The Annals of Statistics
2021, Vol. 49, No. 5, 2803–2824
https://doi.org/10.1214/21-AOS2057

INTEGRATIVE METHODS FOR POST-SELECTION INFERENCE UNDER CONVEX CONSTRAINTS

BY SNIGDHA PANIGRAHI[1], JONATHAN TAYLOR[2] AND ASAF WEINSTEIN[3]

[1]*Department of Statistics, University of Michigan, psnigdha@umich.edu*
[2]*Department of Statistics, Stanford University, jtaylo@stanford.edu*
[3]*Department of Statistics, Hebrew University of Jerusalem, asaf.weinstein@mail.huji.ac.il*

Inference after model selection has been an active research topic in the past few years, with numerous works offering different approaches to addressing the perils of the reuse of data. In particular, major progress has been made recently on large and useful classes of problems by harnessing general theory of hypothesis testing in exponential families, but these methods have their limitations. Perhaps most immediate is the gap between theory and practice: implementing the exact theoretical prescription in realistic situations—for example, when new data arrives and inference needs to be adjusted accordingly—may be a prohibitive task. In this paper, we propose a Bayesian framework for carrying out inference after variable selection. Our framework is very flexible in the sense that it naturally accommodates different models for the data instead of requiring a case-by-case treatment. This flexibility is achieved by considering the full selective likelihood function where, crucially, we propose a novel and nontrivial approximation to the exact but intractable expression. The advantages of our methods in practical data analysis are demonstrated in an application to HIV drug-resistance data.

1. Introduction. Any meaningful statistical problem consists of a model and a set of matching parameters for which decisions are required. In the classical, "textbook" paradigm, the model and this set of parameters are assumed to be chosen independently of the data subsequently used for inference (e.g., estimating the parameters). In practice, however, data analysts often examine some aspect of the data before deciding on a model and the target parameters. Of course, ignoring such form of adaptivity in choosing the model and/or the target parameters, may result in flawed inferential conclusions. Still, most would agree that instructing the analyst to avoid such exploration of the data altogether, is not only impractical, but also not recommended. This realization on the one hand, and a serious concern about a replication problem in science on the other hand, have elicited an effort in the statistical community to develop tools for *selective inference*. In general, such tools allow to take into account the fact that the same data used to select target parameters is used when providing inference, thereby attempting to restore validity of inference following selection.

A common approach is to *condition* on selection, in other words, base inference on the likelihood of the observed data when truncated to all realizations that would result in the analyst posing the same question. A conditional approach has been pursued in many existing works under a sequence model, where each observation corresponds to a single parameter and the parameters are not assumed to have any relationship with one another. For example, a stylized problem is estimating θ_i from the conditional distribution of $Y_i \sim \mathcal{N}(\theta_i, 1)$ given that $Y_i > c$, where c is constant or data dependent. These works typically adopt a frequentist approach, for example, Zöllner and Pritchard [21], Weinstein, Fithian and Benjamini [18], Reid, Taylor

Received October 2019; revised January 2021.
MSC2020 subject classifications. Primary 62F30; secondary 62F10, 62G15.
Key words and phrases. Adaptive data analysis, Bayesian inference, carving, conditional inference, convex constraints, selective inference.

and Tibshirani [12]; an exception is Yekutieli [19], that uses a truncated likelihood within a Bayesian framework. More recently, the practicability of the conditional approach has been extended to the realm of linear models and GLMs ([4, 5, 14, 15], among others). At the core of these new contributions is a so-called *polyhedral lemma* [5], asserting that inference for the projection of the mean vector of Y (in a fixed-X homoscedastic, Gaussian linear model with known σ) onto a one-dimensional subspace can be reduced to the problem of inference for a univariate truncated normal variable.

The polyhedral lemma of Lee et al. [5] was indeed a significant step forward, but it has its limitations. First, the selection rule is assumed to rely on exactly the same set of observations used for subsequent inference, failing to accommodate realistic situations where new samples are available for inference at a later stage. Second, for the polyhedral lemma to yield a uniformly most powerful unbiased (UMPU) test, we need to assume a fixed-X setting and a specific model for the data entailing that the expectation of the response is an arbitrary vector in $\mathbb{R}^n$. Lastly, the target of inference in Lee et al. [5] is restricted to a real-valued linear function of the model parameters. If these assumptions are violated, the polyhedral lemma may not be applicable or it may lead to suboptimal procedures.

1.1. *Our contributions*. In the current paper, we address the problem of selective inference in regression when one moves away from the aforementioned assumptions. To circumvent the limitations related to the polyhedral lemma, instead of working with univariate projections we propose to work directly with a full selective likelihood function. We make this operational by developing a *tractable approximation* to the exact truncated likelihood function, the latter admitting no closed form and presenting serious computational objections. This approximation is the crux of our proposed methodology, and much of our effort will be focused on developing the approximation and proving that it enjoys desirable statistical properties.

Our approach allows considerably more flexibility as compared to that of Lee et al. [5]. In our setup, we divide the data into two parts, where selection operates only on the first part but inference can use both parts. This is often encountered naturally when fresh data arrives after the model selection stage. While valid inference can be provided using only the held-out data and without any adjustment, Fithian, Sun and Taylor [4] show that *data carving*—incorporating also the "leftover" information from the first part—generally improves statistical accuracy. By casting our problem in the *randomized response* framework of Tian and Taylor [14], our methods easily handle data carving.

Another difference between our work and, for example, Lee et al. [5], is that we allow the researcher to refine the choice of the model after seeing the output of the automatic algorithm. Indeed, this kind of flexibility is essential in several practical scenarios. Suppose that we run Lasso to select among p main effects, but after observing the output, we want to add interactions between some of the selected variables; including the interactions may account for some unexplained variance [20]. In other cases, domain-specific information might be available a priori for the predictors. For instance, in statistical genomics, it is common to use existing biological annotations to inform models. Such information may suggest to include a group of genes known to interact with each other, even if some of them were not selected by the automatic protocol.

