}^{LASSO} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1, \tag{2.2}$$

where $\hat{\beta}^{LASSO}$ are the LASSO estimates, that assigns the selected variables to have the coefficients, $\hat{\beta}^{LASSO} \neq 0$, and the remaining variables are "eliminated", i.e., $\hat{\beta}^{LASSO} = 0$. Notice that we do not consider including the intercept term, β_0 for the sake of alignment to our defined regression model in (2.1).

We denote variables selected by LASSO by $\hat{M} = \{j : \hat{\beta}_j \neq 0\}$, where $j \subseteq \{1, \ldots, p\}$. Note that under a set of m models, $\mathcal{M} = \{M_1, \ldots, M_m\}$ with non-zero selection probability, $\hat{M}$ may not necessarily correspond to the true model, $M_{true} \in \mathcal{M}$. Once our variables are selected, we will need to decide on how are we going to form our targets for inference. Liu et al. [2018] provides two types of target formations. Here, we briefly describe these two types and present their differences as a starting point of our remaining workflow. In essence, we ask the question whether our targets, determined through the regression coefficients, are defined with respect to the full model with all the variables, or the partial model with only the selected variables.

Inference for the full and selected regression coefficients

Recall that in the classical inference setting, we fix a model *prior* to viewing the data, and our goal is to infer the full regression coefficients:

$$\beta^F = (X^T X)^{-1} X^T \mu. \tag{2.3}$$

For brevity, we sometimes refer β^F as β. Now, in the post-selection inference context, we can also summarize our relationship between Y and each variable, X_j through β_j. The key difference is that now, LASSO has identified the subset of components of β for us to estimate, the selected regression coefficients, i.e., $\{\beta_j : j \in \hat{M}\}$, but, we do not want LASSO to inform us how we should treat the relationship between X and Y. However, naively constructing confidence intervals and p-values based on the classical theory becomes invalid without conditioning on the selected variables, $\{j \in \hat{M}\}$. Fithian et al. [2014] and Liu et al. [2018] formulates the requirement to satisfy the properties for valid inference using conditional inference. We pursue the modification of (2.3) when splitting strategies are involved, in the case of data splitting in Section 2.2 and (U, V) decomposition in Section 2.4.

Inference for the partial regression coefficients

In contrast, it is convenient to infer the coefficients corresponding to $\hat{M}$ selected by LASSO. Say, we have the selected model as $\hat{M} = M$. Therefore, in this case, we do not need to provide estimates of the coefficients as in (2.3), hence, the name partial regression coefficients or submodel regression coefficients. Here, the inference is equivalent to projecting μ onto the variable X_M Berk et al. [2013]:

$$\beta_M = (X_M^T X_M)^{-1} X_M^T \mu. \tag{2.4}$$

Note that unlike making inference for β, making inference for β_M can be random due to the randomness in $\hat{M}$, provided that the penalty parameter, λ is not fixed, e.g. obtained by cross-validation. In fact, for a linear regression model with p variables, there are 2^p possible model combinations upon variable selection. The conditional inference framework including Fithian et al. [2014], Lee et al. [2016], Tian and Taylor [2018], Rasines and Young [2021], makes inference based on the target in (2.4), as we shall also illustrate in the upcoming sections.

2.2. Data splitting

Simply, data splitting is an "information" splitting strategy that involves partitioning the data randomly into two subsets. This approach is well-known, and is primarily used in machine learning as a technique to prevent model overfitting and generalizing predictive performance. The first subset is allocated for training the model, and the hold-out test set is for assessing the model performance.

In selective inference, data splitting turns into allocating the first subset for model selection, the hold-out subsets are used for inference. Formally, the data splitting procedure is as follows:

Step 1: For a data pair $(X, y) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$:

A random partition, $\mathcal{P}$ splits the data into subsets of sizes n_1 and $n_2 = n - n_1$: $(X^{(1)}, y^{(1)}) \in \mathbb{R}^{n_1\times p} \times \mathbb{R}^{n_1}$ and $(X^{(2)}, y^{(2)}) \in \mathbb{R}^{n_2\times p} \times \mathbb{R}^{n_2}$. Each data point is equally likely to be assigned to either one of the subsets.

Step 2: Apply a model selection procedure to $(X^{(1)}, y^{(1)})$ to choose a submodel $\hat{M}(X^{(1)}, y^{(1)})$.

Step 3: Depending on our inferential objectives, we formulate our estimates differently on $(X^{(2)}, y^{(2)})$.

Notice that unlike (2.3) and (2.4), we choose our inferential objectives based on only $(X^{(2)}, y^{(2)})$. If it is based on (X, y), then this is instead referred to as data carving Fithian et al. [2014]. We consider modifications of (2.3) and (2.4) in the context of data splitting.

To make inference for the full regression coefficients *based on* $(X^{(2)}, y^{(2)})$, then (2.3) becomes:

$$\beta^{DS}(X^{(2)}) = (X^{(2)T}X^{(2)})^{-1}X^{(2)T}\mu_2, \tag{2.5}$$

where $\mu_2 = \mathbb{E}[Y^{(2)}] = X^{(2)}\beta$, and β is now $\beta^{DS}(X^{(2)})$, that is, the target parameters are now a function of $X^{(2)}$. However, we are interested in making inference for selected regression coefficients of $\hat{M}$. So, we make inference for the subset of coefficients: $\{\beta_j^{DS}(X^{(2)}) : j \in \hat{M}\}$.

If we make inference for the partial regression coefficients *based on* $(X^{(2)}, y^{(2)})$, then, (2.4) becomes :

$$\beta_M^{DS}(X^{(2)}) = (X_M^{(2)T}X_M^{(2)})^{-1}X_M^{(2)T}\mu_2, \tag{2.6}$$

where μ_2 is the same as defined before. As highlighted by Rasines and Young [2021], we stress that $\beta_M^{DS}(X^{(2)})$ is strictly not the same target as the "full" partial regression coefficients, $\beta_M^{DS}(X)$, unless $X = X^{(2)}$. This is because the model selection outcome depends on the data provided to LASSO. We restrict our target coefficients based on the chosen model, $\hat{M} = M$. In another chosen model, say, M', our target coefficients changes accordingly in the sense that $\beta_{M'}^{DS}(X^{(2)}) \neq \beta_M^{DS}(X^{(2)})$. Technically, to construct valid inference, we require to condition not only on the variable $j \in \hat{M}$ but also to the specific model that is selected, $\hat{M} = M$.

Assuming that the splitting subsets, $(X^{(1)}, y^{(1)})$ and $(X^{(2)}, y^{(2)})$ are independent, we can make inference based on the classical theory as if the set of chosen variables $\hat{M}$ are fixed in the first place. In constructing valid confidence intervals, data splitting yields a conditional coverage guarantee that satisfies:

$$\mathbb{P}(\beta_j^{DS}(X^{(2)}) \in C_j | j \in \hat{M}) \geq 1 - \alpha, \qquad \text{(Inference for selected regression coefficients)}$$

$$\mathbb{P}(\beta_{j,M}^{DS}(X^{(2)}) \in C_{j,M} | j \in \hat{M}, \hat{M} = M) \geq 1 - \alpha, \qquad \text{(Inference for partial regression coefficients)}$$

where C_j and $C_{j,M}$ are the confidence regions for coefficient, j in the two different target formulations. Here, α is the required desired significance level, and $(1 - \alpha)$ is the nominal coverage level. The conditional requirements can be relaxed due to independence.

So far, we have assumed that data samples are sufficient enough for splitting. When we are low in sample size, n, data splitting becomes very undesirable as this results in inconsistent estimates of the target parameters. Additionally, Fithian et al. [2014] argues that data splitting discards the available information that remains after conditioning on the selected $\hat{M}$, thereby advocating data carving instead.

While data splitting can be very simple to implement, it is not without the drawbacks above that reduces model selection and inferential power. Specifically, this includes model selection instabilities, thus, prone to choosing the wrong models for inference, and wider confidence intervals Fithian et al. [2014], Tian and Taylor [2018], Rasines and Young [2021]. However, one major advantage of data splitting is that it can be applied for any model selection procedure.

2.3. Selective inference for the LASSO with Polyhedral Lemma

In the condition on selection framework, Lee et al. [2016] and Markovic et al. [2017] show that many variable selection procedures can have convex "selection events". For example, using the LASSO solves the convex optimization problem at a fixed λ to yield a unique sparse solution, provided that $n \geq p$.

Given observed y, the selection event, $\{y : \hat{M}(X, y) = M\}$ can be characterized by a polyhedron of the form: $\{y : Ay \leq b\}$, where A and b can be derived analytically. Tibshirani and Taylor [2011] and Lee et al. [2016] presented the derivations from the Karush-Kuhn Tucker conditions for LASSO.

We briefly review the work of Lee et al. [2016] which characterizes LASSO selection events for post-selection inference through the polyhedral method. The theory focuses on inference for the submodel in (2.4). Once again, suppose our parameters of interest are in the submodel after fitting LASSO:

$$\beta_M = \{\beta_{j,M} : \forall j \in \hat{M}, \hat{M} = M\} \in \mathbb{R}^{|M|},$$

such that $\{\hat{M} = M\} = \{j : \hat{\beta}_{j,M} \neq 0\}$ is dependent on the LASSO selection outcome and $|M|$ is the number of selected variables. For a particular selected variable j, our goal is to construct a valid $100(1-\alpha)\%$ selective confidence interval on $\beta_{j,M}$ that satisfies:

$$\mathbb{P}(\beta_{j,M} \in C_{j,M} | j \in \hat{M}, \hat{M} = M) \geq 1 - \alpha.$$

Note that $\beta_{j,M}$ is a particular case for the linear contrast $\eta_{j,M}^T \mu$ since

$$\beta_{j,M} = \mathbb{E}[X_{j,M}^{\dagger} Y] = X_{j,M}^{\dagger} \mu \ := \eta_{j,M}^T \mu,$$

where $X_{j,M}^{\dagger} = e_j^T (X_M^T X_M)^{-1} X_M^T$ is the Moore-Penrose pseudo-inverse matrix of the j^{th} column of X_M. The vector $e_j \in \mathbb{R}^{|M|}$ has all 0's but 1 at the j^{th} component. The subscript of M in $\eta_{j,M}$ emphasizes the dependence on the selected model.

Unfortunately, it turns out that when we condition only on $\{j \in \hat{M}, \hat{M} = M\}$, the distribution of Y conditions on the union of up to $2^{|M|}$ polyhedrons, which is difficult to characterize. Ideally, for accurate inference, we require only to condition the selected model, but this requires sampling methods to approximate the sampling distribution of $\eta_{j,M}^T Y | \{j \in \hat{M}, \hat{M} = M\}$. This distributional form is deferred to section 4.

Following Fithian et al. [2014], we require conditioning on a "refined" event to explicitly characterize our distribution. We have to condition on $\{\hat{S} = s\}$, the signs of $\hat{\beta}_{j,M}$:

$$\hat{S} = \{sign(\hat{\beta}_{j,M}) : \forall j \in \hat{M}, \hat{M} = M\} \in \{-1, 1\}^{|M|}, \tag{2.7}$$

which leads to restricting only to a single polyhedron. Now, we have that $\{j \in \hat{M}, \hat{M} = M, \hat{S} = s\} = \{Ay \leq b\}$. Both A an b are now dependent on M and s, so that, $A = A(M, s)$ and $b = b(M, s)$. Moving forward, we omit the functional dependency for brevity.

While $\{Ay \leq b\}$ is now well-defined, it turns out that a tractable distribution is well-characterized only when we condition on the residual of Y from projecting onto $\eta_{j,M}$. Mathematically, the residual is $(I_n - P_{\eta_{j,M}})Y = z$. This is because the end points of the distribution, $[\mathcal{V}^-(z), \mathcal{V}^+(z)]$ depends on the value of z (cf. Figure 2 Lee et al. [2016]). Explicit forms of $\mathcal{V}^-(z)$ and $\mathcal{V}^+(z)$ can be found in Lemma 5.1 in Lee et al. [2016].

Now, given the conditions above, the conditional distribution for $\eta_{j,M}^T Y$ is now tractable:

$$\eta_{j,M}^T Y | \{j \in \hat{M}, \hat{M} = M, \hat{S} = s, (I_n - P_{\eta_{j,M}})Y = z\} \sim \mathcal{TN}(\eta_{j,M}^T \mu, \sigma^2 \left\|\eta_{j,M}\right\|_2^2, [\mathcal{V}^-(z), \mathcal{V}^+(z)]), \tag{2.8}$$

following a univariate normal distribution truncated to the interval $[\mathcal{V}^-(z), \mathcal{V}^+(z)]$. The form, $\mathcal{TN}(a, b, [c, d])$ implies $\mathcal{N}(a, b)$ truncated to the interval $[c, d]$.

With (2.8), this allows us to make exact inferences (hypothesis tests and confidence intervals) when paired with LASSO as the model selection method. For example, the $100(1-\alpha)\%$ confidence intervals for $\beta_{j,M}$, $[L, U]$ can be constructed from the pivotal quantities that satisfy the following equations Lee et al. [2016]:

$$F^{[\mathcal{V}^-(z),\mathcal{V}^+(z)]}_{L,\sigma^2\|\eta_{j,M}\|_2^2}(\eta_{j,M}^T y) = 1 - \frac{\alpha}{2}, \tag{2.9}$$

$$F^{[\mathcal{V}^-(z),\mathcal{V}^+(z)]}_{U,\sigma^2\|\eta_{j,M}\|_2^2}(\eta_{j,M}^T y) = \frac{\alpha}{2}. \tag{2.10}$$

where $F_{a,b}^{[c,d]}$ is the cumulative distribution function (CDF) of $\mathcal{N}(a,b)$, truncated to the interval $[c,d]$.

Despite providing exact inference, this approach is not without shortcomings. Kivaranovic and Leeb [2021] points out that the interval, $[L,U]$ can result in infinite expected confidence interval length when the truncation set $[\mathcal{V}^-(z),\mathcal{V}^+(z)]$ is bounded from above and below (cf. Figure 1 Kivaranovic and Leeb [2020] and Figure 4 Lee et al. [2016]). Notice also that (2.8) can be very restrictive as it conditions on refined events (up to signs). As we condition more on the selection stage, Fithian et al. [2014] argues that less information is allocated for inference, so we are less certain on our estimates, which adds up to longer intervals.

This leads to the developments in randomization approaches Tian and Taylor [2018] Rasines and Young [2021], demonstratably providing shorter intervals and coverage improvements.

2.4. Randomized selective inference and (U,V) decomposition

Initially addressed by Tian and Taylor [2018], randomization schemes attempt to tackle the drawbacks in data splitting and data carving approaches highlighted in Section 2.2 and 2.3. In linear regression, we can introduce an i.i.d additive noise vector, $W \in \mathbb{R}^n$ to Y, an extension to (2.1). Mathematically:

$$Y^* = Y + W, \quad W \sim \mathbb{Q}, \tag{2.11}$$

where $\mathbb{Q}$ is a specified distribution that generates the randomized noise. In particular, when this is paired with the polyhedral lemma with LASSO described in Section 2.3, the conditional distribution is expressed as follows:

$$\eta_{j,M}^T Y | \{j \in \hat{M}, \hat{M}(X,Y^*) = M, \hat{S}(X,Y^*) = s, (I_n - P_{\eta_{j,M}})Y = z\}. \tag{2.12}$$

A natural approach is to generate Gaussian noise: $W \sim \mathcal{N}(0,\sigma^2)$. This has shown to be advantageous in terms of parameter estimation consistency and power (cf. Figure 2 and Figure 3 in Tian and Taylor [2018]) compared to data splitting and data carving (within the conditional inference framework). Tian and Taylor [2018] note that the explicit conditional form in (2.12) is analytically intractable, resorting to hit and run Bélisle et al. [1993] and Hamiltonian Monte-Carlo sampling Pakman and Paninski [2014].

