Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Approximate Selective Inference via Maximum Likelihood

Snigdha Panigrahi & Jonathan Taylor

To cite this article: Snigdha Panigrahi & Jonathan Taylor (2022): Approximate Selective Inference via Maximum Likelihood, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2081575

To link to this article: https://doi.org/10.1080/01621459.2022.2081575

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–11: Theory and Methods
https://doi.org/10.1080/01621459.2022.2081575

Taylor & Francis
Taylor & Francis Group

Approximate Selective Inference via Maximum Likelihood

Snigdha Panigrahi[a] and Jonathan Taylor[b]

[a]Department of Statistics, University of Michigan, Ann Arbor, MI; [b]Department of Statistics, Stanford University, Stanford, CA

ABSTRACT

Several strategies have been developed recently to ensure valid inference after model selection; some of these are easy to compute, while others fare better in terms of inferential power. In this article, we consider a selective inference framework for Gaussian data. We propose a new method for inference through approximate maximum likelihood estimation. Our goal is to: (a) achieve better inferential power with the aid of randomization, (b) bypass expensive MCMC sampling from exact conditional distributions that are hard to evaluate in closed forms. We construct approximate inference, for example, *p*-values, confidence intervals etc., by solving a fairly simple, convex optimization problem. We illustrate the potential of our method across wide-ranging values of signal-to-noise ratio in simulations. On a cancer gene expression dataset we find that our method improves upon the inferential power of some commonly used strategies for selective inference. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2019
Accepted May 2022

KEYWORDS
Conditional inference; Data adaptivity; Maximum likelihood; Multiple queries; Post-selection inference; Randomization; Selective MLE

1. Introduction

Querying the data has become a fairly common practice for anyone who wishes to learn a model from a range of different candidates. Naively using the same data twice, first to learn a model and then infer for the selected parameters, tends to inflate their estimated effects. As an example of a query, consider a variable selection algorithm with a shrinkage penalty (Tibshirani 1996; Fan and Li 2001; Yuan and Lin 2006); the algorithm learns a set of variables (or features) into a model. Ignoring the dependence of the model (and its parameters) on the outcome of the query while calculating p values, confidence intervals, credible intervals etc. undermines inference after selection; see Benjamini and Yekutieli (2005), Leeb and Pötscher (2005), Leeb and Pötscher (2006), and Berk et al. (2013) for a demonstration of the concerns here. The result is usually an increased chance of finding a statistically significant result when the selected variable in fact has no effect.

Various strategies for selective inference offer different solutions by characterizing the dependence between the learned models and data. Some of the strategies are easier to implement, while others fare better in inferential power. In this article, we introduce a new method for selective inference through approximate maximum likelihood estimation. Building on the recent work by Tian and Taylor (2018), our method allows us to: (a) harness Gaussian randomization variables toward better inferential power after selection, and simultaneously (b) bypass expensive MCMC sampling from intractable conditional distributions. Below, we provide a brief, informal overview of our method in a standard setup of linear regression.

An informal overview of our method. Consider a regression problem in which we observe a response vector $y \in \mathbb{R}^n$ and a matrix of p predictors $X \in \mathbb{R}^{n\times p}$. Let $\omega \in \mathbb{R}^p$ be drawn from a centered Gaussian distribution with known covariance $\Sigma_{\mathbb{W}}$. For fixed values $\lambda \in \mathbb{R}^+$, $\epsilon \in \mathbb{R}^+$, we solve the following query:

$$\underset{o}{\text{minimize}}\ \frac{1}{2}\|y - Xo\|_2^2 + \lambda\|o\|_1 + \frac{\epsilon}{2}\|o\|_2^2 - \omega^{\intercal} o. \qquad (1)$$

We call the optimization in (1) a "randomized LASSO" query; so named because of the randomization variable ω added to the objective of the canonical LASSO. Suppose the query selects a nonempty set of variables $\mathrm{E} \subseteq \{1, 2, \ldots, p\}$. Following selection, we describe our response variable through the model: $y = X_{\mathrm{E}}\beta_{\mathrm{E}} + \mathrm{e}$, $\mathrm{e} \sim N(0, \sigma^2 I_n)$, where I_n is the identity matrix with n rows and columns.

A natural ask in the learned model is inference for the partial regression coefficients after adjusting for their dependence on data through E. One concrete course of action is to condition on selection, specifically, base inference on the likelihood of the observed data when conditioned on the event

$$\left\{(y, \omega) : \widehat{\mathrm{E}}(y, \omega) = \mathrm{E}\right\},$$

where $\widehat{\mathrm{E}}$ represents the (data-dependent) selected set of variables. Maximizing the conditional likelihood function gives us $\widehat{\beta}_{\mathrm{E}}^{\mathrm{mle}}$, the maximum likelihood estimate (MLE) for β_{E}. Taking the Hessian of the negative log-likelihood at the MLE yields us $I(\widehat{\beta}_{\mathrm{E}}^{\mathrm{mle}})$, the observed Fisher information matrix. The (approximate) confidence intervals resulting from our method take

CONTACT Snigdha Panigrahi psnigdha@umich.edu Department of Statistics, University of Michigan, Ann Arbor, MI.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

the form

$$\widehat{\beta}^{\text{mle}}_{j\cdot \text{E}} \pm z_{1-q/2} \cdot \sqrt{I^{-1}_{j,j}(\widehat{\beta}^{\text{mle}}_{\text{E}})}, \tag{2}$$

where $\widehat{\beta}^{\text{mle}}_{j\cdot \text{E}}$ is the jth component of $\widehat{\beta}^{\text{mle}}_{\text{E}}$, $I^{-1}_{j,j}(\widehat{\beta}^{\text{mle}}_{\text{E}})$ is the (j,j)th entry of $I^{-1}(\widehat{\beta}^{\text{mle}}_{\text{E}})$ and z_{1-q} is the $(1-q)$th quantile of a standard normal distribution for $q \in (0,1)$.

The intervals proposed in (2) are seemingly straightforward if only we could directly calculate the two estimates in the expression, the MLE and the observed Fisher information matrix. But as it turns out, the conditional likelihood and subsequently the two estimates based on it do not admit expressions in closed forms. In the remaining development, we solve this challenge head-on in two steps. First, we construct an approximate, statistically consistent proxy for the likelihood function after conditioning on the selection event. Then, we provide a tractable system of estimating equations to obtain the MLE and the observed Fisher information matrix from our proxy likelihood. At the core of the proposed estimating equations is a fairly simple, convex optimization problem in relatively few dimensions.

Comparison with common baselines. Continuing with the regression setup, we compare our proposal with some common baselines in a simulated experiment and relate our method with existing work. In Table 1, four methods for selective inference including our method are evaluated on three criteria after selecting variables using a fixed value of tuning parameter: average coverage of interval estimates with nominal false coverage rate (FCR) level of 0.10; average length of the interval estimates; the power of detecting true associations after applying the selective inference strategy. The data in this experiment obeys a linear model with a 300-by-100 design matrix X and Gaussian errors, such that the rows of X are iid copies of a correlated multivariate normal vector; the model coefficient vector β has 6 nonzero components that are linearly varying in magnitude, and the setting corresponds to a relatively weak signal-to-noise ratio value. The simulation setting is described more precisely later in the article and relative comparisons between all the methods are made for a wide range of values for signal-to-noise ratio.

The first baseline, "Lee et al.," proposed by Lee et al. (2016) reduces inference to a truncated normal variable through the Polyhedral Lemma. The second baseline, "Split" uses a randomly chosen one-third of the data samples for inference after applying the LASSO to the remaining two-thirds of the data, (see, e.g. Cox 1975; Hurvich and Tsai 1990). Based on Liu, Markovic, and Tibshirani (2018), the third baseline "Liu et al." conditions on (strictly) less information than "Lee et al." by choosing to infer for the parameters associated with the selected variables in the full model: $y \sim N(X\beta, \sigma^2 I_n)$. All the other strategies in this example use the learned model: $y \sim N(X_{\text{E}}\beta_{\text{E}}, \sigma^2 I_n)$ based on the selected set of variables.

Table 1. Comparison with baselines.

Method	Coverage $100 \cdot (1 - \text{FCR})\%$	Lengths	Power	% of infinitely long intervals
MLE (Our method)	90.92%	8.31	85%	0
Lee et al.	85.60%	∞	77%	3.7
Split	88.40%	14.83	56%	0
Liu et al.	82.68%	9.67	64%	0

We note that FCR is (roughly) attained at the nominal level by all the strategies, except "Liu et al." falls slightly short of the mark in this setting. Our method, namely "MLE", delivers the shortest intervals with the highest power. In the last column of the table, we also indicate the percentage of intervals that resulted with infinite length. Among all the methods, "Lee et al." produces some infinitely long intervals; this observation is consistent with the established fact in Kivaranovic and Leeb (2018) that the intervals based on the Polyhedral Lemma do not have a finite expected value in the Gaussian regression setting. Both "Split" and "Liu et al." overcome the drawbacks of "Lee et al." by setting aside more information for inference. The former strategy does so by reserving a randomly chosen subsample for inference, while the latter achieves an increase in power through a larger truncation set.