Finally, while the flavor is frequentist as far as using a truncated likelihood, we follow the ideas of Yekutieli [19] and incorporate a prior on the parameters of the adaptively chosen model. This allows us to give inference for general functions of the parameter vector by integrating out nuisance parameters. By adopting formally a Bayesian approach, we are able to exploit the usual advantages of the Bayesian machinery. For instance, credible intervals can be estimated easily by finding appropriate quantiles in the posterior sample of the parameter.

2. Preliminaries. This section is divided into three subsections: the first sets up the mathematical framework, the second summarizes the main results in the paper and the third gives an outline of the remainder of the paper.

2.1. *Formal framework.* We introduce a general framework for inference after model selection that, as usual, includes two steps. The first step entails selecting variables with a procedure acting on the first split of the data. In the second stage, we assume a linear model relating the response to the selected variables, incorporate a prior on the corresponding coefficient vector and work toward a selection-adjusted posterior. For any matrix $\boldsymbol{D} \in \mathbb{R}^{k\times l}$, any subset $N \subseteq \{1,\dots,k\}$ and any subset $M \subseteq \{1,\dots,l\}$, we denote by $\boldsymbol{D}^N$ the $|N| \times l$ matrix obtained from $\boldsymbol{D}$ by keeping only the *rows* with indices in N, and we denote by $\boldsymbol{D}_M$ the $k \times |M|$ matrix obtained from $\boldsymbol{D}$ by keeping only the *columns* with indices in M. Also, $\mathcal{N}_k(\cdot;\boldsymbol{\eta},\boldsymbol{\Gamma})$ is the density of an k-dimensional multivariate normal vector with mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{\Gamma}$.

Let

$$(X_{i1},\dots,X_{ip},Y_i) \sim P, \quad i = 1,\dots,n,$$

be $(p+1)$-dimensional i.i.d. vectors with $Y_i \in \mathbb{R}$. Denote $\mathrm{y} = (Y_1,\dots,Y_n)^T$, and denote by $\mathbf{X}$ the $n \times p$ matrix with $(X_{i1},\dots,X_{ip})$ in its ith row. For a subset $\mathcal{S} \subset \{1,\dots,n\}$ determined independently of the data, let $(\mathbf{X}^{\mathcal{S}}, \mathrm{y}^{\mathcal{S}})$ be a subsample of size $|\mathcal{S}| = n_1$. We will refer to $(\mathbf{X}^{\mathcal{S}}, \mathrm{y}^{\mathcal{S}})$ as the original data henceforth. In the selection step, the statistician has access only to the original data, and she starts with applying some predetermined variable selection rule, a mapping

$$(\mathbf{X}^{\mathcal{S}}, \mathrm{y}^{\mathcal{S}}) \longmapsto \widehat{E} \subseteq \{1,\dots,p\}.$$

The statistician is allowed at this point to refine the selection by dropping variables from (or adding variables to) the output $\widehat{E}$. Specifically, let $\widehat{E}' \subseteq \{1,\dots,p\}$ be a subset obtained by applying an arbitrary but deterministic function to $\widehat{E}$. We denote by E and E' the realizations of the variables $\widehat{E}$ and $\widehat{E}'$, respectively.

In the second stage, inference is provided for P assuming that

$$\mathrm{y}|\mathbf{X} \sim \mathcal{N}_n\bigl(y; \mathbf{X}_{E'}\beta^{E'}, \sigma^2\mathbf{I}\bigr). \tag{1}$$

If not indicated otherwise we treat $\sigma^2 = \sigma^2_{E'}$ as known, although in Sections 5 and 8 we explain how our methods easily accommodate the case of unknown σ^2. Naive inference would now proceed under the model (1), ignoring the fact that E' was chosen after seeing (part of) the data. Instead, the fact that E' is data-dependent will be accounted for by *truncating* the likelihood in (1) to the event $\{\widehat{E} = E\}$. Formally, this means that, conditionally on $\mathbf{X}$, (1) should be replaced by

$$\mathrm{y}|\mathbf{X}, \widehat{E} = E \sim \frac{\mathcal{N}_n(y; \mathbf{X}_{E'}\beta^{E'}, \sigma^2\mathbf{I})}{\int \mathbb{1}_{\{\widehat{E}=E\}}(\mathrm{y}) \cdot \mathcal{N}_n(\mathrm{y}; \mathbf{X}_{E'}\beta^{E'}, \sigma^2\mathbf{I})\,d\mathrm{y}} \mathbb{1}_{\{\widehat{E}=E\}}(y), \tag{2}$$

remembering that $\widehat{E}'$ is determined by $\widehat{E}$, and $\widehat{E}$ is in turn measurable with respect to $(\mathbf{X}^{\mathcal{S}}, \mathrm{y}^{\mathcal{S}})$.

In principle, we could now base post-selection inference for $\beta^{E'}$ on (2). Unfortunately, in practice the denominator in (2) is intractable, and we will seek simplifications. Note that if no selection were involved in specifying E', standard inference for $\beta^{E'}$ would rely on the distribution of the least squares estimator,

$$\widehat{\beta}^{E'} = \mathbf{X}^{\dagger}_{E'}\mathrm{y},$$

where $\mathbf{X}_{E'}^{\dagger}$ is the pseudo-inverse of $\mathbf{X}_{E'}$; if $\mathbf{X}_{E'}$ has full rank, this is $(\mathbf{X}_{E'}^T\mathbf{X}_{E'})^{-1}\mathbf{X}_{E'}^T$. If it were possible to express the selection event in terms of $\widehat{\beta}^{E'}$ and nothing else, this would give rise to a truncated distribution for $\widehat{\beta}^{E'}$. This is, however, not the case for the selection rules considered in the current paper for two reasons. First of all, because we implement data carving, the choice of E cannot be a deterministic function of $\widehat{\beta}^{E'}$ since the latter is a function of the full data $(\mathbf{X}, \mathrm{y})$ rather than $(\mathbf{X}^{\mathcal{S}}, \mathrm{y}^{\mathcal{S}})$. Second, even if we allowed selection to be based on the full data $(\mathbf{X}, \mathrm{y})$, we cannot restrict it to be a function of $\widehat{\beta}^{E'}$ alone. For example, Lasso selection cannot be expressed in terms of only the least squares statistic even if selection operates on the entire data set.