However, (2.12) is rather restricting because we require to choose the LASSO to perform model selection. Rasines and Young [2021] shows that it is possible to avoid such restrictions by the (U,V) decomposition. The proposed strategy randomizes responses in two following forms: $U = Y + \gamma W$ and $V = Y - \gamma^{-1} W$.

It can be shown that they are both independent and normally distributed (see Appendix (A.2)):

$$U \sim \mathcal{N}(\mu, (1+\gamma^2)\sigma^2 I_n), \tag{2.13}$$
$$V \sim \mathcal{N}(\mu, (1+\gamma^{-2})\sigma^2 I_n). \tag{2.14}$$

U is allocated for selection, while V is allocated for inference. Therefore, it is also a splitting strategy like data splitting in section 2.2. Now, $\gamma > 0$ is fixed to control the amount of (Fisher) information split between selection and inference stages. Rasines and Young [2021] proposed that when $\gamma = \sqrt{\frac{1}{f} - 1}$, where f is the splitting fraction or the proportion of samples assigned to the selection stage, then it offers more balanced information split compared to data splitting. Loosely speaking, this combines the simplicity of data splitting and the benefits of randomization, offering superiority against data splitting in terms of selection and inferential power.

With data pair (X, U), we apply a model selection procedure such as LASSO and choose a submodel $\hat{M}(X, U)$. As U and V are independent, this resonates with the data splitting approach, but here, we are not wasting any samples at all. Therefore, we can safely base our inference using (X, V).

If we base our inference for β^F, then, $\beta^{UV}(X) = (X^T X)^{-1} X^T \mu$ which is equivalent to (2.3). For making inference for the selected regression coefficients, we simply take the subset of our coefficients: $\{\beta_j^{UV}(X) : j \in \hat{M}\}$.

On the other hand, if we make inference for the partial regression coefficients *based on* (X, V), then, again, we achieve the same form as (2.4):

$$\beta_M^{UV}(X) = (X_M^T X_M)^{-1} X_M^T \mu. \tag{2.15}$$

Therefore, it is natural that we can also yield valid conditional coverage guarantee that satisfies:

$$\mathbb{P}(\beta_j^{UV}(X) \in C_j | j \in \hat{M}) \geq 1 - \alpha,$$
$$\text{(Inference for selected regression coefficients)}$$
$$\mathbb{P}(\beta_{j,M}^{UV}(X) \in C_{j,M} | j \in \hat{M}, \hat{M} = M) \geq 1 - \alpha.$$
$$\text{(Inference for partial regression coefficients)}$$

Unfortunately, it is obvious to spot that when errors are not normal and error variances are unknown, U and V fails to be normally distributed described in (2.13) and (2.14). Rasines and Young [2021] suggested that under mild conditions, generalized version of (U, V) is asymptotically valid. So far, simulations suggests that through estimating variance, σ^2, (U, V) seems to perform reasonably well against data splitting in all cases.

In the following section, we aim to investigate the robustness of (U, V) under violation of normal error assumption. When the variance is assumed known, we seek to empirically investigate whether there is significant improvement to the performance of (U, V).

3. Sensitivity analyses on (U, V) decomposition

From now on, we will be using the short form "(U, V)" to refer to (U, V) decomposition and "DS" to refer to data splitting for simplicity.

3.1. Previous work

Before we present our work, we briefly describe the simulation setup and summarize key results from the inference sections in the original paper by Rasines and Young [2021]. Precisely, we refer to the results from making inference for selected regression coefficients (from full regression coefficients) and the projection parameters (partial regression coefficients). Although the interpretation is not necessarily equivalent, the inferential objectives described by [Rasines and Young, 2021] corresponds to considering low-dimensional $(n > p)$ and high-dimensional settings $(n < p)$. This is because we can also infer for the partial regression coefficients in the low-dimensional setting.

In their simulation study, they considered the standard normal linear regression setting (2.1). In both inference scenarios, they set $n = 200$. For the design matrix X, they considered for each row independently sampled from $\mathcal{N}(0, \Gamma)$ with Γ having a Toeplitz covariance structure at each replication of the simulations, using total of $B = 5000$ replications. Specifically, the (i, j) entry of X is in the form $\rho^{|i-j|}, \rho \geq 0$. The higher the value of ρ, the higher the pairwise correlation between two variables. The true coefficient vector, $\beta = \{1, -1, 0.5, -0.5, 0.2, -0.2, 0, \dots, 0\} \in \mathbb{R}^p$, where the remaining $(n-6)$ coefficients are zeros. The dimensionality, $p = 30$ for $n > p$ case and $p = 400$ for $n < p$ case. The true variance is fixed at $\sigma^2 = 1$, but is regarded as estimated for all replications.

Two model selection methods were considered: (1) Fixed-X knockoff selection Barber and Candès [2015], and (2) Stability selection paired with LASSO Meinshausen and Bühlmann [2010]. These methods are primarily useful to control the expected number of false positive selections (false discoveries) but at the expense of computational costs. As fixed-X knockoffs deals with $n > p$ case, both methods were considered when making inference for selected regression coefficients, but only stability selection paired with LASSO was considered in making inference for the partial regression coefficients.

(U, V) and DS were compared over two varying conditions: The proportion of data for selection (splitting fraction), $f \in \{\frac{1}{2}, \frac{3}{4}\}$ and $\rho \in \{\frac{1}{2}, \frac{3}{4}\}$. Performance of both methods were assessed on their simulated coverages and interval lengths of their corresponding absolute value of regression coefficient values, $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. In both inferential objectives, they have shown that both splitting strategies reached close to nominal coverage at 90%, with no visible dominance over another. In terms of interval lengths, they found that when $n > p$ and $\rho = 0.5$, both (U, V) and DS reported longer interval lengths, but (U, V) produces shorter lengths than DS overall. When $f = \frac{3}{4}$, they found bigger length differences between the two strategies, concluding obvious advantage of (U, V) over DS in such cases. However, for the case $n < p$, the average lengths were fairly similar between the two strategies.

Nonetheless, the authors concluded that (U, V) remains advantageous against DS because (U, V) largely reported superior selection power than DS (cf. Figure 1 and Figure 2 Rasines and Young [2021]).

3.2. Our work

Our work aligns closely to the settings in the original paper. Here, we also wish to compare (U, V) with DS. We place our attention to assessing the coverage probabilities and confidence interval lengths when making inference for selected regression coefficients $(n > p)$ and partial regression coefficients $(n < p)$.

We seek to understand whether (U, V) maintains good performance in cases when the error distribution of $\mathcal{E}$ is no longer normal and also whether it improves considerably when σ^2 is known.

In Section 3.1, we consider generating non-normal errors. In Section 3.2, we resort back to normally distributed errors. Instead of estimating σ^2, we assume that σ^2 is known. We investigate the performance impact on (U, V) under known low and high σ^2.

The simulation settings are similar to previous works. X is generated as described in Section 3.1, and we use $n = 200, p = 30$ and $n = 100, p = 150$ for inference cases $n > p$ and $n < p$ respectively. We also set $\beta = \{1, -1, 0.5, -0.5, 0.2, -0.2, 0, \ldots, 0\}$ and compute confidence intervals at 90% nominal level. We use $B = 2000$ replications of simulations for all cases considered.

In order to make comparisons manageable, we consider averaging all f and ρ combinations for all simulations, where $f \in \{\frac{1}{2}, \frac{3}{4}\}$ and $\rho \in \{\frac{1}{2}, \frac{3}{4}\}$. This is because our main focus is to experiment the effects of non-normal errors violation and meeting a known variance assumption, instead of reviewing the effects of different f and ρ, which has been previously concluded before.

When σ^2 is considered unknown and $n > p$, we estimate σ^2 using the residual sum of squares formula:

$$\hat{\sigma}^2_{LD} = \frac{1}{n-p} \|Y - X\beta\|_2^2 . \tag{3.1}$$

Otherwise, when $n < p$, we consider the alternative variance estimation proposal by Reid et al. [2016]:

$$\hat{\sigma}^2_{HD} = \frac{1}{n - p_{\hat{\lambda}}} \parallel Y - X\hat{\beta}^{LASSO}_{\hat{\lambda}} \parallel_2^2 . \tag{3.2}$$

Here, $\hat{\beta}^{LASSO}_{\hat{\lambda}}$ corresponds to (2.2) with $\hat{\lambda}$ chosen by 10-fold cross validation and $p_{\hat{\lambda}}$ corresponds to the number of active variables remaining after carrying out LASSO. To deal with possible negative degrees of freedom in the denominator, we set $max(n - p_{\hat{\lambda}}, 1)$.

In terms of the model selection procedure, we settle with using only the LASSO in (2.2) in the low-dimensional setting $(n > p)$. In addition to saving computational costs, we find that upon replicating the simulations in the paper, using the LASSO still achieves the desired conclusions where the dominance of (U, V) still holds over DS. This also allows us to verify the claim that (U, V) holds for any model selection procedure.

However, in the high-dimensional setting $(n < p)$, we stick to the choice of implementing the stability selection rule paired with LASSO. Although the computational demands are higher, it is evidently worth avoiding the costly variable selection instabilities caused by implementing LASSO when $n < p$. We could have decided to implement the LASSO accordingly, but we might end up risking on inferring the wrong targets via the coefficients. Liu et al. [2018] highlights the importance of considering the target formation costs when conducting post-selection inference.

The implementations for DS and (U, V) adhere to the definitions described in Section 2.1 and 2.4.

3.3. Generating non-normal errors

We consider $\mathcal{E} \nsim \mathcal{N}(0_n, \sigma^2 I_n)$. We assume that we still generate Gaussian randomized noise, $W \sim \mathcal{N}(0_n, \hat{\sigma}^2 I_n)$ where $\hat{\sigma}^2$ represents (3.1) or (3.2). Following Section 2.4, we can deduce that U and V are no longer normal and are correlated.

We consider generating two types of error distributions which allow us to examine the effects of heavy tail and skewed distributions on the performance of (U, V), with DS as the baseline comparison:

1. t-distribution
2. Skewed normal distribution

In the first case, we generate t-distributed errors with k degrees of freedom where $k \in \{1, 2, 5, 10, \infty\}$. When k approaches ∞, then this approximates the standard normal distribution. When k is set lower, the shape of the t-distribution tend towards heavier tails.

Then, we consider skewed normal errors with varying skewness parameter, ξ, where $\xi \in \{0.1, 0.7, 1, 5, 10\}$. This comes from the class of skewed distribution considered by Fernández and Steel [1998] with the probability density function (pdf):

$$g(\varepsilon_i|\xi) = \frac{2}{\xi + \frac{1}{\xi}} \left\{ f\left(\frac{\varepsilon_i}{\xi}\right) I_{[0,\infty]}(\varepsilon_i) + f\left(\xi\varepsilon_i\right) I_{[-\infty,0]}(\varepsilon_i) \right\}, \tag{3.3}$$

where each error terms, $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d following the conditional pdf $g(\varepsilon_i|\xi)$ and $f(\cdot)$ is the pdf of the standard (univariate) normal distribution.

We consider fixing zero mean and unit variance for all skewed normal error cases. When $\xi < 1$, the distributions are negatively skewed. When $\xi > 1$, the distributions are positively skewed. When $\xi = 1$, this is exactly the standard normal distribution.

3.3.1. t-distributed errors

Inference for selected regression coefficients ($n > p$ case)

We provide two key takeaways from observations according to plot A (row 1) of Figure (3.1) and Table (3.1):

1. *(U, V) demonstrates good coverages under extremely heavy-tailed distributions*

   In the contrived case where we consider the t_1 distribution (Cauchy), we clearly see for both (U, V) and DS, their interval lengths are very long (mostly > 3) and behaves in a volatile way compared to all other settings of k. This is because the t_1 distribution is highly heavy-tailed and occasionally takes on extreme values (a Cauchy distribution with mean and variance are undefined). However, we see from Table (3.1) that (U, V) still yielded impressively close to 90% nominal intervals in most cases of $|\beta_i|$, whereas DS yielded anti-conservative intervals (approximately 80%) despite the interval lengths are longer.

2. *(U, V) is relatively insensitive towards t-distributed errors and dominates DS across all degrees of freedom considered*

   When $k = 2$ (distributional mean is 0 but variance is infinite), we see that the interval lengths narrow down significantly for both (U, V) and DS. Coverage probabilities improves for DS beyond $k \geq 2$, while (U, V) is relatively insensitive towards all ranges of k considered.

   Nonetheless, in heavy tail scenarios, (U, V) still maintains dominance over DS with lengths consistently 20% shorter across most magnitudes of $|\beta_i|$.

Inference for partial regression coefficients ($n < p$ case)

We provide two key takeaways from observations according to plot B (row 2) of Figure (3.1) and Table (3.2):

1. *Results largely consistent with $n > p$ case*

   We see that both (U, V) and DS show similar trends when compared across all values of k and $|\beta_i|$.

2. *(U, V) becomes less dominant over DS*

   In terms of median interval lengths, DS appears to stay relatively similar when compared with previous $n > p$ case, but (U, V) suffers overall increase by around 10% despite maintaining close to the 90% nominal level. While (U, V) remains on the upper-hand against DS, the higher dimensional setting has caused the disparity in interval lengths between (U, V) and DS closing down to approximately 10%.