By analogy with Tian and Taylor (2018), our method uses added randomization to remedy the excessively long intervals produced by "Lee et al."; because we do not condition on the randomization variable itself, inference does not trivially reduce to the Polyhedral Lemma. Our choice of adding a Gaussian randomization variable to the query draws motivation from data carving, a two stage situation where parameters learned on an initial dataset are estimated using new samples augmented with the initial ones (Fithian, Sun, and Taylor 2014; Panigrahi, Zhu, and Sabatti 2019; Panigrahi 2019). In the analysis here, the variance of ω is chosen so that the randomized LASSO (roughly) resembles "Split" in the amount of information used toward learning the model. Our method improves upon "Split" by conditioning upon an event that implies the selection of the variables in the set E; by doing so, our method reuses data used in selection. A direct relation between the power attained with our randomized method and "Liu et al.," however, is lacking. Some gain in power reported for the above setting might be attributed to the use of the learned model by our method as opposed to the full model under which "Liu et al." offers inference.

Other related work. Our maximum likelihood method differs from previous proposals in the tools used and the scope of inference. Existing strategies for selective inference usually require sampling from conditional distributions, either due to the intractability of their exact counterparts or due to the lack of easily available truncation regions. For example, the pivot described in Tian and Taylor (2018) lacks exact expressions and is not readily amenable for computational analyses. The truncation region in Liu, Markovic, and Tibshirani (2018) takes a tractable form for the LASSO; however, this form does not directly generalize to other queries. Other inferential approaches that account for the effects of selection include sampling from a selection-adjusted posterior in Panigrahi, Taylor, and Weinstein (2021) and resampling-based approaches such as bootstrap in McKeague and Qian (2015) and Guo and He (2020). The computing costs of these approaches are especially acute if two or more queries are applied for learning models, and the construct of valid inference must appropriately take into consideration the effect of each such query. Bypassing the requirement to sample from intractable

conditional distributions after selection, our maximum likelihood method relies on the solution to a simple, convex optimization problem. Furthermore, this convex problem assumes a separable form under multiple queries which is amenable to parallel computing. Much of prior work in the area of selective inference has relied on a testing-based approach for real-valued projections of parameters; see for example Yang et al. (2016), Suzumura et al. (2017), and Rügamer and Greven (2018). In contrast, the scope of the present likelihood-based approach allows joint inference for parameter vectors in learned models.

The rest of the article is organized as follows. In Section 2, we introduce our approximate proposal in a univariate file drawer problem. We describe in Section 3 our method of selective inference by deriving a system of estimating equations for the MLE and the observed Fisher information matrix after we solve a convex query. We conduct simulations in Section 4 to study the gains with our method over existing baselines. We apply our method to gene expression data from The Cancer Genome Atlas in Section 5, corroborating some of the numerical findings in the simulated experiments. We conclude with a discussion in Section 6. In the supplementary material, we include proofs for our main results and generalize our method of selective inference for interval estimation after solving multiple convex queries.

2. MLE Inference: A First Example

Before proceeding further, note, we use (a) $\bar{\Phi}(x)$ for the upper tail probability of the standard Gaussian law at $x \in \mathbb{R}$, and (b) $\phi(u; \nu, \Theta)$ for the (multivariate) Gaussian density function with mean vector μ and covariance Θ at the value u; the special symbol $\phi(x; 0, 1)$ denotes the standard Gaussian density for $x \in \mathbb{R}$.

2.1. Univariate Soft-Truncated Likelihood

We consider two independent random variables:

$$Y \sim N(\beta, 1),\ W \sim N(0, \eta^2),$$

where W denotes a Gaussian randomization variable. We pursue inference for β only if:

$$Y + W > \tau, \quad \text{where } \tau = \sqrt{1+\eta^2} \cdot z_{1-q}.$$

We begin by describing a conditional likelihood by conditioning the Gaussian law of Y upon the selection event:

$$\left\{(y, \omega) \in \mathbb{R}^2 : y + \omega > \tau\right\}. \tag{3}$$

Define $O = Y + W - \tau$, which we call an optimization variable in our framework. Because, $O|Y = y \sim N(y - \tau, \eta^2)$ before conditioning on the event in (3) and the selection event, $y+\omega > \tau$, is equivalent to: $o > 0$, the conditional likelihood for Y and O is given by:

$$\left(\bar{\Phi}\left(\frac{(\tau-\beta)}{\sqrt{(1+\eta^2)}}\right)\right)^{-1} \phi(Y; \beta, 1) \cdot \phi(O; Y - \tau, \eta^2) \cdot 1_{(0,\infty)}(O).$$

Marginalizing over the optimization variable O yields us a likelihood function of β, that is equal to:

$$\left(\bar{\Phi}\left(\frac{(\tau-\beta)}{\sqrt{(1+\eta^2)}}\right)\right)^{-1} \phi(Y; \beta, 1) \cdot \bar{\Phi}\left(\frac{1}{\eta}(\tau - Y)\right). \tag{4}$$

Compared to the conditional Gaussian law in the absence of randomization, $Y \mid Y > \tau$, (see, Example 2, Fithian, Sun, and Taylor 2014), a soft-truncating function replaces the indicator $1_{(\tau,\infty)}(Y)$. Hereafter, we refer to the resulting function as a "soft-truncated likelihood."

2.2. Selective MLE

We are now ready to discuss the maximizer of the soft-truncated likelihood in (4) and inspect some properties of this estimate that serve to motivate our approximate pivot in the article. Maximizing the log-likelihood gives us the "selective MLE," $\widehat{\beta}^{\,\text{mle}}$, based on the estimating equation:

$$\nabla\alpha(\widehat{\beta}^{\,\text{mle}}) = Y, \tag{5}$$

where

$$\alpha(\beta) = \frac{1}{2}\beta^2 + \log\bar{\Phi}\left(\frac{(\tau-\beta)}{\sqrt{(1+\eta^2)}}\right).$$

Turning to the distribution of $\widehat{\beta}^{\,\text{mle}}$, we obtain the density for the selective MLE from (4) by applying the simple variable transformation: $\widehat{\beta}^{\,\text{mle}} = \nabla\alpha^{-1}(Y)$, and note that this density is proportional to:

$$|\det(\nabla^2\alpha(\widehat{\beta}^{\,\text{mle}}))| \cdot \phi(\nabla\alpha(\widehat{\beta}^{\,\text{mle}}); \beta, 1) \cdot \bar{\Phi}\left(\frac{1}{\eta}(\tau - \nabla\alpha(\widehat{\beta}^{\,\text{mle}}))\right). \tag{6}$$

Proposition 2.1 obtains an upper bound for the mean squared error of the selective MLE. An immediate consequence of this bound is a global (asymptotic) consistency guarantee for the selective MLE; that is, the guarantee continues to hold for the event in (3) even when it has a vanishing probability as the sample size grows to infinity. Streamlining the main exposition to focus on finite sample results, we defer the proof for asymptotic consistency to the supplementary materials C.

Proposition 2.1. Fix $B = (1+\eta^2)^{-2}\eta^4$. Then, we have:

$$\mathbb{E}\left[(\widehat{\beta}^{\,\text{mle}} - \beta)^2 \mid Y + W > \tau\right] \leq (B)^{-1} \cdot \text{var}(Y \mid Y + W > \tau).$$

Examining for now the asymptotic behavior of selective MLE and the least squares estimate, and the role of the randomization variable W, we undertake a simulation by letting $Y := \sqrt{n}\bar{Y}_n$ with mean $\beta := \sqrt{n}\beta_n$. Figures 1–3 summarize the three primary take-aways from the simulation. We let $\tau = 0$ in (3), and fix $\beta_n = \beta_0 = -0.10$ which is highlighted in the figures via a dotted black line. Notice, our choice of β_n results in rarer events of selection with vanishing probabilities as $n \to \infty$. For the first two figures, the randomization variance η^2 is equal to 1. Based on the density in (6), Figure 1 first studies the behavior of the selective MLE. Matching our theoretical expectations, the plot demonstrates that the selective MLE is a consistent estimate for the parameter; more specifically, we observe a concentration

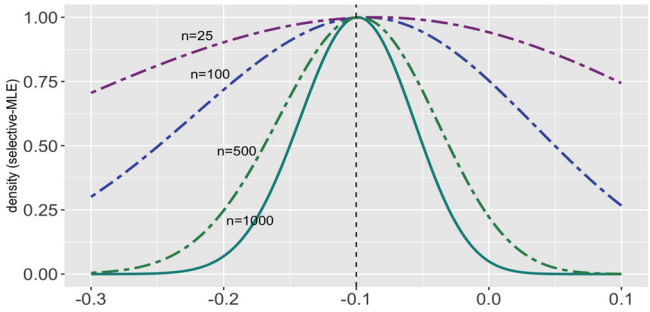

Figure 1. Distribution of the selective MLE under randomization variance $\eta^2 = 1$.

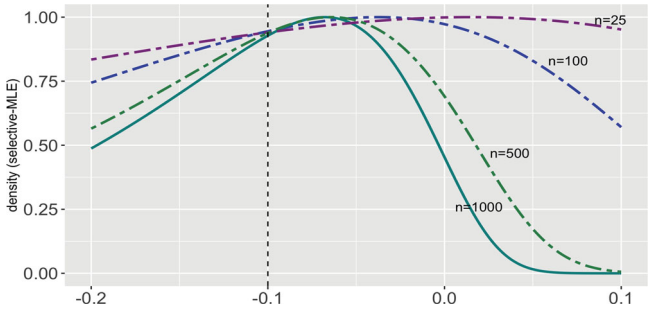

Figure 3. Distribution of the selective MLE under randomization variance $\eta^2 = 0.04$.