Suppose for the moment then that the event $\widehat{E} = E$ could be written in a polyhedral form,

$$\mathbf{A}_E\sqrt{n}\mathrm{T}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E, \tag{3}$$

with $\mathrm{T}_n = (\widehat{\beta}^{E'}, \mathrm{U})^T$ and where U is a vector that may itself depend on E'. The variable Ω_n has the property that (i) it is independent of T_n and (ii) it follows a distribution that does not depend on $\beta^{E'}$. Using the terminology of Tian and Taylor [14], Ω_n is a *randomization* term. As we will see later on, the selection rules that we will be concerned with can indeed be approximately represented in the form (3) with appropriate choices of U, Ω_n and $\mathbf{A}_E$, $\mathbf{B}_E$, b_E. By "approximately represented," we mean that we allow a term that goes to zero in probability in (3) and also that Ω_n satisfies the properties (i) and (ii) above asymptotically, as in the sense formalized in Section 4.

Treating (3) first as if it holds exactly instead of asymptotically, we start by considering the truncated distribution of (T_n, Ω_n),

$$\frac{L^n(\beta^{E'}; t_n, \omega_n)}{\mathbb{P}(\mathbf{A}_E\sqrt{n}\mathrm{T}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E|\beta^{E'})}\mathbb{1}_{\{\mathbf{A}_E\sqrt{n}\mathrm{t}_n + \mathbf{B}_E\omega_n < \mathrm{b}_E\}}(t_n, \omega_n),$$

where $L^n(\beta^{E'}; t_n, \omega_n)$ denotes the marginal likelihood corresponding to (T_n, Ω_n). The term

$$\mathbb{P}\big(\mathbf{A}_E\sqrt{n}\mathrm{T}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E|\beta^{E'}\big), \tag{4}$$

appearing in the denominator of the previous expression, will occur frequently throughout the manuscript and we call it the *adjustment factor*. We emphasize that, when the adjustment factor (4) appears throughout the manuscript, it should be formally understood as equivalent to the probability of the *exact* selection event. As the selection event will have a polyhedral representation only approximately, this is a slight abuse of notation, but we chose to use it anyway to emphasize the form of the selection regions under consideration. Also, to avoid ambiguity, we emphasize $\mathbb{P}(\cdot)$ denotes probability taken under fixed $\beta^{E'}$ but marginalizing over all of the other random variables, that is,

$$\mathbb{P}(A) = \int \mathbb{1}_A(t_n, \omega_n) \cdot L^n\big(\beta^{E'}; t_n, \omega_n\big)\, dt_n\, d\omega_n$$

for any event A measurable with respect to (T_n, Ω_n). As in (4), the dependence on $\beta^{E'}$ is sometimes stressed by writing $\mathbb{P}(\cdot|\beta^{E'})$ instead of $\mathbb{P}(\cdot)$.

By integrating with respect to ω_n, and using the fact that Ω_n is a randomization term, we obtain our *selection-adjusted likelihood*,

$$\begin{aligned} L_S^n(\beta^{E'}; t_n) &:= \frac{L^n(\beta^{E'}; t_n)}{\mathbb{P}(\mathbf{A}_E\sqrt{n}\mathrm{T}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E|\beta^{E'})}\mathbb{P}\big(\mathbf{A}_E\sqrt{n}\mathrm{t}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E|\mathrm{t}_n\big) \\ &\propto \frac{L^n(\beta^{E'}; t_n)}{\mathbb{P}(\mathbf{A}_E\sqrt{n}\mathrm{T}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E|\beta^{E'})}. \end{aligned} \tag{5}$$

Observe that the probability in the numerator of the second term above is taken over Ω_n only, hence this term does not depend on $\beta^{E'}$.

The point of view has thus far been frequentist. Inspired by the basic ideas in Yekutieli [19], we now introduce a prior into our model,

$$\beta^{E'} | \widehat{E} = E \sim \pi. \tag{6}$$

The conditioning on $\widehat{E} = E$ above is consistent with the same conditioning in the likelihood: since we pose a generative model for the data conditionally on selection, the prior as well is a conditional prior. Combining the selection-adjusted likelihood (5) and the prior (6) results formally in the *selection-adjusted posterior* distribution,

$$\pi_S(\beta^{E'} | \mathrm{t}_n) \propto \pi(\beta^{E'}) L_S^n(\beta^{E'}; t_n). \tag{7}$$

We use again a subscript S in π_S to indicate that in the selection-adjusted posterior, π is updated with the selection-adjusted likelihood rather than a marginal likelihood. The idea is to provide adjusted inference for $\beta^{E'}$, or functions thereof, based on (7).

2.2. *Overview of our main results.* Complementing Section 1.1, here we provide a more detailed and slightly more technical account of the main results. The central component of our Bayesian methods is the selection-adjusted likelihood (5), which also presents the primary technical challenge: in order to sample from (7) we need to be able to evaluate the adjustment factor (4), which as a function of $\beta^{E'}$, is generally intractable. Hence, most of our effort will be devoted to finding an amenable approximation as a substitute for (4). In Section 3 we demonstrate our methods in a very special, univariate case where the exact term can actually be computed and, therefore, serve as a reference. Afterwards, we extend the theory to the general case that covers, for example, Lasso selection in a regression problem. Before delving into the details, we present the scope of our theoretical results below while drawing the analogy between the univariate case of Section 3 and the general case of Sections 4 onward:

1. *Representation of the selection region*: we find an explicit polyhedral representation of the selection rule in the form (3) under data carving and a flexible choice for the model as specified by (1). For the univariate case, this (simple) representation is given by equation (10). For selection with Lasso, the (considerably more involved) counterpart is given in Propositions 4.1 and 4.2, and for the two additional examples of the supplement [9] the counterparts are included within.
2. *Consistency of the approximation to the adjustment factor*: as already pointed out, the exact probability (4) is generally too complicated to be handled directly. In Theorem 3.1, we develop an approximation for this probability in the special univariate case and prove that it converges asymptotically to the exact term. This is extended to arbitrary polyhedral selection rules in Theorem 5.1. Notably, our moderate deviations-type approximation in the general case still has the correct limit.
3. *Consistency of the resulting approximate posterior*: we establish frequentist credibility for our Bayesian methods by proving that the resulting (approximate) posterior, obtained by substituting the approximation to the adjustment factor, concentrates around the model parameters as data accumulates ($n \to \infty$). This is formalized for the univariate example in Theorem 6.2, which we extend later to Theorem 6.5 for the general case. To this end,

   (a) *Strong concavity of our approximate selection-adjusted log-likelihood* is established in Proposition 6.1 for the univariate case and in Proposition 6.2 for the general case.

(b) *Consistency of our selection-adjusted MLE* is asserted in Theorem 6.1 for the univariate case and in Theorem 6.4 for the general case.

4. *Convexity of the MAP program*: emphasizing the practicability of the proposed approximation, we represent the MAP estimator as the solution of an amenable convex optimization problem for any log-concave prior. We offer Theorem 6.3 for the univariate example and Theorem 6.6 for the general case.

2.3. *Organization of the paper.* The rest of the paper is organized as follows. In the next two sections, we instantiate the general framework by considering two particular cases: Section 3 begins with a simple univariate example to develop some intuition. In Section 4, we proceed to the main case study, Lasso selection in the linear model, that features all of the complexity of the general problem; further examples appear in the supplement. Section 5 develops an approximation to the selection-adjusted likelihood in the general problem, with the focus on a computationally amenable proxy. We address point estimation based on the approximate posterior in Section 6, with a separate analysis for the univariate case and the general case. In Section 7, we report findings from numerical experiments, analyzing simulated data as well as real data consisting of records for HIV drug resistance. Computational details for implementing our methods are provided in Section 8.

3. A special univariate case. We begin with studying an elementary example where there are no covariates and the selection rule is very simple. Besides priming the sequel, the example below is a case where an exact analysis can be carried out and serve as a true benchmark.

Throughout this section, then suppose that the observations are

$$Y_i \sim \mathcal{N}(\beta, 1), \quad i = 1, \dots, n. \tag{8}$$

For the theoretical analysis that follows, $\beta = \beta_n$ depends in general on n, and we use the parameterization $\sqrt{n}\beta_n \equiv n^{\delta}\beta^*$ with β^* a constant and $\delta \in (0, 1/2]$. Let $\mathcal{S} \subseteq \{1, \dots, n\}$ be a random subset of size $n_1 = \rho n$, and denote

$$\bar{Y}^{\mathcal{S}} := \frac{1}{n_1}\sum_{i\in\mathcal{S}} Y_i, \qquad \bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

We provide inference for β_n conditionally on

$$\sqrt{n_1}\bar{Y}^{\mathcal{S}} > 0, \tag{9}$$

that is, the selection event entails the (scaled) least squares estimate for β_n from the original data exceeds a fixed threshold (zero). We first note that, on defining $W_n := \bar{Y}^{\mathcal{S}} - \bar{Y}_n$, we have

$$\sqrt{n}W_n \sim \mathcal{N}\Big(0, \frac{1-\rho}{\rho}\Big), \qquad \sqrt{n}W_n \perp\!\!\!\perp \sqrt{n}\bar{Y}_n.$$

Therefore, letting

$$\sqrt{n}\widehat{\beta}_n = \sqrt{n}\bar{Y}_n, \qquad \Omega_n = \sqrt{n}W_n, \tag{10}$$

we see that the selection event (9) can be written exactly in the form (3) with $T_n = \widehat{\beta}_n$, and with $\mathbf{A}_E = -1$, $\mathbf{B}_E = -1$, $b_E = 0$.

In this simple example, unlike in the general problem, we have an explicit form for the exact probability of the selection event,

$$\mathbb{P}\big(\mathbf{A}_E\sqrt{n}\mathrm{T}_n + \mathbf{B}_E\Omega_n < \mathrm{b}_E|\beta_n\big) = \mathbb{P}\big(\sqrt{n_1}\bar{Y}^{\mathcal{S}} > 0|\beta_n\big) = \bar{\Phi}(-\sqrt{\rho}\cdot\sqrt{n}\beta_n). \tag{11}$$

The following result suggests an approximation for the selection probability (11) in the simple univariate example, which we will generalize in the next section.

THEOREM 3.1. *Let $\beta = \beta_n = n^{\delta}\beta^*/\sqrt{n}$, where β^* is a constant. Let $\mathcal{K} = (0, \infty)$. Then*

$$\log \mathbb{P}\big(\sqrt{n}(\bar{Y}_n + W_n)/n^{\delta} \in \mathcal{K} | \beta_n\big) \leq -n^{2\delta} \cdot \inf_{(z,w): z+w \in \mathcal{K}} \frac{(z-\beta^*)^2}{2} + \frac{\rho w^2}{2(1-\rho)} \tag{12}$$

for any $n \in \mathbb{N}$ as long as $0 < \delta \leq 1/2$. Furthermore, the corresponding sequence of the logarithms of the selection probabilities satisfies

$$\lim_{n\to\infty} \frac{1}{n^{2\delta}} \log \mathbb{P}\big(\sqrt{n}(\bar{Y}_n + W_n)/n^{\delta} \in \mathcal{K} | \beta_n\big) = - \inf_{(z,w): (z+w) \in \mathcal{K}} \frac{(z-\beta^*)^2}{2} + \frac{\rho w^2}{2(1-\rho)}.$$