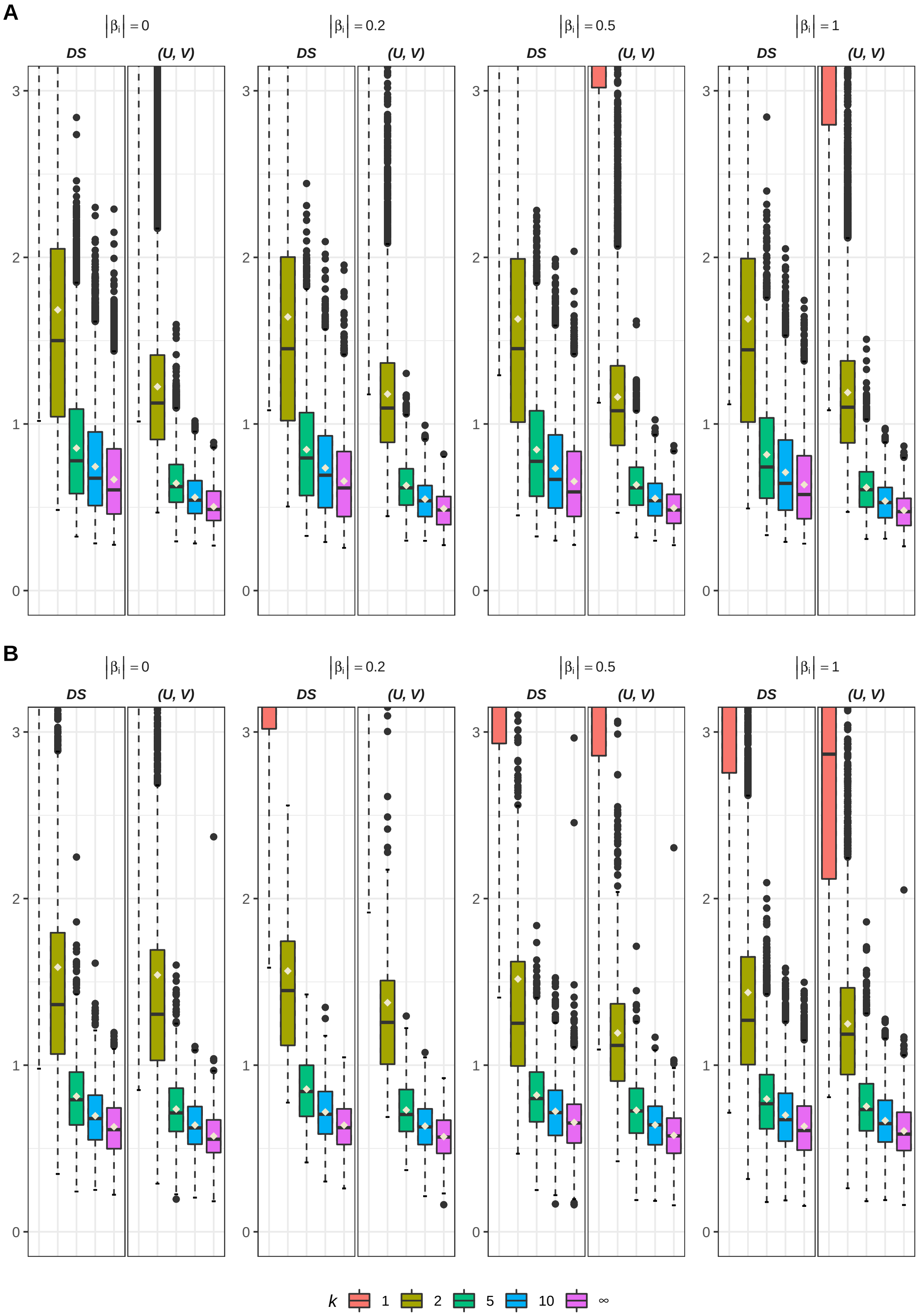

Figure 3.1.: Boxplots of lengths of confidence intervals generated by DS and (U, V) for the selected regression coefficients (Plot A of row 1) and partial regression coeffficients (Plot B of row 2) averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$ and errors are generated under t_k distribution with $k \in \{1, 2, 5, 10, \infty\}$. Lengths shown are not more than 3. Diamond symbol in each boxplot illustrates the mean interval length. Note that our outputs for plot B are categorized based on the absolute value of the selected regression coefficients, $|\beta_i|$ but the intervals computed are still for the partial regression coefficients.

(a) Data Splitting (DS)

	Degrees of freedom, k									
	1		**2**		**5**		**10**		∞	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	6.00	83.47	1.50	88.38	0.78	90.00	0.67	89.83	0.60	89.96
0.2	5.91	81.14	1.45	89.08	0.80	89.62	0.69	90.58	0.62	89.91
0.5	5.78	80.39	1.45	88.48	0.78	90.16	0.67	90.14	0.59	89.95
1.0	5.68	82.97	1.45	89.61	0.74	90.29	0.64	90.09	0.58	90.19

(b) (U, V) decomposition

	Degrees of freedom, k									
	1		**2**		**5**		**10**		∞	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	5.42	90.71	1.13	89.95	0.62	89.95	0.54	89.69	0.49	89.73
0.2	5.24	89.38	1.10	89.59	0.62	90.19	0.54	90.46	0.48	90.22
0.5	4.33	90.72	1.08	90.97	0.62	90.18	0.54	90.18	0.48	89.93
1.0	3.94	89.84	1.10	90.34	0.61	90.22	0.53	90.18	0.47	90.17

Table 3.1.: Reported empirical median confidence interval lengths and mean coverage probabilities generated by (a) DS, and (b) (U, V) for the selected regression coefficients averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$ and errors are generated under t_k distribution with $k \in \{1, 2, 5, 10, \infty\}$.

(a) Data Splitting (DS)

	Degrees of freedom, k									
	1		**2**		**5**		**10**		∞	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	6.65	88.16	1.36	88.51	0.79	89.41	0.68	89.35	0.61	90.06
0.2	4.25	83.78	1.45	93.51	0.84	90.16	0.71	89.35	0.62	85.48
0.5	3.93	85.45	1.25	87.16	0.80	88.81	0.72	88.91	0.65	89.95
1.0	4.01	82.83	1.27	87.57	0.77	89.75	0.67	89.12	0.61	89.23

(b) (U, V) decomposition

	Degrees of freedom, k									
	1		**2**		**5**		**10**		∞	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	6.40	88.95	1.31	88.34	0.71	88.31	0.62	86.98	0.56	87.91
0.2	5.66	89.39	1.26	94.96	0.70	90.38	0.63	85.60	0.57	85.27
0.5	4.07	85.00	1.12	89.36	0.72	90.79	0.64	89.44	0.58	90.50
1.0	2.87	95.04	1.19	89.89	0.73	89.77	0.65	90.03	0.59	89.61

Table 3.2.: Reported empirical median confidence interval lengths and mean coverage probabilities generated by (a) DS, and (b) (U, V) for the partial regression coefficients averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$ and errors are generated under t_k distribution with $k \in \{1, 2, 5, 10, \infty\}$.

3.3.2. Skewed normal errors

Inference for selected regression coefficients ($n > p$ case)

We provide two key takeaways from observations according to plot A (row 1) of Figure (3.2) and Table (3.3):

1. *No influence of skewness towards coverage properties*

   Both methods yield intervals that are robust across all values of ξ and maintains close to 90% nominal coverage.

2. *(U, V) dominates over DS*

   Regardless of the values of ξ and $|\beta_i|$, we can clearly see that (U, V) always yield 20% shorter interval lengths than DS. Quantitatively, (U, V) consistently produces median interval lengths around 0.48, while DS is around 0.60.

   Agreeing with the simulation results from t-distributed errors considered in the previous section, DS produces much greater variability in interval lengths with significantly many more extreme outliers recorded (> 1.5), but (U, V) is relatively more stable.

Inference for partial regression coefficients ($n < p$ case)

We provide two key takeaways from observations according to plot B (row 2) of Figure (3.2) and Table (3.4):

1. *Results largely consistent with $n > p$ case*

   The skewness values, ξ considered have minimal impact on the interval lengths for both splitting strategies. The interval lengths are still fairly consistent.

2. *(U, V) reports signs of coverage miscalibration*

   We discover that (U, V) yields lower coverage probabilities, particularly where $|\beta_i| \in 0, 0.2\}$ for all values of ξ, including the standard normal case ($\xi = 1$). We remark that this is inconsistent with the conclusion by Rasines and Young [2021], since the number of replications used $B = 2000$ is lower than in their choice ($B = 5000$).

   In contrast, DS remains close to 90% nominal coverage.

3. *(U, V) not far better off than DS*

   Once again, consistent with the findings in t-distributed error case, it appears that the overall disparity becomes narrower between (U, V) and DS, this time even less than 10%, especially when $|\beta_i| \in \{0.5, 1.0\}$.

Overall conclusions

We now provide our overall findings with emphasis on the performance of (U, V) decomposition for we generate non-normal errors. Overall, our simulation results suggest that:

1. (U, V) appears to show signs of weakening in inferential power, especially when we consider the high-dimension setting $(n < p)$, when compared against DS.
2. In terms of confidence interval lengths, (U, V) show lesser dominance over DS, (about 10%) when $n < p$ than about 20% when $n > p$ when considering t-distributed and skewed normal errors.
3. (U, V) show evidence of poorer coverages when we consider $n < p$ for small $|\beta_i|$ regardless on the skewness in the generalized skewed normal distribution.
4. (U, V) still maintained reasonably good coverage probabilities up to very heavy tailed distribution $(k > 2)$ sharing similar performance impact as with DS.
5. Despite the slight shortcomings, (U, V) seem to possess fairly excellent coverage properties that are robust to model violations, and still outperforms DS in various parameters considered.

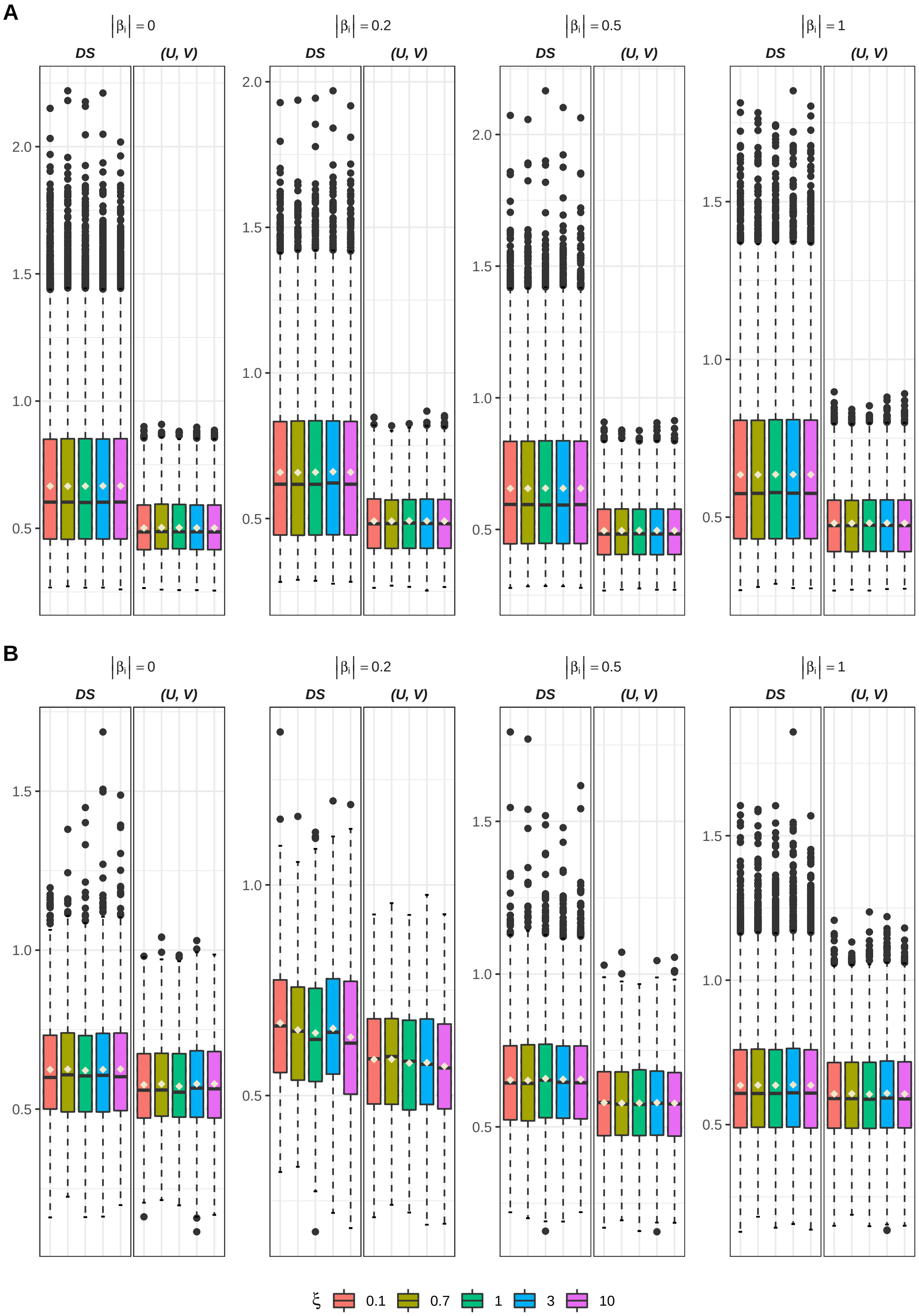

Figure 3.2.: Boxplots of lengths of confidence intervals generated by DS and (U, V) for the selected regression coefficients (Plot A of row 1) and partial regression coeffficients (Plot B of row 2) averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$ and errors generated under skewed normal distribution with skewness parameter, $\xi \in \{0.1, 0.7, 1, 5, 10\}$. Diamond symbol in each boxplot illustrates the mean interval length. Note that our outputs for plot B are categorized based on the absolute value of the selected regression coefficients, $|\beta_i|$ but the intervals computed are still for the partial regression coefficients.

(a) Data Splitting (DS)

	Skewness, ξ									
	0.1		**0.7**		**1**		**5**		**10**	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	0.60	89.84	0.60	89.87	0.60	90.22	0.60	90.02	0.60	90.00
0.2	0.62	90.28	0.62	89.56	0.62	90.81	0.62	89.67	0.62	89.60
0.5	0.60	90.47	0.59	89.73	0.59	89.51	0.59	90.19	0.59	90.57
1.0	0.57	89.19	0.58	89.64	0.58	89.74	0.58	89.65	0.58	89.19

(b) (U, V) decomposition

	Skewness, ξ									
	0.1		**0.7**		**1**		**5**		**10**	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	0.49	89.97	0.49	89.96	0.49	89.95	0.49	90.18	0.49	89.89
0.2	0.48	90.48	0.48	90.08	0.48	90.31	0.48	90.15	0.48	90.08
0.5	0.48	90.08	0.48	89.85	0.48	89.81	0.48	90.20	0.48	90.57
1.0	0.47	90.28	0.47	90.31	0.47	90.19	0.47	89.74	0.47	89.75

Table 3.3.: Reported empirical median confidence interval lengths and mean coverage probabilities generated by (a) DS and (b) (U, V) for the selected regression coefficients averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$ and errors are generated under skewed normal distribution with skewness parameter, $\xi \in \{0.1, 0.7, 1, 5, 10\}$.

(a) Data Splitting (DS)

	Skewness, ξ									
	0.1		**0.7**		**1**		**5**		**10**	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	0.60	89.26	0.61	89.43	0.60	89.43	0.61	89.50	0.60	89.03
0.2	0.67	92.46	0.65	87.98	0.63	88.26	0.65	89.82	0.62	87.85
0.5	0.64	88.18	0.64	88.79	0.65	88.33	0.65	88.32	0.64	88.61
1.0	0.61	89.59	0.61	89.58	0.61	89.05	0.61	89.93	0.61	89.63

(b) (U, V) decomposition

	Skewness, ξ									
	0.1		**0.7**		**1**		**5**		**10**	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	0.56	86.37	0.56	87.27	0.55	86.50	0.57	87.92	0.56	88.07
0.2	0.59	88.57	0.59	88.59	0.58	87.74	0.57	90.09	0.57	88.67
0.5	0.58	89.32	0.57	89.61	0.57	90.01	0.58	89.98	0.57	89.99
1.0	0.59	89.55	0.59	89.58	0.59	89.91	0.59	89.61	0.59	89.40

Table 3.4.: Reported empirical median confidence interval lengths and mean coverage probabilities generated by (a) DS and (b) (U, V) for the partial regression coefficients averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$ and errors are generated under skewed normal distribution with skewness parameter, $\xi \in \{0.1, 0.7, 1, 5, 10\}$.

3.4. Assuming known variance, σ^2

Rasines and Young [2021] show empirically that with $n = 200$ and $p = 30$, estimating the normal variance where the true $\sigma^2 = 1$, the resulting coverage probabilities of confidence intervals are very close to the nominal 90%, and the corresponding lengths are comparatively shorter than DS. However, the σ^2 considered is rather fairly small. What if σ^2 is larger? We also wish to assume that we know the true variance, σ^2 to investigate whether there are significant power gains for (U, V).

We switch back to the normal linear regression setting and examine whether instead of estimating σ^2, we fix σ^2, allowing the classical inferential theory to be applied. While we can be quite certain of coverage guarantees with fixed σ^2, our simulation results suggests that fixing σ^2 generally only leads to improvement in terms of the variability when it comes to the interval lengths, but comparatively similar in terms of means and medians.