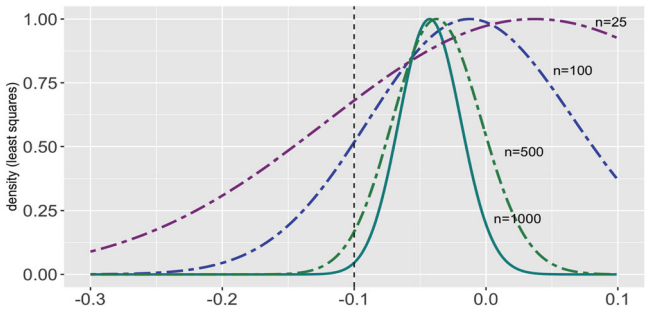

Figure 2. Distribution of the least squares estimate under randomization variance $\eta^2 = 1$.

of the estimate around β_n with increasing n. Next, Figure 2 replaces the selective MLE in the first plot with the least squares estimate. Unlike the selective MLE, the least squares estimate fails to concentrate around the parameter of interest for the same sample sizes. In Figure 3, we reproduce Figure 1, except now we study the behavior of the selective MLE under a very low value of randomization variance, $\eta^2 = 0.04$. An empirical affirmation of the merits of randomization, this plot shows that selective MLE fails to concentrate around β_n in the (almost) absence of randomization.

2.3. Approximate Pivot

Prompted by a concentration of the selective MLE around the parameter of interest, we introduce an approximate pivot in the current section. We propose to approximate the distribution of the MLE by a Gaussian distribution with: (a) mean β, and (b) variance equal to inverse of the observed Fisher information, $I(\widehat{\beta}^{\,\mathrm{mle}})$. Taking a second derivative of the log-likelihood in (4) at the MLE gives us the value of $I(\widehat{\beta}^{\,\mathrm{mle}})$, which is equal to:

$$1 - \frac{(\widehat{\beta}^{\,\mathrm{mle}} - \tau)}{(1+\eta^2)^{3/2}} \cdot \left(\bar{\Phi}\left(\frac{(\tau - \widehat{\beta}^{\,\mathrm{mle}})}{\sqrt{(1+\eta^2)}}\right)\right)^{-1} \phi\left(\frac{(\tau - \widehat{\beta}^{\,\mathrm{mle}})}{\sqrt{(1+\eta^2)}}; 0, 1\right)$$
$$- \frac{1}{(1+\eta^2)} \cdot \left(\bar{\Phi}\left(\frac{(\tau - \widehat{\beta}^{\,\mathrm{mle}})}{\sqrt{(1+\eta^2)}}\right)\right)^{-2} \phi^2\left(\frac{(\tau - \widehat{\beta}^{\,\mathrm{mle}})}{\sqrt{(1+\eta^2)}}; 0, 1\right). \quad (7)$$

The Gaussian approximation described above gives rise to the approximate pivot:

$$\bar{\Phi}\left(\sqrt{I(\widehat{\beta}^{\,\mathrm{mle}})}(\widehat{\beta}^{\,\mathrm{mle}} - \beta)\right). \quad (8)$$

We remark that the distribution of the selective MLE, characterized exactly by the density in (6), can yield us exact maximum likelihood inference. In contrast, the pivot in (8) is only approximate in nature, but, appealingly simple in form. Inference based on the approximate pivot requires us to compute two estimates from the soft-truncated likelihood, namely, the selective MLE and the observed Fisher information.

Before turning to the general development, we explore if the proposed Gaussian approximation mimics the exact distribution of the selective MLE. In Figure 4, we represent the density of the selective MLE in (6), our benchmark, by the gray curve. The panel with $\beta = -3$ results in a rare selection event, while the panel with $\beta = 1.5$ results in a highly probable selection event with little selection bias. Noteworthy, the effectiveness of our pivot is highlighted via a strong agreement of the proposed (approximate) Gaussian density with the exact (benchmark) density of the selective MLE.

3. Maximum Likelihood Inference Post Convex Queries

We develop our method of maximum likelihood inference below, focusing on the randomized LASSO as our leading example. In the supplementary materials, we show that the form of our estimating equations in the primary example generalizes directly to other convex learning queries whose solutions can be similarly characterized through affine Karush-Kuhn-Tucker (K.K.T.) conditions of optimality.

3.1. Framework under Linear Regression

Consider solving the randomized LASSO in (1), where y and ω denote the observed instances of our response variable $Y \in \mathbb{R}^n$ and randomization variable $W \sim N(0_p, \Sigma_{\mathbb{W}}) \in \mathbb{R}^p$, respectively. Let $\widehat{\mathrm{E}}(y, \omega) \subseteq \{1, 2, \ldots, p\}$ denote the active set of variables selected by the randomized LASSO. At the solution of the randomized query, we record the value of the subgradient vector for the ℓ_1 penalty which we represent by $\widehat{\mathrm{S}}(y, \omega)$. Notice, the collection of instances which lead us to observe $\widehat{\mathrm{S}}(y, \omega) = \mathrm{S}$ result in the active set $\widehat{\mathrm{E}}(y, \omega) = \mathrm{E}$, that is,

$$\left\{(y, \omega) \in \mathbb{R}^n \times \mathbb{R}^p : \widehat{\mathrm{S}}(y, \omega) = \mathrm{S}\right\}$$
$$\subseteq \left\{(y, \omega) \in \mathbb{R}^n \times \mathbb{R}^p : \widehat{\mathrm{E}}(y, \omega) = \mathrm{E}\right\}.$$

We turn to a framework for selective inference, allowing our model and parameters to depend on data through the recorded output of our randomized query, S. Consider a prespecified

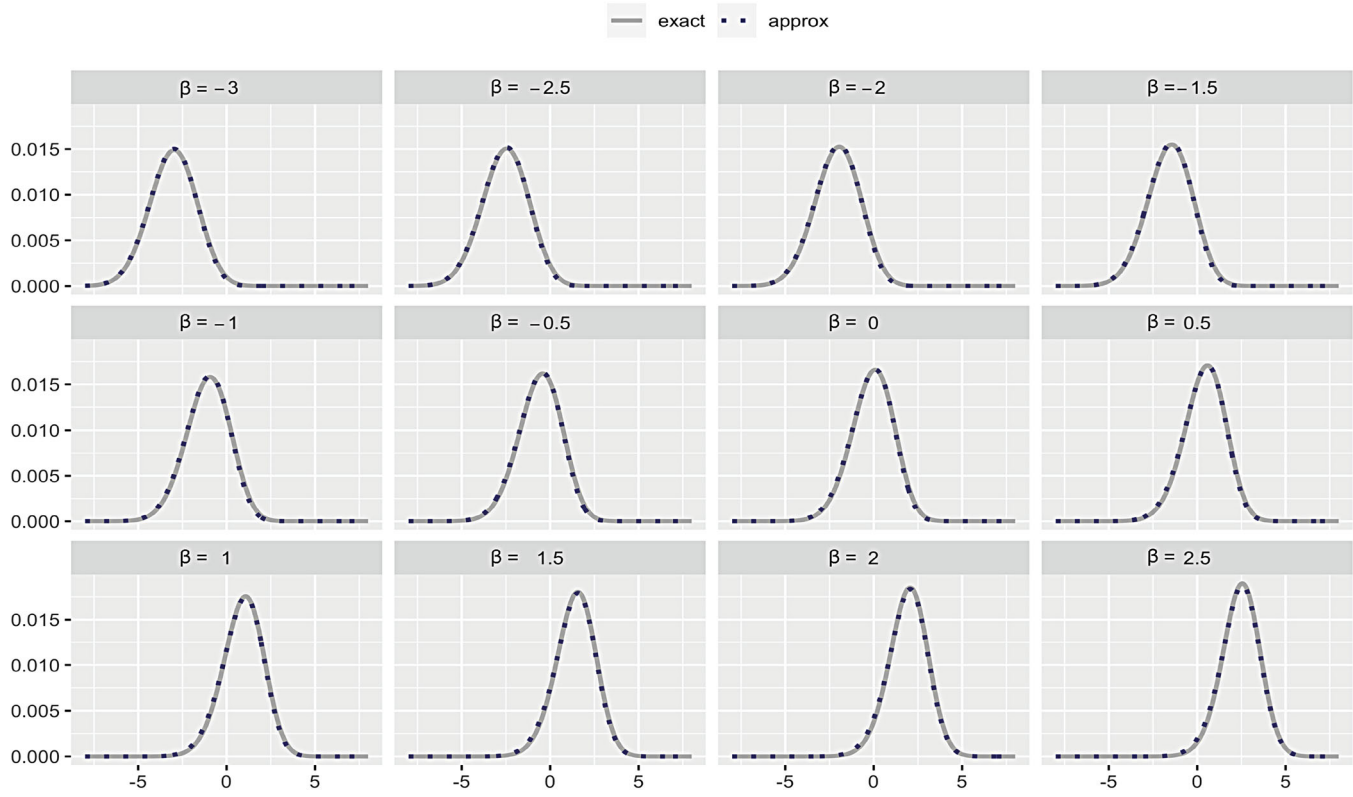

Figure 4. The blue curve represents the normal approximation $N(\beta, I^{-1}(\widehat{\beta}^{\text{mle}}))$ and the gray curve plots the exact density of the MLE in (6).

mapping $\mathcal{H} : \mathrm{S} \to \mathcal{E} \subseteq \{1, 2, \ldots, p\}$. Our model, after observing $\widehat{\mathrm{S}}(y, \omega) = \mathrm{S}$ and subsequently, noting $\mathcal{E} = \mathcal{H}(\mathrm{S})$, is given by

$$\mathbb{M}_{\mathrm{S}} = \left\{ Y \sim N_n(X_{\mathcal{E}}\beta_{\mathcal{E}}, \sigma^2 I), \quad \beta_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|} \right\} \text{ for a fixed } \sigma \in \mathbb{R}^+. \tag{9}$$

The mapping $\mathcal{H}$ grants us the flexibility to use (arbitrary) linear models informed by S, including the selected model in the special case when $\mathcal{E} = \mathrm{E}$, the active set of variables. Suppose, we have a matrix $\mathcal{F}_{\mathrm{S}} \in \mathbb{R}^{d \times n}$, that is allowed to depend on data through S. Then, let

$$\beta_{\mathbb{M}_{\mathrm{S}},\mathrm{S}} = \mathcal{F}_{\mathrm{S}}\mathbb{E}[Y] \in \mathbb{R}^d \tag{10}$$

be our parameter vector of inferential interest.