We call the exponent of the right-hand side of (12) the *Chernoff approximation* to the selection probability (11). Note that in the Chernoff approximation the infimum is taken over a constrained set in $\mathbb{R}^2$. To obtain a more computationally amenable expression, we propose to replace the right-hand side of (12) with the unconstrained optimization problem

$$-n^{2\delta} \cdot \inf_{(z,w)\in\mathbb{R}^2} \left\{ \frac{(z-\beta^*)^2}{2} + \frac{\rho w^2}{2(1-\rho)} + \frac{1}{n^{2\delta}} \psi_{n^{-\delta}}(z+w) \right\}, \tag{13}$$

where $\psi_{n^{-\delta}}$ is a function satisfying that in the limit as $n \to \infty$,

$$\frac{1}{n^{2\delta}} \psi_{n^{-\delta}}(x) \longrightarrow I(x) = \begin{cases} 0 & \text{if } x \in (0, \infty), \\ \infty & \text{otherwise.} \end{cases}$$

Specifically, using

$$\psi_{n^{-\delta}}(x) = \log\big(1 + n^{-\delta}/x\big) \tag{14}$$

whenever $x \in (0, \infty)$ and ∞ otherwise in (13), we obtain our *barrier approximation* to the selection probability (11), so-called because (14) serves as a "soft" barrier alternative to the indicator function $I(x)$. Figure 1(a) compares the approximations in (12) and (13) to the exact expression (11). The plots for both approximations follow the exact curve fairly closely. Notice, the curve depicting the Chernoff approximation lies above that for the true probability, as predicted by Theorem 3.1.

REMARK 3.2. In practice, we will not have access to δ but, by substituting $\sqrt{n}z' = n^{\delta}z$, $\sqrt{n}w' = n^{\delta}w$, (13) is equivalent to

$$-n \cdot \inf_{(z',w')\in\mathbb{R}^2} \big\{ (z'-\beta_n)^2/2 + \rho w'^2/2(1-\rho) + \psi_{n^{-1/2}}(z'+w')/n \big\},$$

and we can simply work with β_n without knowing the underlying parameterization.

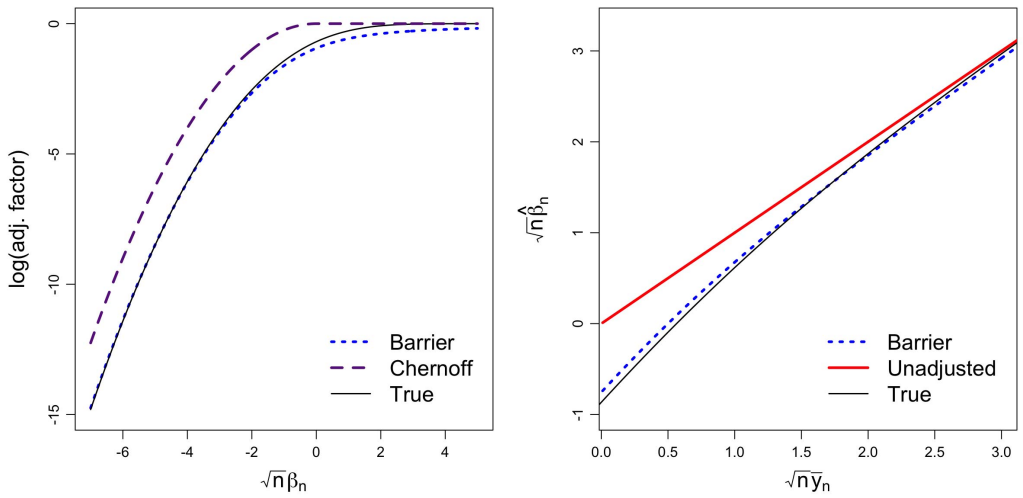
Next, we show that when substituting the barrier approximation for the exact selection probability, the resulting adjusted posterior converges to the actual (exact) selection-adjusted posterior. First, note that the logarithm of the selection-adjusted posterior is

$$\log \pi_S(\beta_n | \bar{y}_n) = \log \pi(\beta_n) - n(\bar{y}_n - \beta_n)^2/2 - \log \mathbb{P}\big(\sqrt{n}(\bar{Y}_n + W_n)/n^{\delta} \in \mathcal{K} | \beta_n\big).$$

On the other hand, corresponding to any barrier function $\psi_{n^{-\delta}}$ is the *approximate* adjusted posterior

$$\begin{aligned} \log \tilde{\pi}_S(\beta_n | \bar{y}_n) &= \log \pi(\beta_n) - n(\bar{y}_n - \beta_n)^2/2 \\ &\quad + n^{2\delta} \cdot \inf_{(z,w)\in\mathbb{R}^2} \left\{ \frac{(z-\beta^*)^2}{2} + \frac{\rho w^2}{2(1-\rho)} + \frac{1}{n^{2\delta}} \psi_{n^{-\delta}}(z+w) \right\}. \end{aligned} \tag{15}$$

Then a consequence of Theorem 3.1 is the followin corollary.



(a) (b)

FIG. 1. (*a*) *Approximations to log-adjustment factor in the univariate case of Section 3. In the legend, "Chernoff" refers to the approximation from* (12), *and "Barrier" to the approximation from* (13); *"True" corresponds to the exact expression* (11). (*b*) *Maximum-likelihood estimation for the univariate Gaussian setting. Broken line corresponds to the approximate MLE, incorporating the proposed approximation to the selection probability. Solid black curve is the exact* (*true*) *MLE, and solid red line is the identity line, representing the unadjusted estimate.*

COROLLARY 3.3. *Let* $\beta = \beta_n$ *be a sequence such that* $\sqrt{n}\beta_n \equiv n^\delta \beta^*$, $\delta \in (0, 1/2]$. *Furthermore, suppose that in* (15) *the function* $\psi_{n^{-\delta}}$ *satisfies that for each* x,

$$n^{-2\delta}\psi_{n^{-\delta}}(x) \longrightarrow I_{\mathcal{K}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{K}, \\ \infty & \text{otherwise} \end{cases}$$

as $n \to \infty$. *Then*

$$\lim_{n\to\infty} \frac{1}{n^{2\delta}} \{\log \pi_S(\beta_n|\bar{y}_n) - \log \widetilde{\pi}_S(\beta_n|\bar{y}_n)\} = 0.$$

4. Selection with the lasso. We now turn to the case of primary interest, specializing the general framework of Section 2 to selection with the Lasso in the linear model. It is important to remark again that our methodology applies to many other selection protocols. For example, in the supplement we show how marginal-screening and logistic Lasso for a binary response, can be formulated to fit our framework.