3.4.1. Theory for confidence intervals

In this subsection, we wish to establish the form of confidence intervals for DS and (U, V) that is not fully accounted for in the original paper when σ^2 is known and unknown under both inference scenarios discussed before. We borrow our discussion in target formulation from Section 2.1 and 2.4 for DS and (U, V) decomposition respectively.

When making inference for the selected regression coefficients, the ordinary least squares (OLS) estimate for the selected regression coefficients for DS and (U, V) are:

$$\begin{aligned}\hat{\beta}_j^{DS}(X^{(2)}) &= e_j^T (X^{(2)T} X^{(2)})^{-1} X^{(2)T} Y^{(2)} \quad \forall j \in \hat{M}_{DS},\\ \hat{\beta}_j^{UV}(X) &= e_j^T (X^T X)^{-1} X^T V \quad \forall j \in \hat{M}_{UV},\end{aligned}$$

where the subscripts under $\hat{M}$ are to emphasize the model selected by DS and (U, V), which are in general not similar.

When σ^2 is known, then both coefficient estimates for DS and (U, V) are normal:

$$\begin{aligned}\hat{\beta}_j^{DS}(X^{(2)}) &\sim \mathcal{N}(\beta_j^{DS}(X^{(2)}), \sigma^2 e_j^T (X^{(2)T} X^{(2)})^{-1} e_j) \quad \forall j \in \hat{M}_{DS},\\ \hat{\beta}_j^{UV}(X) &\sim \mathcal{N}(\beta_j^{UV}(X), \sigma^2 (1+\gamma^{-2}) e_j^T (X^T X)^{-1} e_j) \quad \forall j \in \hat{M}_{UV},\end{aligned}$$

and therefore their valid $100(1-\alpha)\%$ confidence intervals are computed from:

$$\begin{aligned}\hat{\beta}_j^{DS}(X^{(2)}) &\mp z_{1-\frac{\alpha}{2}} \sigma \sqrt{e_j^T (X^{(2)T} X^{(2)})^{-1} e_j} \quad \forall j \in \hat{M}_{DS},\\ \hat{\beta}_j^{UV}(X) &\mp z_{1-\frac{\alpha}{2}} \sigma \sqrt{(1+\gamma^{-2}) e_j^T (X^T X)^{-1} e_j} \quad \forall j \in \hat{M}_{UV},\end{aligned}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution.

When σ^2 is unknown, then, for $n > p$, we have:

$$\begin{aligned}\hat{\beta}_j^{DS}(X^{(2)}) &\mp t_{n-p,1-\frac{\alpha}{2}} \hat{\sigma}_{LD} \sqrt{e_j^T (X^{(2)T} X^{(2)})^{-1} e_j} \quad \forall j \in \hat{M}_{DS},\\ \hat{\beta}_j^{UV}(X) &\mp t_{n-p,1-\frac{\alpha}{2}} \hat{\sigma}_{LD} \sqrt{(1+\gamma^{-2}) e_j^T (X^T X)^{-1} e_j} \quad \forall j \in \hat{M}_{UV},\end{aligned}$$

where $t_{n-p,1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile from a t-distribution with $n-p$ degrees of freedom. Notice that we require to use the same estimates, $\hat{\sigma}_{LD}$ in both splitting approaches.

For inferring the partial regression coefficients, it is trivial to show that both DS and (U, V) are normal as well when σ^2 is known.

If we have $n < p$ and σ^2 is unknown, then their corresponding confidence intervals are:

$$\hat{\beta}^{DS}_{j,M_{DS}}(X^{(2)}) \mp t_{n-p_{\hat{\lambda}},1-\frac{\alpha}{2}}\hat{\sigma}_{HD}\sqrt{e^T_{j,M_{DS}}(X^{(2)}_{M_{DS}}{}^T X^{(2)}_{M_{DS}})^{-1}e_{j,M_{DS}}} \quad \forall j \in \hat{M}_{DS}, \hat{M}_{DS} = M_{DS},$$

$$\hat{\beta}^{UV}_{j,M_{UV}}(X) \mp t_{n-p_{\hat{\lambda}},1-\frac{\alpha}{2}}\hat{\sigma}_{HD}\sqrt{(1+\gamma^{-2})e^T_{j,M_{UV}}(X^T_{M_{UV}}X_{M_{UV}})^{-1}e_{j,M_{UV}}}$$
$$\forall j \in \hat{M}_{UV}, \hat{M}_{UV} = M_{UV},$$

where $e_{j,M} \in \mathbb{R}^{|M|}$ has all 0's but 1 at the j^{th} component of model M. The distributions for the estimates are both now following $t_{n-p_{\hat{\lambda}}}$, where the degrees of freedom (subscript) now varies depending on the estimated model selected by LASSO when estimating $\hat{\sigma}^2_{HD}$. Notice that the degrees of freedom is not dependent on the model selected, $\hat{M}$, where we implement the stability selection paired with LASSO.

With the theory behind constructing the confidence intervals for DS and (U, V), we are now in a position to evaluate their performances. We discuss our findings in the next subsection.

3.4.2. High error variance

We consider $\sigma^2 = 5^2$.

Inference for selected regression coefficients ($n > p$ case)

We provide two key takeaways from observations according to plot A (row 1) of Figure (3.3) and Table (3.5):

1. *Indistinguishable improvement in coverage properties with known* σ^2

   Even by using $\hat{\sigma}^2_{LD}$, it turns out that (U, V) generates roughly similar interval lengths as it were using the true σ^2. This also applies to DS. The variability of lengths produced are smaller for the case of using σ^2, implying greater certainty of coefficient estimates, but the differences between two methods are close to unnoticeable. As expected, using the true σ^2 allows us to produce coverages at approximately nominal 90% but so did using $\hat{\sigma}^2_{LD}$.

2. (U, V) *continues to dominate DS*

   (U, V) consistently reports 25% shorter and lower length variability compared to DS overall, regardless of $|\beta_i|$.

Inference for partial regression coefficients ($n < p$ case)

We provide two key takeaways from observations according to plot B (row 2) of Figure (3.3) and Table (3.5):

1. *An improvement in interval lengths variability*

   We see that using the true σ^2 demonstrates noticeable improvement in the high-dimensional setting while maintaining the desired coverage probabilities. This can be partly attributed to less accurate estimation of σ^2 using $\hat{\sigma}^2_{HD}$. (Note: Surprisingly in the case of $|\beta_i| = 0.2$, we see the emergence of miscalibration (83.57%) for (U, V), possibly due to Monte-Carlo estimation variability. On this premise, we can also see that using $\hat{\sigma}^2_{HD}$ suffers worse (80.77%) for (U, V)). In contrast, while we benefit from shorter interval length ranges when using σ^2, we seem to record higher median lengths.

2. (U, V) *becomes less dominant over DS*

   Once again, regardless of whether we use σ^2 or $\hat{\sigma}^2_{HD}$ for confidence interval estimates, we see the impact of high-dimensionality on the weakening of (U, V): overall length generally increases by 12% for all $|\beta_i|$. The impact is substantial when $|\beta_i|$ is large. Meanwhile, DS records more or less similar lengths.

3.4.3. Low error variance

We consider $\sigma^2 = 1$.

Inference for selected regression coefficients ($n > p$ case)

We provide two key takeaways from observations according to plot A (row 1) of Figure (3.4) and Table (3.6):

1. *Results largely consistent with* $n > p$ *and* $\sigma^2 = 5^2$ *case*

   Our findings are similar as when we set $\sigma^2 = 5^2$. We also see that using the actual σ^2 does not provide significant improvement as we thought would be the case. Owing to simulating smaller $\sigma^2 = 1$, we observe considerably shorter lengths (as much as 41%) for both (U, V) and *DS* regardless of whether we provide σ^2 or $\hat{\sigma}^2_{LD}$.

2. (U, V) *still dominates DS*

   Despite using lower σ^2, DS still frequently produces wide range of interval lengths estimate while (U, V) generally produces far more stable interval lengths. As for coverage probabilities, no issues are detected as all estimates are accurate to 90% nominal levels.

Inference for partial regression coefficients ($n < p$ case)

We provide two key takeaways from observations according to plot B (row 2) of Figure (3.4) and Table (3.6):

1. *Results largely consistent with* $n > p$ *and* $\sigma^2 = 5^2$ *case*

   Our findings are similar as when we set $\sigma^2 = 5^2$ case. As σ^2 is fixed at its true value, we can see that the coverages are closer to 90% nominal level than using $\hat{\sigma}^2_{HD}$, but the impact on coverage between using σ^2 and σ^2_{HD} is practically insignificant.

2. *(U, V) becomes less dominant over DS*

   One key noticeable difference now is that DS produces far less variability in interval lengths while still remaining close to the same lengths as in the $n > p$ case. It appears that with the empirical evidence so far, DS, as a favourably least stringent method can be useful for post-selection inference in high-dimensional settings. In contrast, (U, V) seem to be once again negatively impacted by the high dimensional setting, which increases the overall length and distorts its coverage probabilities. After all, (U, V) still remains dominant against DS overall.

Overall conclusions

Our simulation results suggest that:

1. (U, V) appears to enjoy oracle properties especially when $n > p$, as it seems to perform almost as well as if the true σ^2 is assumed known in advance.
2. The value of σ^2 does not impact the coverage probabilities, only the interval lengths. Estimates are closer to the truth when true σ^2 is small.
3. Once again, we see evidence that while (U, V) still remains comparatively dominant over DS in all considered cases, the dominance becomes less noticeable as we reach higher dimensional settings.

Remark

As a caveat, we note that we only tested a limited set of conditions and that each of our outputs shown is based on only a fixed set of pseudo-random numbers for replicability purposes. Given the results shown, we could have imposed even higher dimension settings ($p \gg n$) and investigated whether (U, V) still maintains robustness in coverage properties. One can also consider a combination of a variety of selective inference methods other than DS to expand the findings.

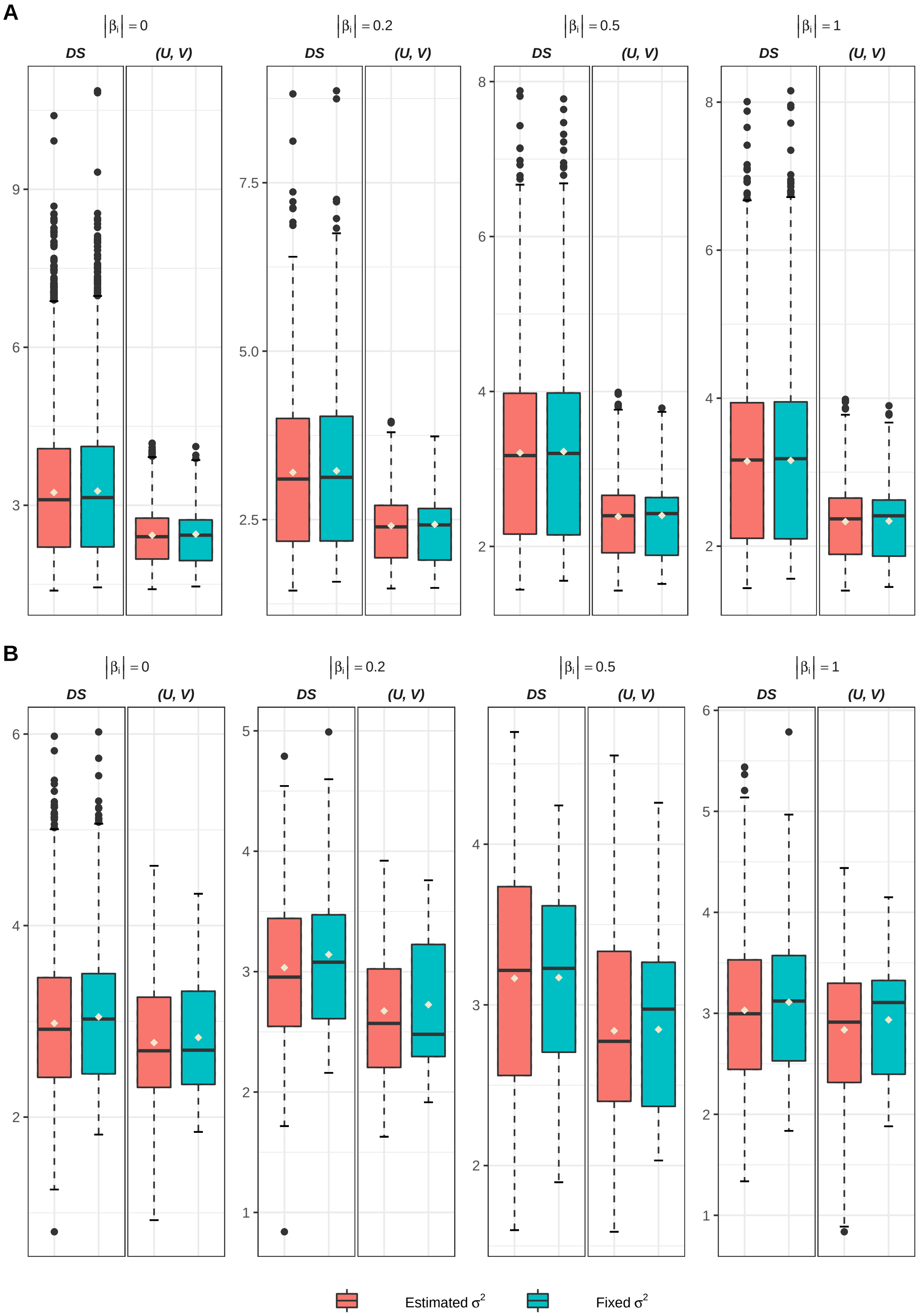

Figure 3.3.: Boxplots of lengths of confidence intervals generated by DS and (U, V) for the selected regression coefficients (Plot A of row 1) and partial regression coefficients (Plot B of row 2) using estimated σ^2 when σ^2 is assumed unknown, and actual σ^2 when σ^2 is assumed known. We set $\sigma^2 = 5^2$. Results are averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$. Diamond symbol in each boxplot illustrates the mean interval length.

(a) Data Splitting (DS)

	Selected regression coefficients				Partial regression coefficients			
	Estimated σ^2		**Fixed** σ^2		**Estimated** σ^2		**Fixed** σ^2	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	3.11	89.99	3.15	90.15	2.92	89.53	3.02	90.47
0.2	3.10	89.17	3.13	89.40	2.95	82.69	3.08	82.69
0.5	3.17	89.62	3.20	89.86	3.21	92.39	3.23	92.39
1.0	3.16	89.57	3.18	89.58	3.00	88.77	3.12	89.84

(b) (U, V) decomposition

	Selected regression coefficients				Partial regression coefficients			
	Estimated σ^2		**Fixed** σ^2		**Estimated** σ^2		**Fixed** σ^2	
$\lvert\beta_i\rvert$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	2.40	89.74	2.43	89.95	2.69	87.89	2.70	89.97
0.2	2.39	90.00	2.42	90.09	2.57	80.77	2.48	83.56
0.5	2.40	90.40	2.42	90.40	2.77	89.22	2.97	92.12
1.0	2.37	89.70	2.41	89.41	2.91	88.93	3.11	90.05

Table 3.5.: Reported empirical median confidence interval lengths and mean coverage probabilities generated by (a) DS and (b) (U, V) for the selected regression coefficients and partial regression coefficients using estimated σ^2 when σ^2 is assumed unknown, and actual σ^2 when σ^2 is assumed known. We set $\sigma^2 = 5^2$. Results are averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$.