In the next step, we form a (multivariate) likelihood function of $\beta_{\mathbb{M}_{\mathrm{S}},\mathrm{S}}$ in the learned model $\mathbb{M}_{\mathrm{S}}$. To do so, for a fixed value S, we consider the following statistic:

$$\widehat{\beta}_{\mathrm{S}} \sim N(\beta_{\mathbb{M}_{\mathrm{S}},\mathrm{S}}, \Sigma_{\mathbb{M}_{\mathrm{S}},\mathrm{S}}).$$

We derive our soft-truncated likelihood by conditioning the Gaussian law of $\widehat{\beta}_{\mathrm{S}}$ upon the selection event:

$$\left\{ (y, \omega) \in \mathbb{R}^n \times \mathbb{R}^p : \widehat{\mathrm{S}}(y, \omega) = \mathrm{S} \right\}. \tag{11}$$

The event in (11) depends not just on $\widehat{\beta}_{\mathrm{S}}$, but further involves a statistic independent of $\widehat{\beta}_{\mathrm{S}}$, which we represent by $\widehat{\beta}^{\perp}_{\mathrm{S}}$. In addition to the above event, we condition on $\widehat{\beta}^{\perp}_{\mathrm{S}}$ to eliminate nuisance parameters from the likelihood.

For ease of exposition, hereafter, we specialize the above framework to a projected parameter in the selected model

$$\left\{ Y \sim N_n(X_{\mathrm{E}}\beta_{\mathrm{E}}, \sigma^2 I), \ \beta_{\mathrm{E}} \in \mathbb{R}^{|\mathrm{E}|} \right\}, \tag{12}$$

which we obtain by applying the specific mapping $\mathcal{H}(\mathrm{S}) = \mathrm{E}$ and fixing $\mathcal{F}_{\mathrm{S}} = (X_{\mathrm{E}}^{\mathsf{T}}X_{\mathrm{E}})^{-1}X_{\mathrm{E}}^{\mathsf{T}} \in \mathbb{R}^{|\mathrm{E}| \times n}$. As noted in Berk et al. (2013) and Lee et al. (2016), our parameter for inference in this model is the projection of the mean for Y onto the subspace spanned by the columns in X_{E}. Immediately, we recognize: $\widehat{\beta}_{\mathrm{S}} = (X_{\mathrm{E}}^{\mathsf{T}}X_{\mathrm{E}})^{-1}X_{\mathrm{E}}^{\mathsf{T}}y$, the least squares statistic refitted to y and X_{E}. Besides the least squares statistic, the likelihood involves

$$\widehat{\beta}^{\perp}_{\mathrm{S}} = -X^{\mathsf{T}}(y - X_{\mathrm{E}}\widehat{\beta}_{\mathrm{S}}),$$

which is independent of $\widehat{\beta}_{\mathrm{S}}$ under the model in (12); we will detail this out in the following section.

3.2. Multivariate Soft-Truncated Likelihood

Our main result in the section, Theorem 1, obtains a soft-truncated likelihood function. As seen in the file drawer example, we begin with a compact representation for our selection event in terms of optimization variables based on the randomized LASSO solution. Introducing some more notations, we let $O_1 \in \mathbb{R}^{|\mathrm{E}|}$ and $O_2 \in \mathbb{R}^{p-|\mathrm{E}|}$ represent the active (nonzero) components of randomized LASSO solution and the subgradient (sub-)vector for the ℓ_1 penalty at the inactive indices in E^c, respectively, and let o_1 and o_2 be the observed instances for these variables. Let $z_{\mathrm{E}} = \mathrm{sign}(o_1) \in \mathbb{R}^{|\mathrm{E}|}$ be the sign vector for the active components of the estimated LASSO solution. The K.K.T. conditions for the randomized LASSO are given by

$$\omega = (\omega_{\mathrm{E}}^{\mathsf{T}} \ \omega_{E^c}^{\mathsf{T}})^{\mathsf{T}} = -X^{\mathsf{T}}X_{\mathrm{E}}\widehat{\beta}_{\mathrm{S}} + \begin{bmatrix} X_{\mathrm{E}}^{\mathsf{T}}X_{\mathrm{E}} + \epsilon I \\ X_{\mathrm{E}^c}^{\mathsf{T}}X_{\mathrm{E}} \end{bmatrix} o_1 + \begin{pmatrix} \lambda z_{\mathrm{E}} \\ o_2 \end{pmatrix} + \widehat{\beta}^{\perp}_{\mathrm{S}},$$

where: $-\text{diag}(z_E)\ o_1 < 0$, and $\|o_2\|_\infty < \lambda$. Because, $S = \begin{pmatrix}\lambda z_E^\intercal & o_2^\intercal\end{pmatrix}^\intercal$, our selection event in (11) is equivalent to the $|E|$ linear constraints: $Uo_1 < v$, for the fixed matrices $U = -\text{diag}(z_E)$, $v = 0_{|E|}$. We note a resemblance with the file drawer example, wherein the selection event is equivalent to the single linear constraint: $o > 0$.

Fixing the matrices:

$$P_S = -X^\intercal X_E,\ Q_S = \begin{bmatrix} X_E^\intercal X_E + \epsilon I \\ X_{E^c}^\intercal X_E \end{bmatrix},\ r_S = \begin{pmatrix}\lambda z_E \\ o_2\end{pmatrix} + \widehat{\beta}_S^\perp,$$

we rewrite the stationary mapping in the K.K.T. condition as

$$\omega = \begin{pmatrix}\omega_E^\intercal \omega_{E^c}^\intercal\end{pmatrix}^\intercal = P_S\widehat{\beta}_S + Q_S o_1 + r_S. \tag{13}$$

Based on

$$\bar{\Sigma}^{-1} = Q_S^\intercal \Sigma_{\mathbb{W}}^{-1} Q_S,\ A = -\bar{\Sigma} Q_S^\intercal \Sigma_{\mathbb{W}}^{-1} P_S,\ b = -\bar{\Sigma} Q_S^\intercal \Sigma_{\mathbb{W}}^{-1} r_S,$$

define:

$$f(\widetilde{\beta}_S) = \int \phi(o_1; A\widetilde{\beta}_S + b, \bar{\Sigma}) \cdot 1_{R_0}(o_1) do_1, \tag{14}$$

where $R_0 = \{o_1 \in \mathbb{R}^{|E|} : Uo_1 < v\}$.

Theorem 1. After conditioning the law of $\widehat{\beta}_S$ upon $\widehat{S}(Y, W) = S$ and $\widehat{\beta}_S^\perp(Y) = \widehat{\beta}_S^\perp$, the soft-truncated likelihood is

$$\left(\int \phi(\widetilde{\beta}_S; J\beta_{\mathbb{M}_S,S} + k, \Sigma) \cdot f(\widetilde{\beta}_S) d\widetilde{\beta}_S\right)^{-1} \phi(\widehat{\beta}_S; J\beta_{\mathbb{M}_S,S} + k, \Sigma) \cdot f(\widehat{\beta}_S),$$

where the matrices Σ, J, k are equal to:

$$\Sigma = (\Sigma_{\mathbb{M}_S,S}^{-1} + P_S^\intercal \Sigma_{\mathbb{W}}^{-1} P_S - A^\intercal \bar{\Sigma}^{-1} A)^{-1},$$
$$J = \Sigma \Sigma_{\mathbb{M}_S,S}^{-1},\ k = \Sigma(A^\intercal \bar{\Sigma}^{-1} b - P_S^\intercal \Sigma_{\mathbb{W}}^{-1} r_S).$$

3.3. Approximate Inference

The likelihood in Theorem 1, though exact, does not directly result in tractable estimating equations for the maximum likelihood estimate and the observed Fisher information matrix. This is because the normalizer for the soft-truncated likelihood lacks a closed-form expression. To circumvent the problem, we propose an approximate proxy for our soft-truncated likelihood based on an upper bound for the normalizer in Proposition 3.1. Later in the supplementary materials, using a large deviations principle, we prove that the approximate proxy converges to the exact likelihood with increasing sample size. Furthermore, we show the maximizer of the approximate likelihood, $\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S}$, is guaranteed to concentrate around the parameter in (10). Endowed with a property we expect with the exact MLE, we call the maximizer of the approximate likelihood an "approximate selective MLE."