From now on, unless specified otherwise, we consider

$$\widehat{E} = \{j : \widehat{\beta}^\lambda_j \neq 0\}, \tag{16}$$

where $\widehat{\beta}^\lambda$ is the solution to

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \frac{1}{2\sqrt{n}\rho} \|\mathrm{y}^{\mathcal{S}} - \mathbf{X}^{\mathcal{S}}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{17}$$

and $\rho = n_1/n$. We emphasize that $\widehat{\beta}^\lambda = \widehat{\beta}^\lambda_{\mathcal{S}}$ is obtained from only the original data $(\mathrm{y}^{\mathcal{S}}, \mathbf{X}^{\mathcal{S}})$, but we suppress the subscript throughout. We also denote by $\widehat{\beta}^\lambda_E$ the vector in $\mathbb{R}^{|E|}$ consisting of only the nonzero coordinates of $\widehat{\beta}^\lambda$. While the general prescription in Section 2 would now entail conditioning on the event $\widehat{E} = E$, as in Lee et al. [5] we in fact condition on the more specific event

$$(\widehat{E}, \widehat{S}^{\widehat{E}}) = (E, s^E), \tag{18}$$

where $\widehat{S}^{\widehat{E}} = \operatorname{sgn}(\widehat{\beta}^{\lambda}_{\widehat{E}})$ is the vector of signs of $\widehat{\beta}^{\lambda}_{\widehat{E}}$. This refinement is important for obtaining a convex truncation region, which is crucial for our methods to be applicable. As in the general setting, the object of inference is $\beta^{E'}$, with $E' \subseteq \{1, \ldots, p\}$ chosen upon observing (18).

To fit the general framework of Section 2, in the previous section we reexpressed the selection event, originally involving $\widehat{\beta}^{\mathcal{S}} = \bar{Y}^{\mathcal{S}}$, as a sum involving $\widehat{\beta}_n = \bar{Y}_n$ and an independent variable W_n. Furthermore, the distribution of W_n was fully specified, in particular it did not depend on the unknown parameter β. We will now develop an analogous asymptotic representation for the Lasso selection event (18). Thus, let

$$\sqrt{n}\mathrm{T}_n := \begin{pmatrix} \sqrt{n}\widehat{\beta}^{E'} \\ \sqrt{n}\mathrm{N}_{-E'} \end{pmatrix} := \begin{pmatrix} \sqrt{n}\widehat{\beta}^{E'} \\ \frac{1}{\sqrt{n}}\mathbf{X}^T_{-E'}(\mathrm{y} - \mathbf{X}_{E'}\widehat{\beta}^{E'}) \end{pmatrix}, \tag{19}$$

where $\mathbf{X}_{-E'}$ is the matrix obtained from $\mathbf{X}$ by deleting the columns with indices in E'. Define also

$$\Omega_n = \sqrt{n}W_n = \frac{\partial}{\partial\beta}\left\{-\frac{1}{2\sqrt{n}\rho}\|\mathrm{y}^{\mathcal{S}} - \mathbf{X}^{\mathcal{S}}\beta\|_2^2 + \frac{1}{2\sqrt{n}}\|\mathrm{y} - \mathbf{X}\beta\|_2^2\right\}\Big|_{\beta=\widehat{\beta}^{\lambda}}. \tag{20}$$

As observed in [6], Ω_n can be treated asymptotically as a randomization term. We verify this in Proposition 4.1 and show that the event (18) can be asymptotically written as in (3) in Proposition 4.2.

The next two propositions formulate the Lasso selection in terms of our framework.

THEOREM 4.1. *Denote $\mu := (\beta^{E'}, 0)^T \in \mathbb{R}^p$ where $\beta^{E'} \in \mathbb{R}^{|E'|}$. Assume $\mathbb{E}_P(\mathbf{X}^T_{E'}\mathbf{X}_{E'}/n)$ is invertible. Define also the matrices*

$$\mathbf{Q} := \mathbb{E}_P(\mathbf{X}^T_{E'}\mathbf{X}_{E'}/n),$$

$$\mathbf{N} := \{\mathbb{E}_P(\mathbf{X}^T_{-E'}\mathbf{X}_{-E'}/n) - \mathbb{E}_P(\mathbf{X}^T_{-E'}\mathbf{X}_{E'}/n)\mathbf{Q}^{-1}\mathbb{E}_P(\mathbf{X}^T_{-E'}\mathbf{X}_{E'}/n)^T\}^{-1}$$

and

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_P & 0 \\ 0 & \mathbf{\Sigma}_{\mathcal{G}} \end{bmatrix}, \qquad \mathbf{\Sigma}_P = \sigma^2 \begin{bmatrix} \mathbf{Q}^{-1} & 0 \\ 0 & \mathbf{N}^{-1} \end{bmatrix}.$$

Then, under the modeling assumption (1),

$$\begin{pmatrix} \sqrt{n}(T_n - \mu) \\ \Omega_n \end{pmatrix} \dot{\sim} \mathcal{N}_{2p}(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}_{\mathcal{G}}$ is a $p \times p$ matrix, and where the symbol "$\dot{\sim}$" means "asymptotically distributed as."