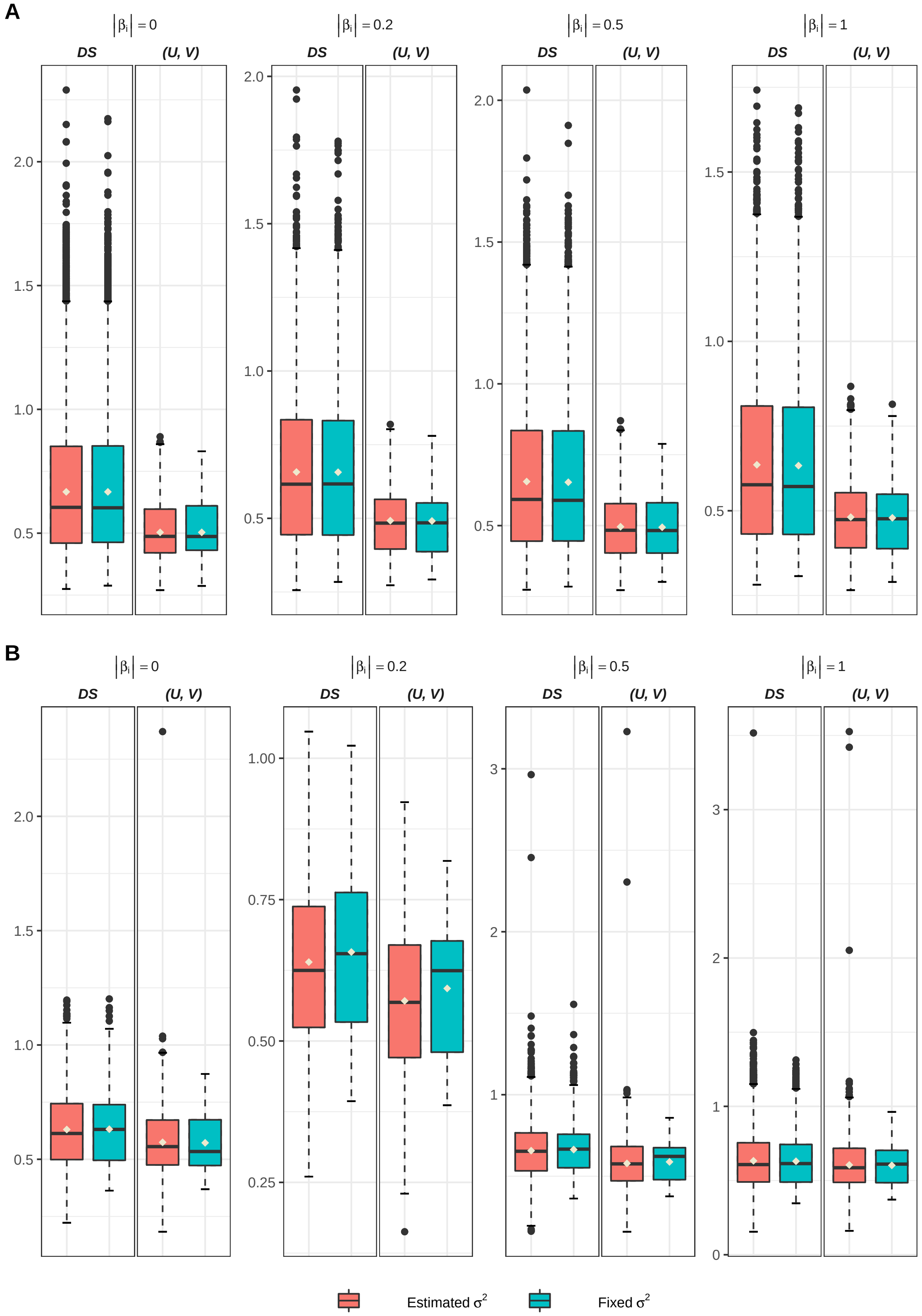

Figure 3.4.: Boxplots of lengths of confidence intervals generated by DS and (U, V) for the selected regression coefficients (Plot A of row 1) and partial regression coefficients (Plot B of row 2) using estimated σ^2 when σ^2 is assumed unknown, and actual σ^2 when σ^2 is assumed known. We set $\sigma^2 = 1$. Results are averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$. Diamond symbol in each boxplot illustrates the mean interval length.

(a) Data Splitting (DS)

	Selected regression coefficients				Partial regression coefficients			
	Estimated σ^2		**Fixed** σ^2		**Estimated** σ^2		**Fixed** σ^2	
$\|\beta_i\|$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	0.60	89.96	0.60	89.95	0.61	90.06	0.63	89.94
0.2	0.62	89.91	0.62	90.05	0.62	85.48	0.65	87.90
0.5	0.59	89.95	0.59	89.73	0.65	89.95	0.67	90.86
1.0	0.58	90.19	0.57	90.08	0.61	89.23	0.61	89.89

(b) (U, V) decomposition

	Selected regression coefficients				Partial regression coefficients			
	Estimated σ^2		**Fixed** σ^2		**Estimated** σ^2		**Fixed** σ^2	
$\|\beta_i\|$	Length	Coverage	Length	Coverage	Length	Coverage	Length	Coverage
0.0	0.49	89.73	0.49	89.77	0.56	87.91	0.53	89.39
0.2	0.48	90.22	0.48	90.21	0.57	85.27	0.62	87.24
0.5	0.48	89.93	0.48	89.8	0.58	90.50	0.62	90.79
1.0	0.47	90.17	0.48	90.07	0.59	89.61	0.61	89.79

Table 3.6.: Reported empirical median confidence interval lengths and mean coverage probabilities generated by (a) DS and (b) (U, V) for the selected regression coefficients and partial regression coefficients using estimated σ^2 when σ^2 is assumed unknown, and actual σ^2 when σ^2 is assumed known. We set $\sigma^2 = 1$. Results are averaged across all combinations of simulation parameters, f and ρ. Lengths are compared across coefficients where $|\beta_i| \in \{0, 0.2, 0.5, 1.0\}$.

4. MC construction of conditional selective inference with LASSO

In Section 2.3, we have discussed making inference for the partial regression coefficients after model selection by LASSO where the selection event is convex. We highlighted that despite the convenient use of the analytical characterization (2.8) to our inferential objectives, it is practically concerning that (2.8) leads us to losing substantial inferential power with imprecise confidence intervals and coverage problems. This is because (2.8) conditions on the signs of the active coefficients, $\{\hat{S} = s\}$, which geometrically implies restricting y into a single polyhedron. Following the principal of conditioning "more" by Fithian et al. [2014], it is not surprising at all, that we suffer the aforementioned consequences.

In fact, Lee et al. [2016] also mentions conditioning only the model selected, $\{j \in \hat{M}, \hat{M} = M\}$, which considers the union of all possible signs of the active coefficients. Following Fithian et al. [2014], the selection event now is less "refined", so we are essentially conditioning "less" and this provides greater inferential power. Mathematically, we can reiterate (2.8) by writing the normal distribution truncated to the union of intervals $\bigcup\limits_s [\mathcal{V}_s^-(z), \mathcal{V}_s^+(z)]$:

$$\eta_{j,M}^T Y | \{j \in \hat{M}, \hat{M} = M, (I_n - P_{\eta_{j,M}})Y = z\} \sim \mathcal{TN}\left(\eta_{j,M}^T \mu, \sigma^2 \left\|\eta_{j,M}\right\|_2^2, \bigcup_s [\mathcal{V}_s^-(z), \mathcal{V}_s^+(z)]\right). \tag{4.1}$$

Similar to Section 2.3, to form the $100(1-\alpha)\%$ confidence intervals for $\eta_{j,M}^T \mu$, we require to find the lower and upper confidence limits, L and U by solving the truncated normal statistic (pivot):

$$F_{L,\sigma^2 \|\eta_{j,M}\|_2^2}^{\cup_s [\mathcal{V}_s^+(z), \mathcal{V}_s^-(z))]}(\eta_{j,M}^T y) = 1 - \frac{\alpha}{2}, \tag{4.2}$$

$$F_{U,\sigma^2 \|\eta_{j,M}\|_2^2}^{\cup_s [\mathcal{V}_s^+(z), \mathcal{V}_s^-(z))]}(\eta_{j,M}^T y) = \frac{\alpha}{2}. \tag{4.3}$$

Lee et al. [2016] mentions that (4.1) is intractable due to difficulties in expressing the form of appropriate end points for $\bigcup\limits_s [\mathcal{V}_s^-(z), \mathcal{V}_s^+(z)]$ for all $s \in \{-1, 1\}^{|M|}$ explicitly, especially when $|M|$ is large. Therefore, a natural remedy to this is by resorting to sampling procedures.

In this section, we advocate the approach of minimizing the conditioning event to $\{j \in \hat{M}, \hat{M} = M\}$. We will not attempt to derive an exact form for the pivots in (4.2) and (4.3), but we hope to propose a proof-of-concept "conditional sampling" strategy to (4.1) and making use of Monte-Carlo approximation to the confidence limits satisfying (4.2) and (4.3). Following the benefits of randomization discussed in Section 2.4, we also provide a very convenient extension to (2.12) with our method.

While it is possible to approximate the conditioning additionally on the signs of the active coefficients, $\{\hat{S} = s\}$, we do not think this is necessary in pursuit of greater inferential precision. Further, conditioning on $\{\hat{S} = s\}$ involves extra computational power for eliminating the active variables corresponding to the coefficients that does not meet the signs in our algorithm. In that case, we might as well revert our approach to considering the original tractable form in (2.8) since it is much more computationally efficient. The *selectiveInference* package for implementing this is readily available in **R** as *fixedLassoInf* (Lee et al. [2016] and Taylor and Tibshirani [2018]).

As a new proposal, it is imperative that we should critically assess its performance. Along with showcasing our approach, we implement evaluation studies and make appropriate comparisons with other methods including Lee et al. [2016] (conditioning on $\{j \in \hat{M}, \hat{M} = M, \hat{S} = s\}$), DS and (U, V). We limit our focus on carrying out inference for partial regression coefficients in terms of interval lengths and coverages.

4.1. Conditional MC rejection sampling method

Our approach is similar to the conditional Monte-Carlo (MC) re-sampling theory which conditions on the sufficient statistics Lindqvist and Taraldsen [2007]. A general sampling framework is considered by Garcia-Angulo and Claeskens [2022] for the case of post-selection inference. Our algorithm serves as an application for the case of the LASSO selection events through the polyhedral method by Lee et al. [2016].

As before, we assume our observed data follows the normal linear model (2.1). The procedure is as follows:

Step 1: For a given observed data, $(X_{obs}, y_{obs}) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$:

Carry out variable selection with LASSO at a fixed penalty parameter, λ, obtaining $(X_{\hat{M}_{obs}}, y_{obs}) \in \mathbb{R}^{n \times |\hat{M}_{obs}|} \times \mathbb{R}^n$ where $\hat{M}_{obs} = \{j : \hat{\beta}_{j,\hat{M}_{obs}} \neq 0\}$, $j \subseteq \{1, \dots, p\}$.

Step 2: Compute $\eta_{\hat{M}_{obs}} = X_{\hat{M}_{obs}}(X^T_{\hat{M}_{obs}} X_{\hat{M}_{obs}})^{-1} \in \mathbb{R}^{n \times |\hat{M}_{obs}|}$.

Substitute to obtain observed statistics for all selected variables: $\hat{\beta}_{\hat{M}_{obs}} \equiv \eta^T_{\hat{M}_{obs}} y_{obs} \in \mathbb{R}^{|\hat{M}_{obs}|}$.

Step 3: Inference for $\beta_{j,\hat{M}_{obs}}$:

Simulating (4.1) *and approximating* $100(1-\alpha)\%$ *confidence interval for* $\beta_{j,\hat{M}_{obs}}$, (L_{obs}, U_{obs}), *approximated by* (L_{est}, U_{est}).

For all $j \in \hat{M}_{obs}$ do

(a) Compute $z_{obs} = (I_n - P_{\eta_{j,\hat{M}_{obs}}})y_{obs}$ where $P_{\eta_{j,\hat{M}_{obs}}} = \frac{\eta_{j,\hat{M}_{obs}} \eta^T_{j,\hat{M}_{obs}}}{\left\|\eta_{j,\hat{M}_{obs}}\right\|_2^2}$, and $\eta_{j,\hat{M}_{obs}}$ is the j^{th} column of $\eta_{\hat{M}_{obs}}$.

(b) Propose an initial estimate for the lower limit, L_0 for the confidence interval.

(c) Obtain B samples of condition on selection distribution (4.1).

```
For $b = 1, 2, \dots, B$ do
  Set "condition" = FALSE.
  While "condition" = FALSE do
  (i) Obtain a simulated sample of $\hat{\beta}^{sim_b}_{j,\hat{M}_{obs}} = \eta^T_{j,\hat{M}_{obs}} y_{sim}$ from
      $\mathcal{N}\left(L_0, \sigma^2 \left\|\eta_{j,\hat{M}_{obs}}\right\|_2^2\right)$, the unconditional normal distribution, where
      $y_{sim} \sim \mathcal{N}(\mu, \sigma^2)$, $\mu$ is unknown and $\sigma^2$ is assumed known.
      Otherwise, plug-in the estimate, $\hat{\sigma}^2_{LD}$ (3.1) when $n > p$ or $\hat{\sigma}^2_{HD}$ (3.2)
      when $n < p$.

  (ii) Solve the conditioning on the residual constraint,
      $\{(I_n - P_{\eta_{j,\hat{M}_{obs}}})Y_{sim} = z_{obs}\}$:
      After observing $\hat{\beta}^{sim_b}_{j,\hat{M}_{obs}}$ in (i), we have:

          $y_{sim} = \frac{\eta_{j,\hat{M}_{obs}} \hat{\beta}^{sim_b}_{j,\hat{M}_{obs}}}{\left\|\eta_{j,\hat{M}_{obs}}\right\|_2^2} + z_{obs}.$

  (iii) Solve the conditioning on the selected model constraint,
      $\{j \in \hat{M}_{sim}, \hat{M}_{sim} = \hat{M}_{obs}\}$:
      Carry out LASSO on $(X_{obs}, y_{sim})$ to find
      $\hat{M}_{sim} = \{j_{sim} : \hat{\beta}^{sim}_{j_{sim},\hat{M}_{sim}} \neq 0\}$.

  (iv) If both constraints in (ii) and (iii) are met, then set
      "condition" = TRUE. Otherwise, we reject $\hat{\beta}^{sim_b}_{j_{sim}}$.
    EndWhile
    Record $\hat{\beta}^{sim_b}_{j_{sim}}$.
 EndFor
 Output: $\{\hat{\beta}^{sim_b}_{j,\hat{M}_{obs}}\}_{b=1}^B \in \mathbb{R}^B$.
```

These are the samples from the simulated condition on selection distribution:

$$\hat{\beta}^{sim}_{j,\hat{M}_{obs}} | \{j \in \hat{M}_{sim}, \hat{M}_{sim} = \hat{M}_{obs}, (I_n - P_{\eta_{j,\hat{M}_{obs}}})Y_{sim} = z_{obs}\}. \tag{4.4}$$

(d) Compute the conditional MC rejection sampling estimate for the requirement in (4.2):

$$F^{MC,\cup_s[\mathcal{V}_s^+(z),\mathcal{V}_s^-(z))]}_{L_0,\sigma^2\|\eta_{j,\hat{M}}\|_2^2}(\eta^T_{j,\hat{M}_{obs}} y_{obs}) \approx \frac{\sum_{b=1}^B I\left(\hat{\beta}^{sim_b}_{j,\hat{M}_{obs}} < \hat{\beta}^{obs}_{j,\hat{M}_{obs}}\right)}{B}. \tag{4.5}$$

(e) Perform bisection search until (4.5) is equivalent or approximately close to $1 - \frac{\alpha}{2}$, (see Appendix (1.1)). Report for L_{est}.