Proposition 3.1. Let R be a convex and compact subset of $\mathbb{R}^{|E|} \times \mathbb{R}^{|E|}$. Suppose, $\widehat{\beta}_S$ and O_1 are drawn from a Gaussian distribution with the following likelihood:

$$\phi(\widehat{\beta}_S; J\beta_{\mathbb{M}_S,S} + k, \Sigma) \cdot \phi(O_1; A\widehat{\beta}_S + b, \bar{\Sigma}).$$

Then, $\log \mathbb{P}\left[\begin{pmatrix}\widehat{\beta}_S^\intercal, O_1^\intercal\end{pmatrix}^\intercal \in R\right]$ is bounded from above by

$$-\inf_{(\widetilde{\beta}_S, o_1) \in R} \left\{\frac{1}{2}(\widetilde{\beta}_S - J\beta_{\mathbb{M}_S,S} - k)^\intercal \Sigma^{-1}(\widetilde{\beta}_S - J\beta_{\mathbb{M}_S,S} - k) + \frac{1}{2}(o_1 - A\widetilde{\beta}_S - b)^\intercal \bar{\Sigma}^{-1}(o_1 - A\widetilde{\beta}_S - b)\right\}.$$

Recall, $R_0 = \{o_1 \in \mathbb{R}^{|E|} : Uo_1 < v\}$. We apply the bound in Proposition 3.1 to obtain the following proxy for the exact log-likelihood:

$$\log \phi(\widehat{\beta}_S; J\beta_{\mathbb{M}_S,S} + k, \Sigma) + \inf_{(\widetilde{\beta}_S, o_1) \in \mathbb{R}^{|E|} \times R_0} \left\{\frac{1}{2}(\widetilde{\beta}_S - J\beta_{\mathbb{M}_S,S} - k)^\intercal \Sigma^{-1}(\widetilde{\beta}_S - J\beta_{\mathbb{M}_S,S} - k) + \frac{1}{2}(o_1 - A\widetilde{\beta}_S - b)^\intercal \bar{\Sigma}^{-1}(o_1 - A\widetilde{\beta}_S - b)\right\},$$

after ignoring constants free of the parameter vector.

Remark 1. Observe, $\mathbb{R}^{|E|} \times R_0$, the subset of $\mathbb{R}^{|E|} \times \mathbb{R}^{|E|}$ associated with the our selection event is clearly not compact. While compactness is a requirement to prove that the approximation in Proposition 3.1 is an upper bound for the normalizer of the likelihood, in practice, we may consider a sufficiently large compact, convex subset such that the probability of the associated event converges to the actual probability with increasing sample size.

As noted in Panigrahi and Taylor (2018), we can further modify the approximation in Proposition 3.1 to solve an unconstrained optimization via a barrier penalty that reflects the same constraints, but allocates a higher preference to the optimizing variables within the selection region. Letting $\mathcal{B}_{U;v}(o_1)$ denote a barrier penalty for the constraints $Uo_1 < v$, the final expression for our approximate log-likelihood agrees up to an additive constant with

$$\log \phi(\widehat{\beta}_S; J\beta_{\mathbb{M}_S,S} + k, \Sigma) + \inf_{(\widetilde{\beta}_S, o_1)} \left\{\frac{1}{2}(\widetilde{\beta}_S - J\beta_{\mathbb{M}_S,S} - k)^\intercal \Sigma^{-1}(\widetilde{\beta}_S - J\beta_{\mathbb{M}_S,S} - k) + \frac{1}{2}(o_1 - A\widetilde{\beta}_S - b)^\intercal \bar{\Sigma}^{-1}(o_1 - A\widetilde{\beta}_S - b) + \mathcal{B}_{U;v}(o_1)\right\}. \tag{15}$$

Based on the approximate likelihood in (15), the results in Theorems 2 and 4 give us compact estimating equations for the two ingredients of approximate maximum likelihood inference.

Theorem 2. Consider the optimization problem

$$o_1^*(\widehat{\beta}_S) = \underset{o_1}{\text{argmin}} \frac{1}{2}(o_1 - A\widehat{\beta}_S - b)^\intercal \bar{\Sigma}^{-1}(o_1 - A\widehat{\beta}_S - b) + \mathcal{B}_{U;v}(o_1). \tag{16}$$

Then, maximizing the approximate log-likelihood in (15) yields us the following estimating equation for the approximate selective MLE:

$$\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S} = J^{-1}\widehat{\beta}_S - J^{-1}k + \Sigma_{\mathbb{M}_S,S} A^\intercal \bar{\Sigma}^{-1}(A\widehat{\beta}_S + b - o_1^*(\widehat{\beta}_S)).$$

In line with Proposition 2.1 for the file drawer example, Theorem 3 provides a bound for the mean squared error of the approximate selective MLE. The bound in this result allows us to formalize a global consistency guarantee for our estimate in supplementary material C.

Theorem 3. Let the smallest eigen values for $(\Sigma^{-1}_{\mathbb{M}_S,S} + P_S^\intercal \Sigma^{-1}_{\mathbb{W}} P_S)^{-1}$ and $\Sigma^{-1}_{\mathbb{M}_S,S}$ be λ_0 and λ_1, respectively. Fix $B = (\lambda_0 \cdot \lambda_1)^2$. Based on the real-valued mapping:

$$\alpha(\eta_{\mathbb{M}_S,S}) = \frac{1}{2}\eta^\intercal_{\mathbb{M}_S,S}\Sigma\eta_{\mathbb{M}_S,S} - \inf_{(\widetilde{\beta}_S,o_1)\in\mathbb{R}^{|E|}\times R_0} \Big\{\frac{1}{2}(\widetilde{\beta}_S - \Sigma\eta_{\mathbb{M}_S,S})^\intercal\Sigma^{-1}(\widetilde{\beta}_S - \Sigma\eta_{\mathbb{M}_S,S}) + \frac{1}{2}(o_1 - A\widetilde{\beta}_S - b)^\intercal\bar{\Sigma}^{-1}(o_1 - A\widetilde{\beta}_S - b) + \mathcal{B}_{U;\nu}(o_1)\Big\},$$

we have

$$\mathbb{E}\Big[\|\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S} - \beta_{\mathbb{M}_S,S}\|_2^2 \mid \widehat{S}(Y,W) = S,\ \widehat{\beta}^\perp_S(Y) = \widehat{\beta}^\perp_S\Big] \le (B)^{-1}\mathbb{E}\big[\|\widehat{\beta}_S - \nabla\alpha(\Sigma^{-1}(J\beta_{\mathbb{M}_S,S} + k))\|_2^2 \mid \widehat{S}(Y,W) = S,\ \widehat{\beta}^\perp_S(Y) = \widehat{\beta}^\perp_S\big].$$

We provide a proxy for the observed Fisher information matrix based on the estimate in Theorem 2.

Theorem 4. Let $o_1^*(\widehat{\beta}_S)$ be the solution to the optimization problem in (16). The observed Fisher information $I(\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S})$ for the approximate log-likelihood in (15) is

$$\Sigma^{-1}_{\mathbb{M}_S,S}\left(\Sigma^{-1} + A^\intercal\bar{\Sigma}^{-1}A - A^\intercal\bar{\Sigma}^{-1}(\bar{\Sigma}^{-1} + \nabla^2\mathcal{B}_{U;\nu}(o_1^*(\widehat{\beta}_S)))^{-1}\bar{\Sigma}^{-1}A\right)^{-1}\Sigma^{-1}_{\mathbb{M}_S,S}.$$

We summarize in Algorithm 1 our steps for maximum likelihood inference. Our primary computational step is the simple, $\mathbb{R}^{|E|}$-dimensional, convex optimization problem **(O)** resulting in **(S-MLE)** and **(FI)**. Emphasized earlier, the form of the estimating equations for the randomized LASSO generalizes to convex queries with affine K.K.T. conditions of optimality as in (13). In supplementary materials B.2, we illustrate how our method applies to: (a) variable screening based on marginal correlations (Lee and Taylor 2014); (b) variable selection via SLOPE (Bogdan et al. 2015).

Algorithm 1 ALGORITHM 1: Approximate maximum likelihood inference post a convex query

Require: Query, $\omega \sim N(0, \Sigma_{\mathbb{W}})$
Observe: $\widehat{S} = S$
Implied parameters **(P)**: Compute matrices: $\bar{\Sigma}, A, b, \Sigma, J, k$
Optimization **(O)**: $o_1^*(\widehat{\beta}_S) = \underset{o_1}{\text{argmin}}\frac{1}{2}(o_1 - A\widehat{\beta}_S - b)^T\bar{\Sigma}^{-1}(o_1 - A\widehat{\beta}_S - b) + \mathcal{B}_{U;\nu}(o_1)$.
Selective MLE **(S-MLE)**: $\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S} = J^{-1}\widehat{\beta}_S - J^{-1}k + \Sigma_{\mathbb{M}_S,S}A^\intercal\bar{\Sigma}^{-1}(A\widehat{\beta}_S + b - o_1^*(\widehat{\beta}_S))$
Inverse info **(FI)**: $I^{-1}(\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S}) = \Sigma_{\mathbb{M}_S,S}\Big(\Sigma^{-1} + A^T\bar{\Sigma}^{-1}A - A^T\bar{\Sigma}^{-1}(\bar{\Sigma}^{-1} + \nabla^2\mathcal{B}_K (o_1^*(\widehat{\beta}_S)))^{-1}\bar{\Sigma}^{-1}A\Big)\Sigma_{\mathbb{M}_S,S}$
MLE-based inference:
for all j in selected set E **do**
(*p*-value for $\beta_{j;\mathbb{M}_S,S}$**)** : $2\min\Big(\bar{\Phi}\Big(\widehat{\beta}^{\text{mle}}_{j;\mathbb{M}_S,S}/\sqrt{I^{-1}_{j,j}(\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S})}\Big), \Phi\Big(\widehat{\beta}^{\text{mle}}_{j;\mathbb{M}_S,S}/\sqrt{I^{-1}_{j,j}(\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S})}\Big)\Big)$
(interval for $\beta_{j;\mathbb{M}_S,S}$**)** : $\Big(\widehat{\beta}^{\text{mle}}_{j;\mathbb{M}_S,S} - z_{1-q/2}\cdot\sqrt{I^{-1}_{j,j}(\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S})}, \widehat{\beta}^{\text{mle}}_{j;\mathbb{M}_S,S} + z_{1-q/2}\cdot\sqrt{I^{-1}_{j,j}(\widehat{\beta}^{\text{mle}}_{\mathbb{M}_S,S})}\Big)$
end for