Before we prove Proposition 4.2, we let $\mathcal{I}^{E,E'} \in \mathbb{R}^{|E|\times(p-|E'|)}$ and $\mathcal{J}^{E,E'} \in \mathbb{R}^{(p-|E|)\times(p-|E'|)}$ be matrices with entries that are all zero except for the (j, j) entries, which are defined as follows: if $E = \{k_1, k_2, \ldots, k_{|E|}\}$ and $E^c = \{l_1, l_2, \ldots, l_{p-|E|}\}$, then

$$\mathcal{I}^{E,E'}_{j,j} = \begin{cases} 1 & \text{if } k_j \notin E', \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \ldots, \min(|E|, p - |E'|);$$

$$\mathcal{J}^{E,E'}_{j,j} = \begin{cases} 1 & \text{if } l_j \notin E', \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \ldots, \min(p - |E|, p - |E'|).$$

Define the following moments: $\mathbf{P}^{E,E'} = \mathbb{E}_P(\mathbf{X}^T_E\mathbf{X}_{E'}/n)$, $\mathbf{F}^{E,E'} = \mathbb{E}_P(\mathbf{X}^T_{-E}\mathbf{X}_{E'}/n)$, $\mathbf{Q}^E = \mathbb{E}_P((\mathbf{X}^T_E\mathbf{X}_E/n)$, $\mathbf{C}^E = \mathbb{E}_P(\mathbf{X}^T_{-E}\mathbf{X}_E/n)$.

THEOREM 4.2. *The event* $(\widehat{E}, \widehat{S}^E) = (E, s^E)$ *is equivalent to*

$$\mathbf{A}_E \sqrt{n} T_n + \mathbf{B}_E \Omega_n + o_p(1) < b_E, \tag{21}$$

where

$$\mathbf{A}_E = \begin{bmatrix} -\operatorname{diag}(s^E)(\mathbf{Q}^E)^{-1}\mathbf{P}^{E,E'} & -\operatorname{diag}(s^E)(\mathbf{Q}^E)^{-1}\mathcal{I}^{E,E'} \\ \mathbf{F}^{E,E'} - \mathbf{C}^E(\mathbf{Q}^E)^{-1}\mathbf{P}^{E,E'} & \mathcal{J}^{E,E'} - \mathbf{C}^E(\mathbf{Q}^E)^{-1}\mathcal{I}^{E,E'} \\ -\mathbf{F}^{E,E'} + \mathbf{C}^E(\mathbf{Q}^E)^{-1}\mathbf{P}^{E,E'} & -\mathcal{J}^{E,E'} + \mathbf{C}^E(\mathbf{Q}^E)^{-1}\mathcal{I}^{E,E'} \end{bmatrix},$$

$$\mathbf{B}_E = \begin{bmatrix} -\operatorname{diag}(s^E)(\mathbf{Q}^E)^{-1} & \mathbf{0} \\ -\mathbf{C}^E(\mathbf{Q}^E)^{-1} & \mathbf{I} \\ \mathbf{C}^E(\mathbf{Q}^E)^{-1} & -\mathbf{I} \end{bmatrix}, \qquad b_E = \lambda \begin{pmatrix} -\operatorname{diag}(s^E)(\mathbf{Q}^E)^{-1} s^E \\ \mathbf{1} - \mathbf{C}^E(\mathbf{Q}^E)^{-1} s^E \\ \mathbf{1} + \mathbf{C}^E(\mathbf{Q}^E)^{-1} s^E \end{pmatrix}.$$

5. Approximations to the selection-adjusted posterior. Extending the results of Section 3, we will now present a computationally tractable approximation for the adjustment factor (4) in the general problem. The central result of this section, Theorem 5.1 below, is a moderate deviations-type of result implying a relatively simple approximation to the adjustment factor that has the correct (exact) limit.

5.1. *An approximation to the adjustment factor.* To make explicit the dependence of the parameter vector in (2) on n, we use the notation $\beta_n^{E'}$ instead of $\beta^{E'}$ and represent μ (defined in Proposition 4.1) by μ_n. Furthermore, without loss of generality we take the (known) variance term to be $\sigma^2 = 1$; the case of unknown σ^2 is discussed after stating the theorem.

We first recall the representation from the proof of Proposition 4.1

$$\begin{pmatrix} \sqrt{n} T_n \\ \Omega_n \end{pmatrix} - \sqrt{n} \begin{pmatrix} \mu_n \\ 0 \end{pmatrix} = \sqrt{n} \bar{Z}_n + E_n,$$

where $E_n = o_p(1)$ and $\mu_n = (\beta_n^{E'}, 0)^T \in \mathbb{R}^p$, and $\bar{Z}_n$ is the mean of n i.i.d. centered observations. Suppose that (1) holds for some vector $\beta_n^{E'}$ satisfying $\sqrt{n}\beta_n^{E'} = n^\delta \beta^* \in \mathbb{R}^{|E'|}$, where β^* does not depend on n, and where $\delta \in (0, 1/2)$. We will need the following assumptions.

ASSUMPTION 1. Defining W_n through $\Omega_n := \sqrt{n} W_n$, assume that

$$\mathbb{E}_P\big[\exp(\lambda_0 \| X(Y - X_{E'}{}^T \beta_n^{E'}) \|)\big], \mathbb{E}_P\big[\exp(\eta_0 \| X(Y - X_E{}^T \mathbb{E}[\widehat{\beta}_E^\lambda]) \|)\big] < \infty$$

for some $\lambda_0, \eta_0 > 0$, where $(X, Y) = (X_1, \ldots, X_p, Y) \sim P$.