Step 4: Repeat **Step 3** for finding U_{est} by replacing L_0 with U_0 in **Step 3(b)**. Replace the indicator function in (4.5) to $I\left(\hat{\beta}^{sim_b}_{j,\hat{M}_{obs}} > \hat{\beta}^{obs}_{j,\hat{M}_{obs}}\right)$ for equivalent approximation to $1 - \frac{\alpha}{2}$ (see Appendix (1.2) for full implementation).

Randomization

Our MC method provides an easy extension to incorporating randomization to the responses which approximates the analytically intractable conditional on selection distribution (2.12). The algorithmic procedures are largely similar with the exception that we incorporate a randomized noise vector, $W_{MC} \sim \mathbb{Q}$ to y_{sim} in **Step 3(c)(iii)** before carrying out LASSO for model selection, where $\mathbb{Q}$ is a specified noise distribution. Notice that we do not replace the constraint in **Step 3(c)(ii)** to $\{(I_n - P_{\eta_{j,\hat{M}_{obs}}})(Y_{sim} + W_{MC}) = z_{obs}\}$ so that we respect the sufficiency principle necessary for the conditional MC, i.e., we do not want to perturb the form of z_{obs} which is the sufficient statistic for the $(n - |\hat{M}_{obs}|)$ nuisance parameters, $(I_n - P_{\eta_{j,\hat{M}_{obs}}})\mu$.

4.2. Preliminary evaluation study

4.2.1. Simulation setup

We demonstrate our proposed numerical procedure in Section 4.1, along with empirical assessment of its performance based on interval lengths and coverages. We compare with the analytical approach (conditioning on $\{j \in \hat{M}, \hat{M} = M, \hat{S} = s\}$) in (2.8). To supplement our comparison, we also involved DS and (U, V) methods considered beforehand.

We focus our experiment on the low-dimensional case $(n > p)$, leaving evaluation on the high-dimensional case for future work. Aligning with the framework by Lee et al. [2016], we limit ourselves to conducting inference for the partial regression coefficients based on the submodel selected by LASSO.

We now describe the parameter configurations for the experiment. We consider the normal linear model setup. We set $n = 100, p = 3$ and $\sigma^2 = 1$. The design matrix, X is set to be lightly correlated $(\rho = 0.1)$ with the Toeplitz matrix form considered in Section 3.1. For the true coefficent vector, we set $\beta = \{1, 0.8, 0.2\}$. This allows us to shed some light on the performance of our method when variables corresponding to strong and weak effects are selected. For DS and (U, V), we consider setting $f = \frac{1}{2}$. The randomized noise for (U, V) is $W_{UV} \sim \mathcal{N}(0, 1)$. As before, we set the confidence interval nominal levels at 90%.

We consider constraining ourselves to $R = 200$ replications of the data pair, (X, Y). For each $r \in \{1, \ldots, R\}$, our data is fit into LASSO. Recall that the framework by Lee et al. [2016] requires a deterministic LASSO selection event (using a fixed penalty, λ), for characterizing the distributions in (2.8), (2.12) and (4.1). We adhere to providing fixed λ by considering the universal penalty parameter, λ_{univ} proposed by Liu et al. [2018], where $\lambda_{univ} = \sqrt{\frac{2log(p)}{n}}$ for LASSO. For the case of DS, we replace n with n_1, where $n_1 \subset \{1, \ldots, n\}$ and $|n_1| = \frac{n}{2}$. This setting is independent of data (X, Y), and provides a deterministic way to carry out fixed-λ inference for conditional MC. Liu et al. [2018] mentioned that using λ_{univ} results in consistent variables selected by LASSO, therefore, lowering our target formation costs.

As for the conditional MC, we consider obtaining a fixed $B = 100$ samples of $\hat{\beta}^{sim}_{j,\hat{M}_{obs}}$ for characterizing (4.5). For the randomized case, we consider examining the effect of injecting normal noise, $W_{MC} \sim \mathcal{N}(0, 2.5^2\hat{\sigma}^2)$ into y_{sim}. The choice of the variance is arbitrary and is independent to W_{UV}.

4.2.2. Results

We remark that the performance of the MC method is dependent on B and R. Therefore, we provide cautious interpretation for our simulation results. Figure (4.1) depicts the boxplots of intervals lengths produced by each method considered so far. Specifically, the selective inference methods compared are: Lee's which conditions on M and s ($Lee_{M,s}$), conditional MC which conditions on M only (MC_M), conditional MC with randomized responses ($MC_{M,r+}$), DS, and (U, V).

Aligning with the presentation of results in Section 3, we examine the interval lengths output grouped by making inference for the partial regression coefficients of the selected variables that corresponds to true large and small coefficients. The numerical results are displayed in Table (4.1). We point out two key takeaways:

1. *Lee's method reports unreliable interval lengths and coverages while MC methods reveal considerable improvements*

   We can see that the MC methods ($MC_M, MC_{M,r+}$) which conditions only on $\{j \in \hat{M}, \hat{M} = M\}$ produces interval lengths that are considerably shorter than "Lee's method" ($Lee_{M,s}$) which conditions on both $\{j \in \hat{M}, \hat{M} = M, \hat{S} = s\}$. For $Lee_{M,s}$, we identified 1% infinite lengths in the large effect case ($\beta_i \in \{1, 0.8\}$), and increased to 2.7% for small effect case ($\beta_i = 0.2$), out of the variables that are selected to represent our submodel at each replication. This is attributed by the fact that the observed statistics for $\hat{\beta}_{j,\hat{M}_{obs}}$ are near at the endpoints of the truncation intervals in (2.8) causing numerical instability issues for $Lee_{M,s}$ (see left panel of Figure 1 Kivaranovic and Leeb [2020]). In other words, the polyhedron constraint $\{Ay \leq b\}$ is not satisfied at certain replications, $r \subset \{1, \ldots 200\}$.

2. *Randomization appears to be effective at providing shorter confidence intervals*

   Even with small B and R, we can see that randomization demonstrates smaller variability and shorter confidence intervals compared to the non-randomized version of the MC method. Analogously, we can see that (U, V) depicts similar effects when compared against DS as it also incorporates randomizing responses. Remarkably, $MC_{M,r+}$ provides substantially shorter lengths than all other methods especially when considering $\beta_i = 0.2$.

It seems that our MC method outperforms other methods provided that we consider non-zero true regression coefficients. However, the key question is whether under infinite number of replications R, and number of simulated conditional on selection distribution samples, B, our MC method will display asymptotic properties. Certainly, this can be further complicated by consideration of manipulating other parameters involved in the study.

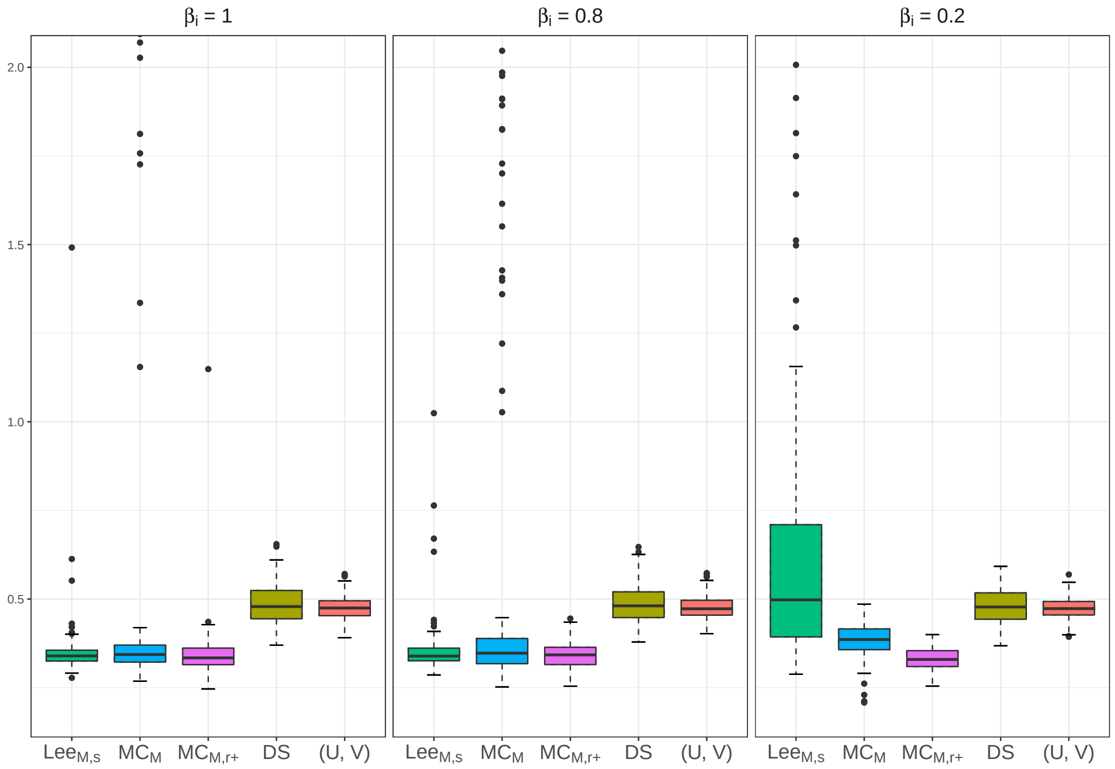

Figure 4.1.: Boxplots of interval lengths of $MC_M, MC_{M,r+}, DS$, and (U, V) respectively under $R = 200$ replications, evaluated for $\beta_i \in \{1, 0.8, 0.2\}$ under $B = 100$. Lengths beyond 2 are omitted from display.

	$\beta_i = 1$			$\beta_i = 0.8$			$\beta_i = 0.2$		
Method	Length	Coverage	Inf. length	Length	Coverage	Inf. length	Length	Coverage	Inf. length
$Lee_{M,s}$	0.340	87.50	1.00	0.339	87.00	1.00	0.499	91.89	2.70
MC_M	0.343	86.50	0.00	0.347	87.00	0.00	0.386	93.92	0.00
$MC_{M,r+}$	0.334	87.00	0.00	0.342	89.00	0.00	0.330	95.95	0.00
DS	0.478	98.20	0.00	0.481	87.00	0.00	0.479	88.50	0.00
(U, V)	0.473	88.19	0.00	0.472	87.50	0.00	0.475	86.00	0.00

Table 4.1.: Reported empirical median confidence interval lengths, mean coverage probabilities, and percentage of infinite lengths for $Lee_{M,s}$, MC_M, $MC_{M,r+}$, DS, and (U, V). Out of $R = 200$, the selection probability for variable 1 ($\beta_i = 1$), variable 2 ($\beta_i = 0.8$), and variable 3 ($\beta_i = 0.2$) are 1, 1, and 0.74 respectively.

4.3. Further performance evaluation under low-dimensional setting

The bisection search for appropriate confidence intervals demands high computational power for implementing the MC method. Consequently, computing large quantities of lengths can be critically limited at this stage.

We consider broadly evaluating the coverage performance of our MC methods under a few Monte-Carlo experiments on different regression coefficient sizes. To avoid confusion on terminology, we state "MC" to refer our proposed method in section 4.1, while "Monte-Carlo" to refer to the Monte-Carlo run for R replications. We also briefly examine whether randomization provides further improvements in interval lengths for larger R and B.

Coverage study

To evaluate the coverages for our MC method, we consider varying the parameters B and R. We consider $B \in \{200, 1000\}$ and run five Monte-Carlo simulations, each with $R = 1000$ replications. To avoid overly large computations, we provide a simple setup where: $n = 100, p = 3, \beta = \{1, 0.5, 0.2\}, \rho = 0.1, \sigma^2 = 1$, and 90% nominal confidence level. We assume σ^2 is known. In addition to the preliminary study in section 4.2.2, we consider extending comparison on low and large randomized noise for the MC method this time. Denote them by, $MC_{M,r}$ and $MC_{M,r+}$ with $W_{MC} \sim \mathcal{N}(0, \hat\sigma^2)$ and $W_{MC} \sim \mathcal{N}(0, 2.5^2\hat\sigma^2)$ respectively. σ^2 is estimated by (3.1). As a measure of uncertainty, we compute the Monte-Carlo error estimate of the coverage level Koehler et al. [2009].

Define the Monte-Carlo estimate of the coverage level of $\beta_{j,\hat M_{obs}}$ across all replications, $r = 1, \ldots, R$ in one simulation as:

$$\hat\psi^{MC}_{coverage} = \frac{1}{R}\sum_{r=1}^{R} I\left(\frac{\alpha}{2} \le F^{MC,\cup_s[\mathcal{V}_s^-(z),\mathcal{V}_s^+(z))]}_{\beta_{j,\hat M^r_{obs}},\sigma^2\|\eta_{j,\hat M^r_{obs}}\|_2^2}(\hat\beta_{j,\hat M^r_{obs}}) \le 1 - \frac{\alpha}{2}\right).$$

Notice that the superscript r on $\hat M^r_{obs}$ emphasizes that not all replications of the data (X, y) resulted in the submodel that includes the variable j selected with associated coefficient $\hat\beta_{j,\hat M^r_{obs}}$. Here, the significance level, $\alpha = 0.1$.

Then, the Monte-Carlo error (MCE) is the standard deviation of $\hat\psi^{MC}_{coverage}$:

$$MCE = \sqrt{Var(\hat\psi^{MC}_{coverage})}. \tag{4.6}$$

Our result using $B = 200$ is shown in Figure (4.2). We summarize our findings:

1. *Fast convergence to nominal levels for moderate to large effects*

   We see that the coverage levels of $MC_M, MC_{M,r}$ and $MC_{M,r+}$ quickly stabilizes close to the nominal levels after surpassing $R = 750$. $Lee_{M,s}$ reports slight under-coverages. As we have only conducted 5 repeated Monte-Carlo runs, the MCE is fairly large. Even so, we intuitively expect that the lengths produced by $Lee_{M,s}$ would be long enough to cover the true partial regression coefficients.

2. *Randomization provides close to exact and conservative coverage levels*

   When considering randomization ($MC_{M,r}$ and $MC_{M,r+}$), we also see consistent coverages at far shorter lengths than without. For $\beta_i = 0.2$, we observed obvious conservative coverages at around 0.95, above the specified nominal level. Based on our length results in Section 4.2.2, they are considerably shorter than without randomization. This provides indication that $MC_{M,r}$ and $MC_{M,r+}$ may enjoy higher degree of estimation accuracy than we control to be (90%).

Figure (4.3) shows the results when we increase B from 200 to 1000. Suprisingly, we do not see noticeable difference when applying $B = 1000$. Our observation reflects possibly faster convergence to the respective coverages by the MC methods, as detailed by lower MCE (improvement of 0.01 to 0.02). However, producing similar results with higher computational trade-off means that choosing higher B may be practically less favourable than choosing lower B.

Interval length study

To investigate whether randomization indeed substantially improves the interval lengths, we consider increasing R from 200 to 300, B from 100 to 200 and set $\beta = \{1, 0.5, 0.2\}$ to compared against the preliminary results in Section 4.2.2. We consider using λ_{univ} as before.

Our findings based on Figure (4.4) and Table (4.2) show consistency with our preliminary results in Section 4.2.2. We summarize our overall main takeaways:

1. *Randomization is effective at shortening of confidence intervals*

   As can be compared between the two pairs: (1) MC_M and $MC_{M,r+}$, (2) DS and (U, V), demonstrates the shortening of confidence intervals benefited from randomizing responses, confirming the conclusions by Tian and Taylor [2018].

   Although infinite lengths are not detected, it is certain that MC_M resulted in undesirably many lengthy intervals, and $MC_{M,r+}$ substantially eliminates this phenomenon.