4. Simulation Experiments

We explore the potential of our method for a wide range of signal-to-noise ratio (SNR) values in a linear regression setting. Following closely the setup in Hastie, Tibshirani, and Tibshirani (2017), in every round of experiment, we simulate each $\mathbb{R}^p$-valued row of the design matrix X from $N(0, \Sigma(\rho))$ where the (i,j)th entry of $\Sigma(\rho) = \rho^{|i-j|}$. We then draw the response as $Y|X \sim N(X\beta, \sigma^2 I)$. Fixing $n = 300$, $p = 100$, $\rho = 0.35$, we consider a linearly-varying coefficient vector with $s = 6$ nonzero equally spaced components that have magnitudes: $-10,-6,-2,2,6,10$. We vary the noise level σ^2 to match the SNR value: $\text{SNR} = \sigma^{-2}\cdot(\beta^\intercal\Sigma\beta)$, and vary SNR in the set $\{0.15, 0.21, 0.26, 0.31, 0.42, 0.71, 1.22, 2.07, 3.52\}$.

Dividing our experiments into two regimes, namely randomized and nonrandomized, we run a canonical LASSO query (without randomization)

$$\underset{o\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|y - Xo\|_2^2 + \lambda\|o\|_1, \tag{17}$$

and a randomized LASSO query in (1) with $\omega \sim N(0, \eta^2 I_p)$ and $\epsilon = n^{-1/2}$. The randomization variance η^2 is chosen so that $(\hat{\sigma})^{-2}\eta^2 = 0.50$, using the estimated noise level in the data: $\hat{\sigma}^2 = (n-p)^{-1}\|(I - X(X^TX)^{-1}X^T)y\|^2$. Based on an asymptotic equivalence between data splitting and a Gaussian randomization scheme (see Proposition 4.1, Panigrahi, Taylor, and Weinstein 2021), our choice of randomization variance roughly matches the amount of information used up in selection when two-thirds of the samples are allocated for the LASSO. For each query, we carry out three common schemes to choose λ. Our first choice is a theoretical value proposed in Negahban et al. (2009) and is given by $\lambda_{\text{theory}} = \mathbb{E}[\|X^T\Psi\|_\infty]$ where $\Psi \sim \mathbb{N}(0, \hat{\sigma}^2 I)$. Our second and third choices are obtained from cross-validation. Denoted by $\lambda_{\text{cv.min}}$ and $\lambda_{\text{cv.1se}}$, the tuning parameters are associated with the lowest cross-validated error and error within 1 standard error of the best model, respectively.

4.1. Methods and Metrics

In our experiments, we illustrate maximum likelihood inference for two sets of parameter vectors after selection: (a) the partial regression coefficients in the selected model, obtained by letting $\mathcal{H}(S) = E$ and $\mathcal{F}_S = (X_E^\intercal X_E)^{-1}X_E^\intercal$; (b) the selected set of parameters in the full model, obtained by letting $\mathcal{H}(S) = \{1, 2, \ldots, p\}$ and $\mathcal{F}_S = \mathcal{L}_E(X^TX)^{-1}X^\intercal$, where $\mathcal{L}_E \in \mathbb{R}^{|E|\times p}$ is a matrix of all zeros except for the indices

$$\mathcal{L}_E(k, j_k) = 1 \quad \text{for} \quad k \in \{1, \ldots, |E|\},\ E = \{j_1, \ldots, j_{|E|}\}.$$

We call the former parameter vector "Partial" and the latter "Full" in our depictions.

In the linear regression setting described above, we compare the relative performance of our proposed method with "Lee

et al.," "Liu et al.," "Split," and "Naive." Our method follows Algorithm 1 with the barrier penalty

$$\mathcal{B}_{U;v}(o_1) = \sum_j \log\left(1 + \frac{1}{v_j - U_j^\intercal o_1}\right),$$

where U_j is the jth row of U and v_j is the jth component of v. First, we report the average coverage of the interval estimates produced by each strategy across all simulations. Each simulation records the proportion of intervals that cover our target parameters in a single round of experiment. The nominal level of FCR aimed by the interval estimates is 10%. Note, selective inference produced by "Liu et al." is tied only to the "Full" parameter vector. The coverage for "Naive" intervals, not adjusted for any selection, underscores the extent of selection bias for a specific value of SNR. Next, we provide a breakdown of all methods in terms of their inferential power. To this end, we record the average lengths of the interval estimates and their power which we define to be the proportion of signals detected by a strategy from the ones successfully screened by the query. Detected variables here count the variables for which the corresponding interval estimates do not cover zero. Faced with two different regimes, we depict the power comparisons for the randomized and nonrandomized estimates under certain best-case scenarios for each estimate. The best-case scenarios in our experiment are led by an assessment of predictive risks for a point estimate associated with each strategy, measured through the relative risk metric:

$$\mathcal{R}(\widehat{o}^\lambda, \beta) = (\beta^\intercal \Sigma \beta)^{-1} \cdot \left\{(\widehat{o}^\lambda - \beta)^\intercal \Sigma (\widehat{o}^\lambda - \beta)\right\},$$

where $\widehat{o}^\lambda$ is the estimate and β is the parameter vector. We consider the LASSO solution as a natural point estimate for the parameter vector β in the nonrandomized regime and associate the LASSO estimate with both strategies "Lee et al." and "Liu et al.". For "Split," we consider the least squares estimate obtained after refitting the selected model to the remaining one-third of the data samples and append it with zeros for the indices not selected by the LASSO. Note, the selective MLE, an immediate byproduct of Algorithm 1, appended with zeros for the inactive indices serves as a point estimate for our proposal. We discuss the detailed findings of our experiments next.

4.2. Findings and Interpretation

Supporting the validity of inference after selection, Figure 5 highlights the performance of the different interval estimates for the "Full" and "Partial" parameters. The three columns in the plot are associated with the three different choices of tuning parameter: λ_{theory}, $\lambda_{\text{cv.1se}}$, and $\lambda_{\text{cv.min}}$. We remark that none of the methods adjust for the adaptivity involved in the choice of the cross-validated tuning parameters. Coherent with expectations, the interval estimates for all the methods approximately attain the nominal FCR level 10% at λ_{theory}; "Lee et al." fails to yield valid inference at $\lambda_{\text{cv.1se}}$ and $\lambda_{\text{cv.min}}$ since it does not account for the fact that the tuning parameters for the LASSO were chosen based on the specific data through cross-validation. The effectiveness of the normal approximation for the proposed method "MLE" is largely ascribed to the soft-truncated likelihood, due to the use of randomization in the LASSO query. Besides, the accuracy of the large deviations-type approximation for the exact likelihood is maintained under moderate dimensions. Interestingly, "MLE" and "Liu et al." recover the nominal levels even at the cross validated tuning parameter; this is seen in the second and third columns of the plot. Although there lacks a formal justification for this observation, heuristically, the use of randomization in the LASSO query limits the role of the data-dependent regularization. A similar justification possibly holds