2. *$MC_{M,r+}$ provides an appealing alternative over (U, V)*

   Most notably, when compared with (U, V), $MC_{M,r+}$ dominates (U, V). We recall that this can be explained by conditioning on "less" heuristic justified by Fithian et al. [2014]. (U, V) discards randomized response information upon selection, but $MC_{M,r+}$ incorporates the selection information through the conditional of
   $(Y_{sim} + W_{MC})| \{(Y_{sim} + W_{MC}) \in \hat{M}_{obs}\}$.

   However, we require to pay the price for implementing large Monte-Carlo replications, R to compete with (U, V). Nonetheless, we wish to emphasize that the MC method can be an appealing alternative over (U, V) if computational costs is not an issue.

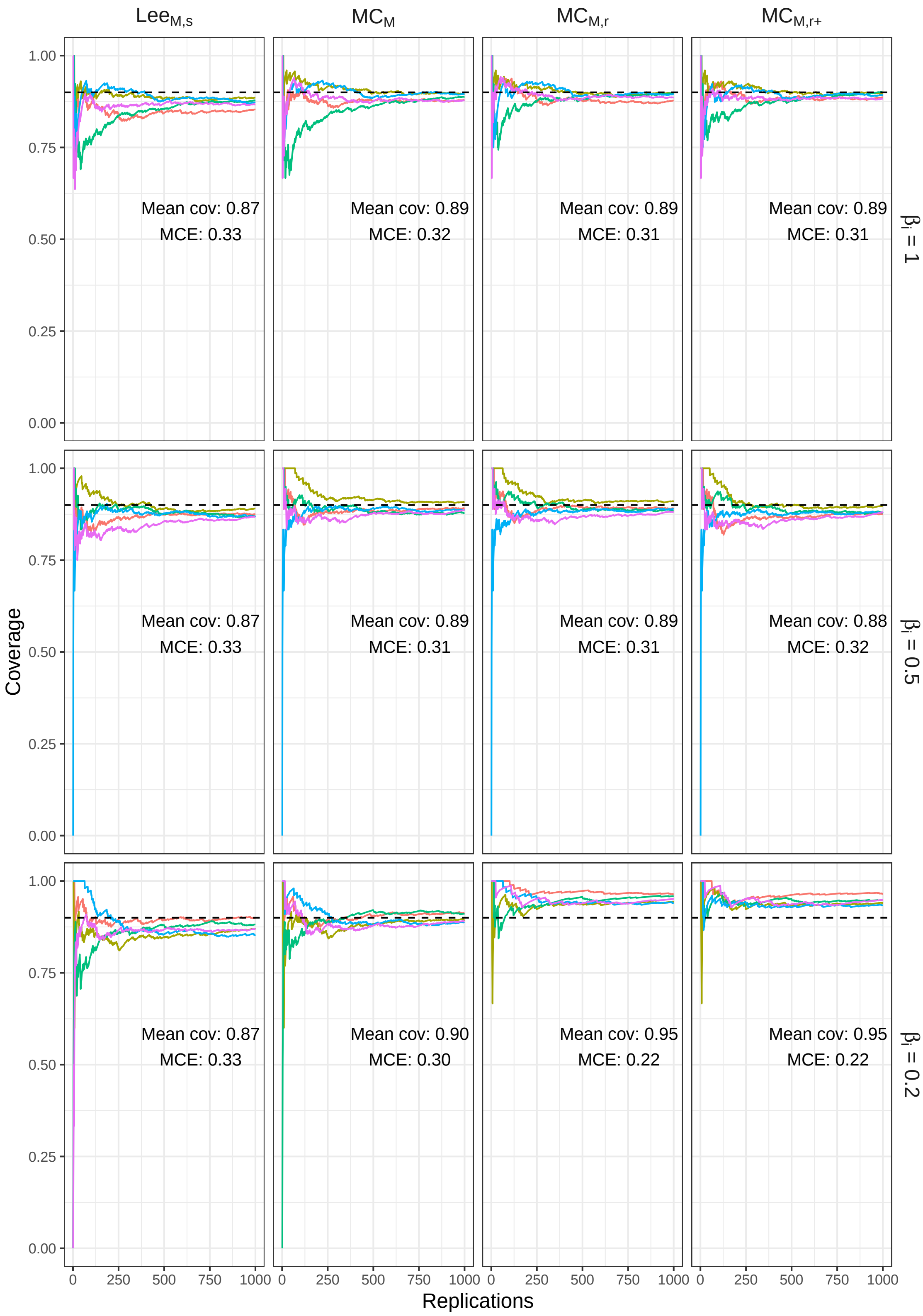

Figure 4.2.: Monte-Carlo estimates of the coverage levels of $Lee_{M,s}$, MC_M, $MC_{M,r}$, $MC_{M,r+}$ respectively under $R = 1000$ replications for 5 simulations runs, evaluated for $\beta \in \{1, 0.5, 0.2\}$ under $B = 200$. Horizontal dotted line indicate the nominal coverage level at 90%. Over 5 Monte-Carlo runs, out of $R = 1000$, the selection probability of variable 1, 2, and 3 are overall 1, 1, and 0.75 respectively.

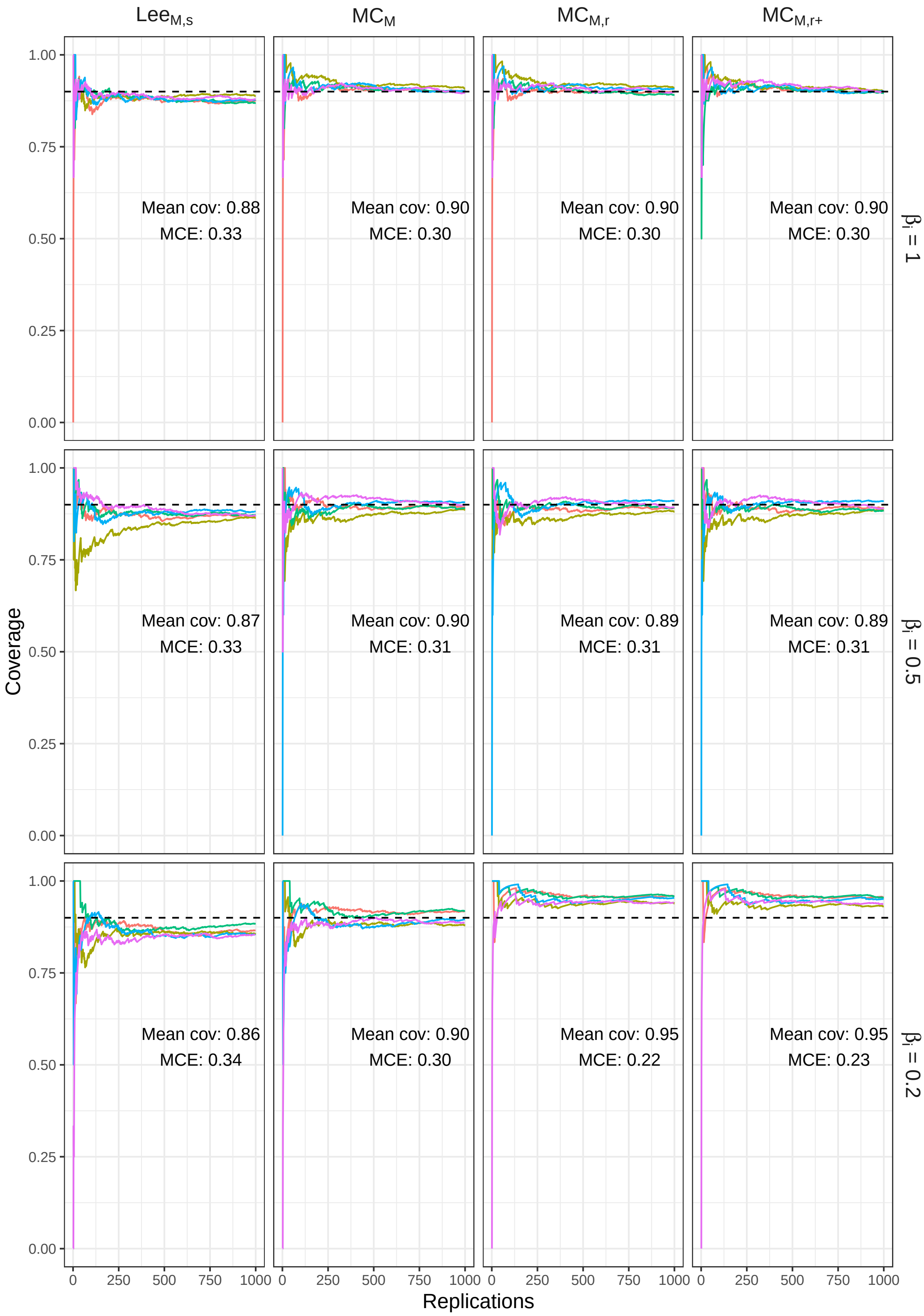

Figure 4.3.: Monte-Carlo estimates of the coverage levels of $Lee_{M,s}$, MC_M, $MC_{M,r}$, $MC_{M,r+}$ respectively under $R = 1000$ replications for 5 simulations runs, evaluated for $\beta \in \{1, 0.5, 0.2\}$ under $B = 1000$. Horizontal dotted line indicate the nominal coverage level at 90%. Over 5 Monte-Carlo runs, out of $R = 1000$, the selection probability of variable 1, 2, and 3 are overall 1, 1, and 0.75 respectively.

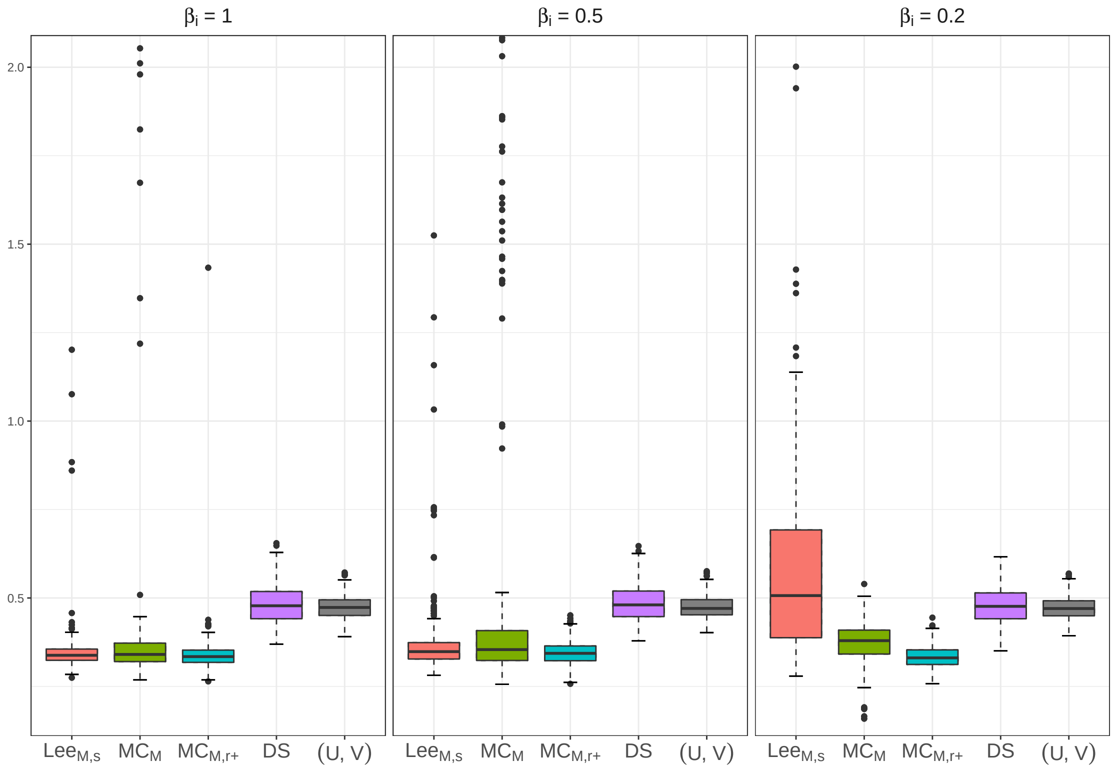

Figure 4.4.: Boxplots of interval lengths of $MC_M, MC_{M,r+}, DS$, and (U, V) respectively under $R = 300$ replications, evaluated for $\beta_i \in \{1, 0.5, 0.2\}$ under $B = 200$. Lengths beyond 2 are omitted from display.

	$\beta_i = 1$			$\beta_i = 0.5$			$\beta_i = 0.2$		
Method	Length	Coverage	Inf. length	Length	Coverage	Inf. length	Length	Coverage	Inf. length
$Lee_{M,s}$	0.339	87.67	3.33	0.350	87.67	3.67	0.513	86.90	5.24
MC_M	0.340	90.33	0.00	0.354	92.33	0.00	0.379	89.52	0.00
$MC_{M,r+}$	0.334	88.33	0.00	0.344	90.33	0.00	0.331	92.58	0.00
DS	0.476	96.82	0.00	0.481	86.55	0.00	0.478	89.00	0.00
(U, V)	0.470	88.83	0.00	0.471	85.86	0.00	0.473	85.33	0.00

Table 4.2.: Reported empirical median confidence interval lengths, mean coverage probabilities, and percentage of infinite lengths for $Lee_{M,s}$, MC_M, $MC_{M,r+}$, DS, and (U, V). Out of $R = 300$, the selection probability for variable 1 ($\beta_i = 1$), variable 2 ($\beta_i = 0.5$), and variable 3 ($\beta_i = 0.2$) are 1, 1, and 0.76 respectively.

5. Conclusions

In Section 3 of our work, we conduct sensitivity analyses on (U, V) decomposition, a randomized splitting strategy proposed by Rasines and Young [2021] as a selective and inferential power enhancement device against data splitting (DS). Using DS as our baseline for comparison, we evaluate its inference stage performance under heavy-tailed and skewed distributional simulated data and for the case where the normal variance, σ^2 is known. Our findings are two-fold:

1. (U, V) decomposition (and DS) are practically immune to heavy-tailed and skewed data.
2. (U, V) decomposition (and DS) possess close to oracle properties when they rely on variance estimation especially in low-dimensional settings.

Overall, (U, V) decomposition still effectively generates shorter interval lengths than DS while maintaining close to nominal coverage levels. When σ^2 is unknown, we see that provided we are considering a low-dimensional $(n > p)$ setup and σ^2 is small, Theorem 1 of Rasines and Young [2021] can be relaxed without asymptotic condition, i.e. $n \to \infty$.

In Section 4, we introduce the condition on model only Monte-Carlo (MC) rejection sampling based approach, MC_M that follows upon the framework by Lee et al. [2016]. We empirically reveal that our method can potentially provide significant improvement in shortening of confidence interval lengths compared to $Lee_{M,s}$, DS and (U, V) while maintaining optimal coverage levels when making inference for partial regression coefficients. Relying on our toy preliminary findings, we certainly do not find signs of infinite lengths for our method especially when we incorporate randomization. Of course, there is a possibility for generating infinite lengths if nonsensical parameters are provided for the bisection algorithm, specifically through initializing "bad" initial estimates and choosing large error tolerance.

Undeniably, our MC method has its underlying complications. As demonstrated by our preliminary Monte-Carlo simulations, we can see that our MC method could only provide best results when the selected variables for the submodel have coefficients that are not close to zero. We can expect that the lengths are substantially shorter than (U, V) under "ideal" scenarios. By "ideal", we mean that our limited preliminary results are based on $n > p$ setup, and σ^2 is assumed known. We did not examine the effectiveness of MC_M for correlated X and post-selection inference when λ is not fixed.

Furthermore, the main critique is the computational infeasibility of the MC method to provide appealing arguments against other methods. Two computational issues are found when implementing the algorithm: (1) Low or no simulated condition on selection samples especially for relative small coefficients, prompting early sampling termination. (2) Slowness in bisection search for confidence limits. To meet the computational demands of the MC method, future extensions could include parallel computation of replications, or using a high performance computing technology, to allow more detailed analysis. Along with these tools, it is exciting to also improve the efficiency over the bisection search for approximating the limits of confidence intervals. Alternatively, one could apply a stochastic-based approximation algorithm such as the Robbins-Monro (RM) Search Algorithm Buckland [1985] which conducts independent searches on the confidence interval. Another proposed interpolation approach introduced by O'Gorman [2018] circumvents the lengthy computational steps of RM, while maintaining the coverage probabilities and could be extended to our setting.

Considering the flexibility of the MC method, smooth integration to non-normal distributional settings and non-convex model selection representations (beyond LASSO) will be of great interest for exploration in the future. Given that various selective inference methods beyond our scope of consideration exist (see Liu et al. [2018], Liu et al. [2022]), extensive comparison study is required to further assess the relative performance of our MC method. While substantial empirical evaluations provides extensive insights, balance on establishing definitive asymptotic theory should also be taken care for rigorous assessment of the method.

At this stage, we certainly still recommend (U, V) decomposition as a theoretically and empirically validated, flexible, and computationally viable approach compared to other methods. If many significant variables are expected and the observational data is normal, the MC method may be very useful for providing better inferential power than (U, V) at the expense of computational demands. $Lee_{M,s}$ should certainly be avoided in most cases due to restrictive assumptions coupled with unacceptably long intervals.

A. Appendix

A.1. Bisection algorithm

The following bisection algorithm searches the lower and upper confidence limits, (L_{est}, U_{est}) to approximate the pivotal quantities in (4.2) and (4.3).

Algorithm 1.1 Bisection algorithm for the lower bound

```
Input: Pivot, $F(\cdot) \to \mathbb{R}$; Left estimate, $a^{(0)} \in \mathbb{R}^{(-\infty,0)}$, Right estimate, $b^{(0)} \in \mathbb{R}^{(0,\infty)}$;
    Error tolerance, $\epsilon$; Significance level, $\alpha$.
Output: Lower limit estimate, $L_{est}$.
Goal: Solve $g(\cdot) = F(\cdot) - (1 - \frac{\alpha}{2}) = 0$.

Evaluate $g(c)$ with $c = \frac{a^{(0)}+b^{(0)}}{2}$.
if $g(c) = 0$ then return $c$.
if $g(c) < 0$ then
    $a^* = a^{(0)}$.
    while $g(a^*) < 0$ do
        Evaluate $g(a^*)$ with $a^* := 2a^*$.
        if $a^* = -\infty$ then return $L_{est} = a^* = -\infty$.
    Evaluate $g(a^{(1)})$ with $a^{(1)} = a^*$.
    Evaluate $g(b^{(1)})$ with $b^{(1)} = c$.
else
    $b^* = b^{(0)}$
    while $g(b^*) > 0$ do
        Evaluate $g(b^*)$ with $b^* := 2b^*$.
        if $b^* = \infty$ then return $L_{est} = b^* = \infty$.
    Evaluate $g(a^{(1)})$ with $a^{(1)} = c$.
    Evaluate $g(b^{(1)})$ with $b^{(1)} = b^*$.
for $t = 1, 2, \ldots$ do
    Evaluate $g(c')$ with $c' = \frac{a^{(t)}+b^{(t)}}{2}$.
    if $g(c') = 0$ then return $L_{est} = c'$.
    if $(g(c') \cdot g(a^{(t)}) < 0)$ then
        $(a^{(t+1)}, b^{(t+1)}) = (a^{(t)}, c')$..
    else
        $(a^{(t+1)}, b^{(t+1)}) = (c', b^{(t+1)})$
    Evaluate $g(a^{(t+1)})$ with $a^{(t+1)} = c'$.
    if $|b^{(t+1)} - a^{(t+1)}| < \epsilon$ then return
        $L_{est} = \frac{a^{(t+1)}+b^{(t+1)}}{2}$.
```

To find U_{est}, we simply obtain the survival function form of the pivot, i.e., we solve for $F^* := 1 - F^{\cup_s[\mathcal{V}_s^-(z),\mathcal{V}_s^+(z))]}_{U_{est},\sigma^2\|\eta_{j,\hat{M}_{obs}}\|_2^2}(\eta^T_{j,\hat{M}_{obs}} y) \approx 1 - \frac{\alpha}{2}$.

Algorithm 1.2 Bisection algorithm for the upper bound

```
Input: Pivot, $F^*(\cdot) \to \mathbb{R}$; Left estimate, $a^{(0)} \in \mathbb{R}^{(-\infty,0)}$, Right estimate, $b^{(0)} \in \mathbb{R}^{(0,\infty)}$;
  Error tolerance, $\epsilon$; Significance level, $\alpha$.
Output: Upper limit estimate, $U_{est}$.
Goal: Solve $g(\cdot) = F'(\cdot) - (1 - \frac{\alpha}{2}) = 0$.

Evaluate $g(c)$ with $c = \frac{a^{(0)}+b^{(0)}}{2}$.
if $g(c) = 0$ then return $c$.
if $g(c) > 0$ then
    $a^* = a^{(0)}$.
    while $g(a^*) > 0$ do
        Evaluate $g(a^*)$ with $a^* := 2a^*$.
        if $a^* = -\infty$ then return $U_{est} = a^* = -\infty$.
    Evaluate $g(a^{(1)})$ with $a^{(1)} = a^*$.
    Evaluate $g(b^{(1)})$ with $b^{(1)} = c$.
else
    $b^* = b^{(0)}$.
    while $g(b^*) < 0$ do
        Evaluate $g(b^*)$ with $b^* := 2b^*$.
        if $b^* = \infty$ then return $U_{est} = b^* = \infty$.
    Evaluate $g(a^{(1)})$ with $a^{(1)} = c$.
    Evaluate $g(b^{(1)})$ with $b^{(1)} = b^*$.
for $t = 1, 2, \ldots$ do
    Evaluate $g(c')$ with $c' = \frac{a^{(t)}+b^{(t)}}{2}$.
    if $g(c') = 0$ then return $U_{est} = c'$.
    if $(g(c') \cdot g(a^{(t)}) < 0)$ then
        $a^{(t+1)}, b^{(t+1)} = (a^{(t)}, c')$.
    else
        $a^{(t+1)}, b^{(t+1)} = (c', b^{(t+1)})$.
    Evaluate $g(a^{(t+1)})$ with $a^{(t+1)} = c'$.
    if $|b^{(t+1)} - a^{(t+1)}| < \epsilon$ then return
        $U_{est} = \frac{a^{(t+1)}+b^{(t+1)}}{2}$.
```

Remark

We apply the estimates, $(a^{(0)}, b^{(0)}) = (-sign(x_1)x_1, sign(x_2)x_2)$, where:

$$x_1 = \hat{\beta}_{j,\hat{M}_{obs}} - 10\sigma\sqrt{\left\|\eta_{j,\hat{M}_{obs}}\right\|_2^2}, \tag{A.1}$$

$$x_2 = \hat{\beta}_{j,\hat{M}_{obs}} + 10\sigma\sqrt{\left\|\eta_{j,\hat{M}_{obs}}\right\|_2^2}, \tag{A.2}$$

for our simulation studies as a reasonable starting point for the bisection algorithm. The error tolerance, ϵ is set to 0.01. Intuitively, setting farther estimates from $\hat{\beta}_{j,\hat{M}_{obs}}$ and larger ϵ incur higher computational costs.

A.2. Independence of U and V

Proof

We first show that U and V are jointly multivariate normal and then deduce that U and V are independent if and only if they are uncorrelated.

Assuming $Y = (Y_1, \dots Y_n)^T$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$, then, $U = (U_1, \dots, U_n)^T$ are i.i.d normal distributed as in (2.13). Similarly, $V = (V_1, \dots, V_n)^T$ are i.i.d normal distributed as in (2.14). Alternatively, we can concretely re-express (2.13) and (2.14) as the n-variate forms:

$$U \sim \mathcal{N}_n(\mu_n, \Sigma_U),$$
$$V \sim \mathcal{N}_n(\mu_n, \Sigma_V),$$

where:

- $\mu_n \in \mathbb{R}^n$,
- $\Sigma_U \in \mathbb{R}^{n \times n}$ is the covariance matrix of U and has $(1+\gamma^2)\sigma^2$ in the diagonal and 0's in the off-diagonals,
- $\Sigma_V \in \mathbb{R}^{n \times n}$ is the covariance matrix of V and has $(1+\gamma^{-2})\sigma^2$ in the diagonal and 0's in the off-diagonals.

Using the extension to the definition of bivariate normal distribution, we have that U and V are jointly $2n$-variate normally distributed:

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}_{2n}\left(\mu_{2n}, \begin{pmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{pmatrix}\right),$$

where $\Sigma_{UV} = \Sigma_{VU}^T \in \mathbb{R}^{n \times n}$.

To prove that U and V are independent, it suffices to show that the off-diagonals are zero, i.e., $\Sigma_{UV} = 0_{n \times n}$. The result follows:

$$\begin{aligned}
\Sigma_{UV} &\equiv Cov(U, V) \\
&= \mathbb{E}[(U - \mathbb{E}(U))(V - \mathbb{E}(V))] && \text{(definition of covariance)} \\
&= \mathbb{E}[(U - \mu_n)(V - \mu_n)] \\
&= \mathbb{E}[UV^T - U\mu_n^T - V\mu_n^T + \mu_n\mu_n^T] \\
&= \mathbb{E}[UV^T] - \mu_n(\mathbb{E}[U] + \mathbb{E}[V]) + \mu_n^T\mu_n \\
&= \mathbb{E}[UV^T] - 2\mu_n\mu_n^T + \mu_n\mu_n^T \\
&= \mathbb{E}[(Y + \gamma W)(Y - \gamma^{-1}W)] - \mu_n\mu_n^T && \text{(definition of } U \text{ and } V\text{)} \\
&= \mathbb{E}[(YY^T - \gamma^{-1}YW^T + \gamma WY^T - WW^T] - \mu_n\mu_n^T \\
&= \mathbb{E}[Y]\mathbb{E}[Y]^T - \gamma^{-1}\mathbb{E}[Y]\mathbb{E}[W]^T + \gamma\mathbb{E}[W]\mathbb{E}[Y]^T - \mathbb{E}[W]\mathbb{E}[W]^T - \mu_n\mu_n^T \\
& && \text{(i.i.d property of } Y \text{ and } W\text{)} \\
&= \mu_n\mu_n^T - 0 + 0 - 0 - \mu_n\mu_n^T && \text{(E[W] = 0)} \\
&= 0_{n \times n}.
\end{aligned}$$

Now, suppose we assume that U and V are independent. Since Σ_{UV} is the covariance matrix between the components of U and V, then, $\Sigma_{UV} = 0_{n\times n}$.

A.3. Reproducibility

R code for reproducing figures and numerical results are available at:
https://github.com/tsc21-ic/PSI_evaluation_thesis

Remark Results may vary slightly depending on the user's specified seed number.

Bibliography

Barber, R. F. and Candès, E. J. [2015], 'Controlling the false discovery rate via knockoffs', *The Annals of Statistics* **43**(5), 2055–2085.

Begley, C. G. and Ioannidis, J. P. [2015], 'Reproducibility in science: improving the standard for basic and preclinical research', *Circulation Research* **116**(1), 116–126.

Bélisle, C. J., Romeijn, H. E. and Smith, R. L. [1993], 'Hit-and-run algorithms for generating multivariate distributions', *Mathematics of Operations Research* **18**(2), 255–266.

Benjamini, Y. and Yekutieli, D. [2005], 'False discovery rate–adjusted multiple confidence intervals for selected parameters', *Journal of the American Statistical Association* **100**(469), 71–81.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. [2013], 'Valid post-selection inference', *The Annals of Statistics* pp. 802–837.

Buckland, S. T. [1985], 'Algorithm as 214: calculation of monte carlo confidence intervals', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **34**(3), 296–301.

Cox, D. R. [1975], 'A note on data-splitting for the evaluation of significance levels', *Biometrika* **62**(2), 441–444.

Fernández, C. and Steel, M. F. [1998], 'On Bayesian modeling of fat tails and skewness', *Journal of the American Statistical Association* **93**(441), 359–371.

Fithian, W., Sun, D. and Taylor, J. [2014], 'Optimal inference after model selection', *arXiv preprint arXiv:1410.2597* .

Garcia-Angulo, A. C. and Claeskens, G. [2022], 'Exact uniformly most powerful postselection confidence distributions', *Scandinavian Journal of Statistics* .

Kivaranovic, D. and Leeb, H. [2020], 'A (tight) upper bound for the length of confidence intervals with conditional coverage', *arXiv preprint arXiv:2007.12448* .

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**(534), 845–857.

Koehler, E., Brown, E. and Haneuse, S. J.-P. [2009], 'On the assessment of monte carlo error in simulation-based statistical analyses', *The American Statistician* **63**(2), 155–162.

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**(3), 907–927.

Lindqvist, B. H. and Taraldsen, G. [2007], Conditional monte carlo based on sufficient statistics with applications, *in* 'Advances in statistical modeling and inference: Essays in honor of Kjell A Doksum', World Scientific, pp. 545–561.

Liu, K., Markovic, J. and Tibshirani, R. [2018], ‘More powerful post-selection inference, with application to the lasso’, *arXiv preprint arXiv:1801.09037* .

Liu, S., Markovic, J. and Taylor, J. [2022], ‘Black-box selective inference via bootstrapping’, *arXiv preprint arXiv:2203.14504* .

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. [2014], ‘A significance test for the lasso’, *The Annals of Statistics* **42**(2), 413.

Markovic, J., Xia, L. and Taylor, J. [2017], ‘Unifying approach to selective inference with applications to cross-validation’, *arXiv preprint arXiv:1703.06559* .

Meinshausen, N. and Bühlmann, P. [2010], ‘Stability selection’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4), 417–473.

O’Gorman, T. W. [2018], ‘Reducing the width of confidence intervals for the difference between two population means by inverting adaptive tests’, *Statistical Methods in Medical Research* **27**(5), 1422–1436.

Pakman, A. and Paninski, L. [2014], ‘Exact hamiltonian monte carlo for truncated multivariate gaussians’, *Journal of Computational and Graphical Statistics* **23**(2), 518–542.

Rasines, D. G. and Young, G. A. [2021], ‘Splitting strategies for post-selection inference’, *arXiv preprint arXiv:2102.02159* .

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

Taylor, J. and Tibshirani, R. [2018], ‘Post-selection inference for-penalized likelihood models’, *Canadian Journal of Statistics* **46**(1), 41–61.

Tian, X. and Taylor, J. [2018], ‘Selective inference with a randomized response’, *The Annals of Statistics* **46**(2), 679–710.

Tibshirani, R. J. and Taylor, J. [2011], ‘The solution path of the generalized lasso’, *The Annals of Statistics* **39**(3), 1335–1371.

Wasserman, L. and Roeder, K. [2009], ‘High dimensional variable selection’, *The Annals of Statistics* **37**(5A), 2178.

White, H. [2000], ‘A reality check for data snooping’, *Econometrica* **68**(5), 1097–1126.

Yong, E. [2012], ‘Bad copy’, *Nature* **485**(7398), 298.