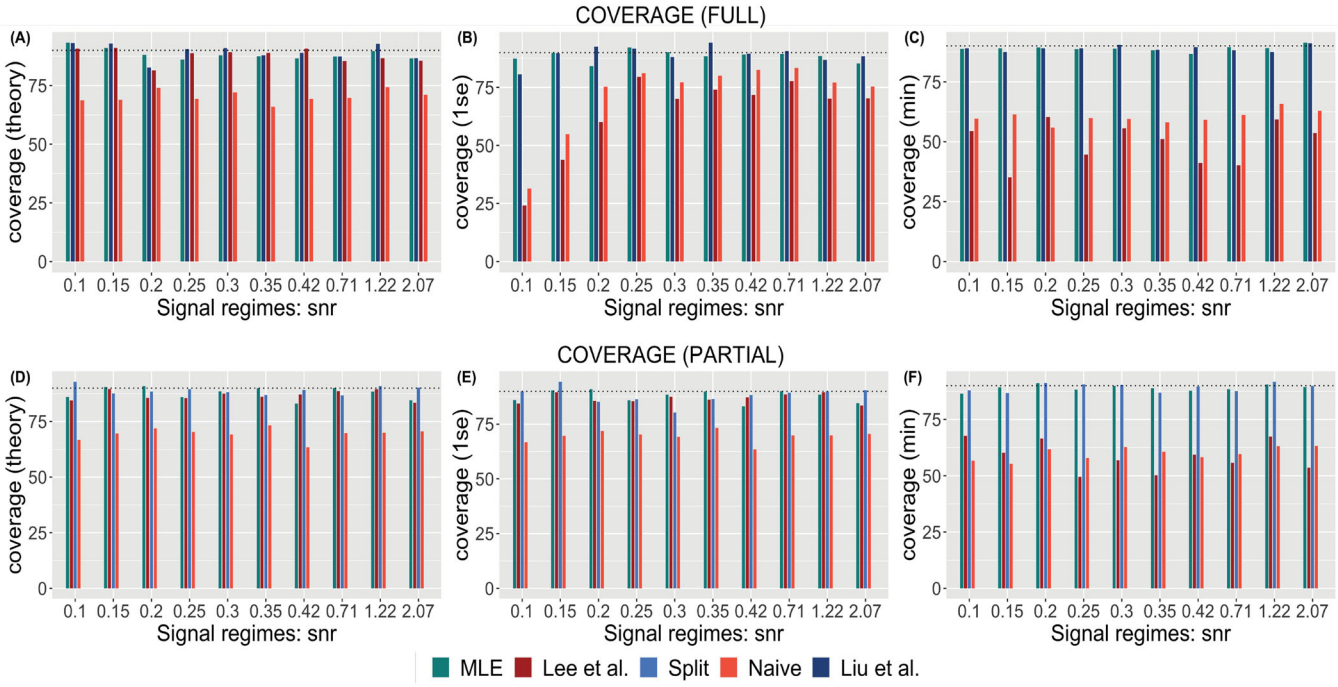

Figure 5. Averaged coverage of interval estimates. The nominal target for coverage is 90%, marked by the dotted horizontal line. (A)–(C) depict coverage for "Full" parameters; (D)–(F) depict coverage for "Partial" parameters.

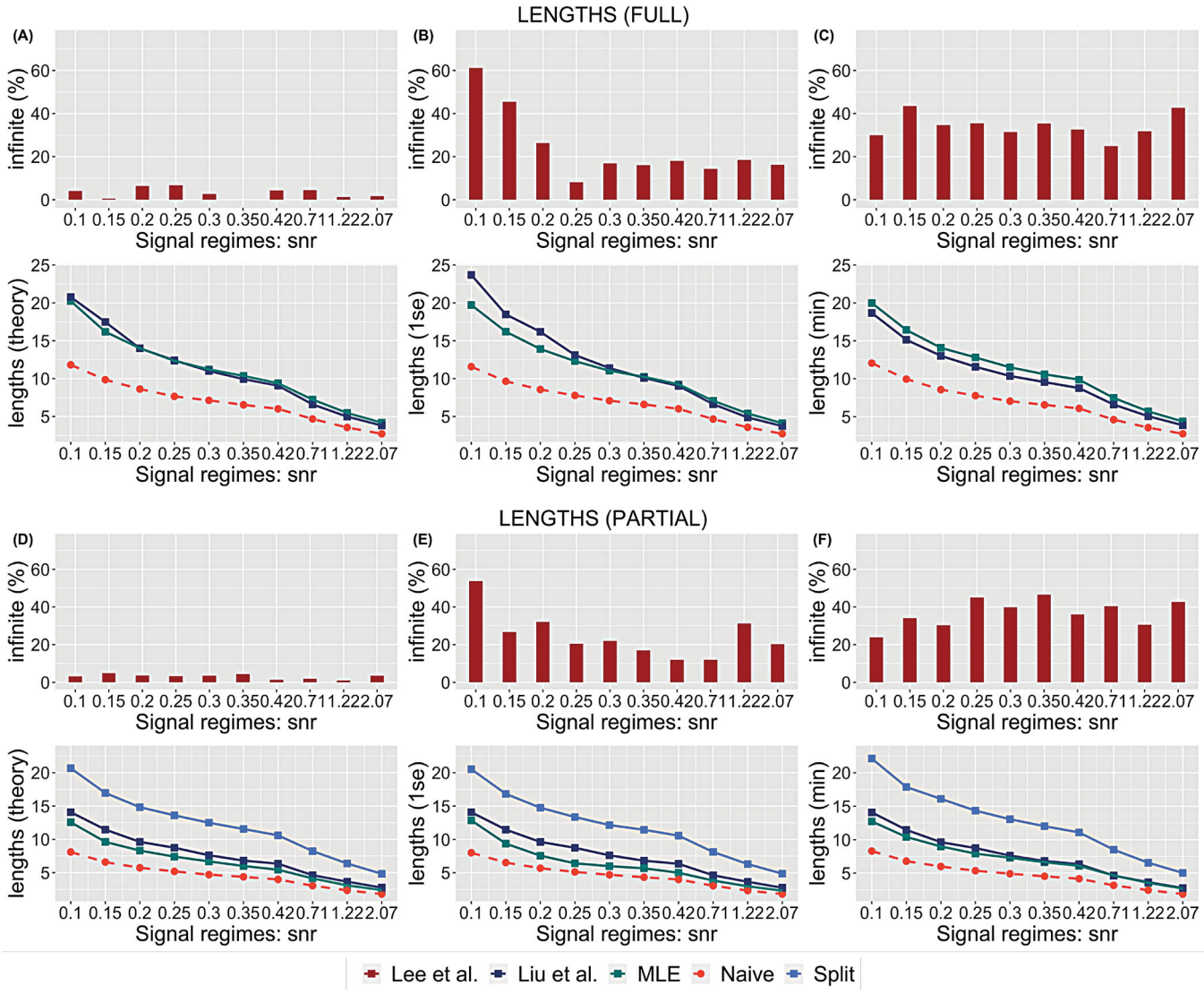

Figure 6. Average lengths of interval estimates. (A)–(C) depict the averaged lengths of intervals for "Full" parameters; (D)–(F) depict the averaged lengths of intervals for "Partial" parameters. The red bars on the top row of each panel reports the percentage of intervals by "Lee et al." which resulted in infinite length.

for "Liu et al.," which achieves the same goal by constructing inference for parameters that are less affected by selection.

Figure 6 highlights the averaged lengths of the interval estimates produced by different methods. The bars in red depict the percentage of intervals with infinite length for "Lee et al.," confirming the conclusions in Kivaranovic and Leeb (2018). The interval estimates produced by all other methods are bounded in length. Consistent with the example presented in the introduction, the interval estimates based on "Liu et al." and "MLE" are comparable for the "Full" parameters. The estimates for the "Partial" parameters produced by "MLE" are shorter than those for the "Full" parameters; this gain in power is in part due to inference in the learned model as opposed to the full model. Assuredly, the new proposal dominates the simple "Split" estimates which roughly use the same information in selection as the randomized query in our method.

Emphasized earlier, for a fair comparison of power between the randomized and nonrandomized estimates, we use their relative risks to guide us to a best-case scenario within each regime. Purely from a predictive lens, the risk assessment across the different values of SNR suggest running the randomized LASSO at $\lambda_{\text{cv.1se}}$ and the canonical LASSO at $\lambda_{\text{cv.min}}$. Taking on direct comparisons for "MLE" after the randomized LASSO at $\lambda_{\text{cv.1se}}$, and "Liu et al.," "Lee et al.," "Split" after the usual LASSO at $\lambda_{\text{cv.min}}$, Figure 7 depicts their relative risks and power under the respective best-case situations. For the moderately high SNR values, the selective MLE proves to be a competing estimate when compared against the LASSO estimate. We note a far superior predictive performance of the LASSO at $\lambda_{\text{cv.min}}$ than the selective MLE in the lower range of SNR values. Our proposal, however, turns out to be a better choice for inference across the range of SNR values, outperforming the non-randomized alternatives in terms of power.

5. Real Data Example

We apply our method to investigate associations between gene expressions and patient survival times for Gliomas (a type of brain tumor) in the TCGA data. With survival times ranging between 1 and 15 years, and some of these tumors quickly

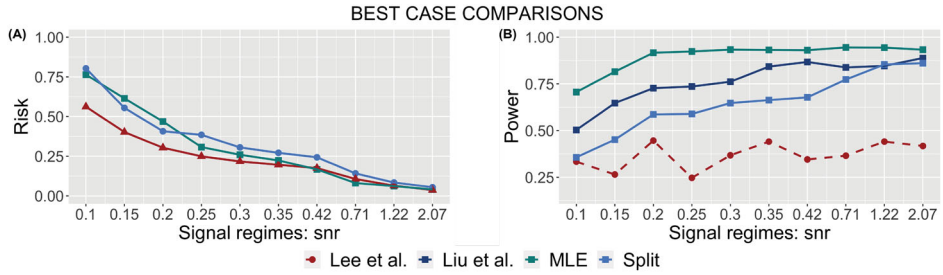

Figure 7. Best-case comparisons between randomized and nonrandomized estimates. (A) depicts the relative-risks for point estimates associated with each method; (B) depicts power of each method in detecting true associations after selection.

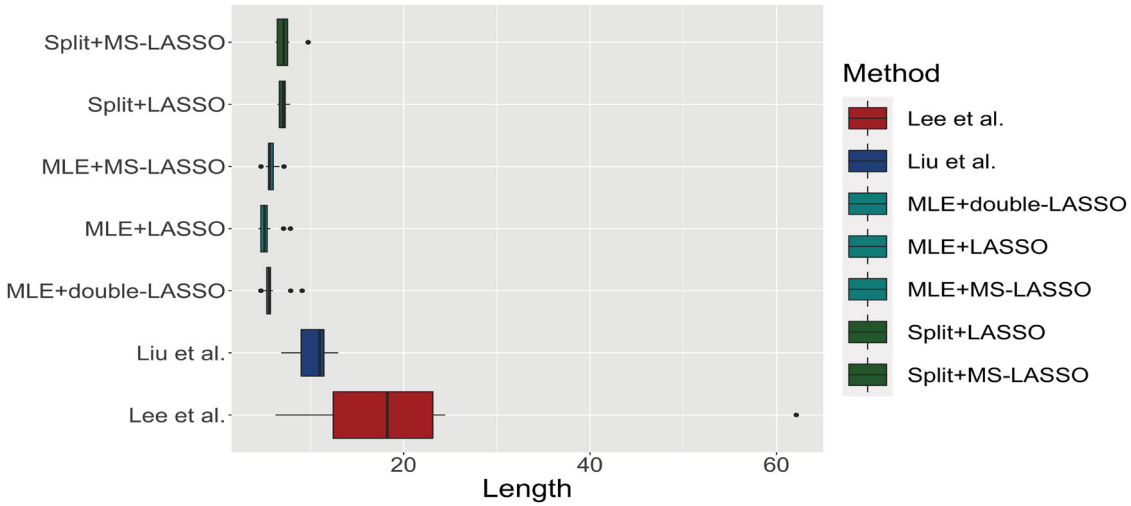

Figure 8. TCGA analysis. Boxplots for lengths of interval estimates by all methods.

progressing to Glioblastoma, genetic associations are increasingly used for prognostic decisions (e.g., Zhang et al. 2019; Panigrahi et al. 2020). In our analysis of 441 samples, we use log-transformed survival times as our response. As potential predictors, we choose the top 2500 predictors with the largest sample variation from a candidate pool of 17,500 molecular measurements of gene expression values (mRNAseq). Before running a meaningful LASSO query, we account for the presence of strongly correlated predictors by further pruning the 2500 predictors to a subset of 140 predictors. We do so by applying the hierarchical clustering scheme in Bien and Tibshirani (2011), followed by collecting the prototype representatives for each resulting cluster of predictors. We consider the following algorithms: (a) the LASSO; (b) two runs of the LASSO; (c) a marginal screening of the predictors followed by the LASSO, adding a Gaussian randomization variable $\omega \sim N(0, \eta^2 I_p)$ to the queries for our method. Consistent with the simulations, we fix $(\hat{\sigma})^{-2}\eta^2 = 1$. We use λ_{theory} in Section 4 to tune the LASSO penalty. We conduct a marginal screening of variables at the nominal level $q = 0.20$ and let the screening threshold be $\zeta = z_{1-q/2} \cdot \sqrt{\hat{\sigma}^2 \text{diag}(X^T X) + \eta^2 1_p}$ for the randomized version of this query.

Figure 8 showcases the distribution of the lengths of interval estimates produced by "MLE," "Lee et al." and "Liu et al." after solving (a). For inference post (b) and (c), we compare "MLE + double-LASSO" and "MLE + MS-LASSO" against "Split + LASSO" where half of the samples are reserved for inference. The conditional prescriptions in "Lee et al." and "Liu et al." do not directly apply to accommodate multiple queries at the time of selection. Corroborating our findings in the simulations, the lengths of the estimates using our proposal are way shorter than those based on "Liu et al." and "Lee et al." Observe, "MLE + double-LASSO" and "MLE + MS-LASSO" outperform "Split + LASSO" with shorter intervals, despite querying the data twice before inference.

6. Discussion

We investigate in the current article a method for selective inference via maximum likelihood estimation. Amenable to a large class of convex queries at the time of selection, we rely on an optimization problem whose solution yields us estimating equations for the MLE and the observed Fisher information matrix, the two main ingredients for the proposed method. The estimating equations easily generalize to multiple convex queries at the time of selection and assume a separable form across the queries. The appeal of our method is 2-fold: (a) the computing costs for selective inference are reduced by orders of magnitude in comparison with MCMC sampling-based alternatives, and (b) statistical power for inference is preserved despite querying the data multiple times through randomized queries.

Future extensions of our proposal include a development of theory to use the method beyond Gaussian data. Along

this direction, uniform guarantees for coverage (see, e.g., Leeb and Pötscher 2005, 2006) require closer investigation. The framework for selective inference in the paper adjusts for queries with affine K.K.T. conditions at the solution. The ability of the proposal to account for learning queries that present non-affine representations for the K.K.T. conditions, for example, the Group LASSO, remain to be explored in the future.

Supplementary Materials

The supplementary materials contain proofs for the technical results, provide additional examples to demonstrate the soft-truncated likelihood, show asymptotic guarantees for the approximate selective MLE, and illustrate the generalization of our method to multiple, convex queries.

Acknowledgments

S.P. would like to sincerely thank and acknowledge Veera Baladandayuthapani and Yujia Pan for their inputs in the analysis of the TCGA dataset. S.P. is immensely thankful to Xuming He and Liza Levina for offering valuable comments on an initial draft of the article. The authors thank the anonymous reviewers for their many insightful suggestions on earlier drafts of the article.

Funding

Snigdha Panigrahi acknowledges support by NSF-DMS 1951980 and NSF-DMS 2113342. Jonathan Taylor acknowledges support in part by ARO grant 70940MA.

References

Benjamini, Y., and Yekutieli, D. (2005), "False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters," *Journal of the American Statistical Association*, 100, 71–81. [1]

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. (2013), "Valid Post-Selection Inference," *The Annals of Statistics*, 41, 802–837. [1,5]

Bien, J., and Tibshirani, R. (2011), "Hierarchical Clustering with Prototypes via Minimax Linkage," *Journal of the American Statistical Association*, 106, 1075–1084. [10]

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015), "Slope Adaptive Variable Selection via Convex Optimization," *The Annals of Applied Statistics*, 9, 1103–1140. [7]

Cox, D. (1975), "A Note on Data-Splitting for the Evaluation of Significance Levels," *Biometrika*, 62, 441–444. [2]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1]

Fithian, W., Sun, D., and Taylor, J. (2014), "Optimal Inference After Model Selection," arXiv preprint arXiv:1410.2597. [2,3]

Guo, X., and He, X. (2020), "Inference on Selected Subgroups in Clinical Trials," *Journal of the American Statistical Association*, 116, 1–19. [2]

Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017), "Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso," arXiv preprint arXiv:1707.08692. [7]

Hurvich, C. M., and Tsai, C. (1990), "The Impact of Model Selection on Inference in Linear Regression," *The American Statistician*, 44, 214–217. [2]

Kivaranovic, D., and Leeb, H. (2018), "Expected Length of Post-Model-Selection Confidence Intervals Conditional on Polyhedral Constraints," arXiv preprint arXiv:1803.01665. [2,9]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-Selection Inference with the Lasso," *The Annals of Statistics*, 44, 907–927. [2,5]

Lee, J. D., and Taylor, J. E. (2014), "Exact Post Model Selection Inference for Marginal Screening," in *Advances in Neural Information Processing Systems*, pp. 136–144. [7]

Leeb, H., and Pötscher, B. M. (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59. [1,11]

——— (2006), "Can One Estimate the Conditional Distribution of post-Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554–2591. [1,11]

Liu, K., Markovic, J., and Tibshirani, R. (2018), "More Powerful Post-Selection Inference, with Application to the Lasso." arXiv preprint arXiv:1801.09037. [2]

McKeague, I. W., and Qian, M. (2015), "An Adaptive Resampling Test for Detecting the Presence of Significant Predictors," *Journal of the American Statistical Association*, 110, 1422–1433. [2]

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009), "A Unified Framework for High-Dimensional Analysis of *m*-Estimators with Decomposable Regularizers," in *Advances in Neural Information Processing Systems*, pp. 1348–1356. [7]

Panigrahi, S. (2019), "Carving Model-Free Inference," arXiv preprint arXiv:1811.03142. [2]

Panigrahi, S., Mohammed, S., Rao, A., and Baladandayuthapani, V. (2020), "Integrative Bayesian Models using post-selective Inference: A Case Study in Radiogenomics," arXiv preprint arXiv:2004.12012. [10]

Panigrahi, S., and Taylor, J. (2018), "Scalable Methods for Bayesian Selective Inference," *Electronic Journal of Statistics*, 12, 2355–2400. [6]

Panigrahi, S., Taylor, J., and Weinstein, A. (2021), Integrative Methods for Post-selection Inference under Convex Constraints," *The Annals of Statistics*, 49, 2803–2824. [2,7]

Panigrahi, S., Zhu, J., and Sabatti, C. (2019), Selection-Adjusted Inference: An Application to Confidence Intervals for cis-eQTL Effect Sizes," *Biostatistics*, 22, 181–197. [2]

Rügamer, D., and Greven, S. (2018), "Selective Inference after Likelihood- or Test-Based Model Selection in Linear Models," *Statistics & Probability Letters*, 140, 7–12. [3]

Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., and Takeuchi, I. (2017), "Selective Inference for Sparse High-Order Interaction Models," in *International Conference on Machine Learning*, pp. 3338–3347. PMLR. [3]

Tian, X., and Taylor, J. (2018), "Selective Inference with a Randomized Response," *The Annals of Statistics*, 46, 679–710. [1,2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Yang, F., Barber, R. F., Jain, P., and Lafferty, J. (2016), "Selective Inference for Group-Sparse Linear Models," in *Advances in Neural Information Processing Systems*, pp. 2469–2477. [3]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [1]

Zhang, Y., Morris, J. S., Aerry, S. N., Rao, A. U., Baladandayuthapani, V. (2019), "Radio-ibag: Radiomics-based Integrative Bayesian Analysis of Multiplatform Genomic Data," *The Annals of Applied Statistics*, 13, 1957–1988. [10]