ASSUMPTION 2. Additionally, we assume that

$$\lim_{n\to\infty} n^{-2\delta} \cdot \log \mathbb{P}\big[n^{-\delta} \| E_n \| > \epsilon\big] = -\infty \quad \text{for every } \epsilon > 0,$$

and that

$$\lim_{n\to\infty} \frac{1}{n^{2\delta}} \big\{ \log \mathbb{P}(\mathbf{A}_E \sqrt{n} T_n + \mathbf{B}_E \sqrt{n} W_n + o_p(1) < b_E | \beta_n^{E'}) - \log \mathbb{P}(\mathbf{A}_E \sqrt{n} T_n + \mathbf{B}_E \sqrt{n} W_n < 0 | \beta_n^{E'}) \big\} = 0.$$

Assumption 1 is a condition on the existence of an exponential moment. This is a necessary condition for a mean statistic based on i.i.d. observations, in our case $\bar{Z}_n$, to satisfy a moderate deviation principle; see [1, 3]. Assumption 2 is necessary to allow a moderate deviation principle for the statistic $(\sqrt{n}(T_n - \mu_n), \Omega_n)^T$, which is obtained by adding to the centered mean statistic $\sqrt{n}\bar{Z}_n$ a remainder term that is converging in probability to 0. The condition

on our remainder term is typically needed to apply moderate deviations to statistics such as M-estimators; see Arcones [1]. Moderate deviations approximations are used to compute probabilities of the form

$$\mathbb{P}\big(\sqrt{n}(T_n, W_n)^T/n^\delta \in \mathcal{K}\big).$$

If we take $\mathcal{K} = \{(t, w) : \begin{bmatrix}\mathbf{A}_E & \mathbf{B}_E\end{bmatrix}(t, w)^T < 0\}$, then the following condition in this assumption allows us to approximate

$$n^{-2\delta} \cdot \log \mathbb{P}\big(\begin{bmatrix}\mathbf{A}_E & \mathbf{B}_E\end{bmatrix}\sqrt{n}(T_n, W_n)^T/n^\delta + n^{-\delta}o_p(1) < n^{-\delta}b_E|\beta_n^{E'}\big)$$

by

$$n^{-2\delta} \cdot \log \mathbb{P}\big(\begin{bmatrix}\mathbf{A}_E & \mathbf{B}_E\end{bmatrix}\sqrt{n}(T_n, W_n)^T/n^\delta < 0|\beta_n^{E'}\big).$$

We are now ready to state

THEOREM 5.1. *Denote* $\mathcal{H}_n = \{(b, \eta, w) : \mathbf{A}_E(b, \eta)^T + \mathbf{B}_E w < n^{-\delta}b_E\}$. *Under the assumptions* 1 *and* 2, *we have*

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n^{2\delta}} \log \mathbb{P}\big(\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E|\beta_n^{E'}\big)\\ + \inf_{(b,\eta,w)\in\mathcal{H}_n} \frac{(b-\beta^*)^T\mathbf{Q}(b-\beta^*)}{2} + \frac{\eta^T\mathbf{N}\eta}{2} + \frac{w^T\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}w}{2} = 0,\end{aligned} \tag{22}$$

where the optimization variables $(b, \eta, w) \in \mathbb{R}^{E'} \times \mathbb{R}^{p-|E'|} \times \mathbb{R}^p$, *and matrices* $\mathbf{Q}$, $\mathbf{N}$ *and* $\boldsymbol{\Sigma}_{\mathcal{G}}$ *are defined in Proposition* 4.1.

REMARK 5.2. The polyhedron in Theorem 5.1 can in fact be more generally replaced with any open and convex subset $\mathcal{K} \subset \mathbb{R}^{2p}$.

REMARK 5.3. If σ^2 is unknown, we can put

$$\mathcal{K} = \big\{(t, w) : \mathbf{A}_E\sqrt{n}t + \mathbf{B}_E\sqrt{n}w \le b_E/\sigma\big\}$$

and compute the probability with respect to the law of $(\sqrt{n}T_n/\sigma, \sqrt{n}W_n/\sigma)^T$ where $\mu_n := (\beta_n^{E'}/\sigma, 0)^T$. In this case, we note that (22) becomes

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n^{2\delta}} \log \mathbb{P}\big(\mathbf{A}_E(\sqrt{n}T_n/\sigma) + \mathbf{B}_E(\sqrt{n}W_n/\sigma) < b_E/\sigma|\beta_n^{E'}, \sigma\big)\\ + (\sigma^2)^{-1} \cdot \inf_{(b,\eta,w)\in\mathcal{H}_n} \frac{(b-\beta^*)^T\mathbf{Q}(b-\beta^*)}{2} + \frac{\eta^T\mathbf{N}\eta}{2} + \frac{w^T\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}w}{2} = 0,\end{aligned} \tag{23}$$

where $\mathcal{H}_n$ is defined in Theorem 5.1. Hence our approximation readily accommodates the case of unknown σ^2. In Section 8, we use this to outline a scheme for posterior sampling when imposing a joint prior on $(\beta^{E'}, \sigma)^T$.

Theorem 5.1 suggests using the negative of the second term in (22), scaled by $n^{2\delta}$, as an approximation in computing the log of the probability in (4), whenever $\sqrt{n}(T_n, W_n)$ satisfies a central limit property and has exponential moments. From now on, we write the selection event as $\{\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E\}$, ignoring the $o_p(1)$ with a slight abuse of notation.

5.2. *A refinement of the approximation.* The optimization variables in the approximation implied by Theorem 5.1 are constrained to the set $\mathcal{H}_n$. In the spirit of the development in Section 3, we propose to relax the computational burden by solving an unconstrained version of the optimization problem in (22). The resulting expression, which we refer to as a *barrier* version of our approximation, will be used in place of the log of the probability in (4) when implementing our methods. To proceed with our construct, we first introduce a change of variable for the optimization arguments to simplify the constraints in the left-hand side of (22).

THEOREM 5.1. *For $0 < \delta < 1/2$ and for a sequence $\beta_n^{E'}$ such that $\sqrt{n}\beta_n^{E'} = n^{\delta}\beta^*$, define a change of variable $\omega \mapsto o$ through*

$$n^{\delta} w = n^{\delta} P_E \begin{pmatrix} b \\ \eta \end{pmatrix} + n^{\delta} Q_E o + r_E,$$

where

$$P_E = -\begin{bmatrix} \mathbf{P}^{E,E'} & \mathcal{I}^{E,E'} \\ \mathbf{F}^{E,E'} & \mathcal{J}^{E,E'} \end{bmatrix}, \qquad Q_E = \begin{bmatrix} \mathbf{Q}^E & 0 \\ \mathbf{C}^E & I \end{bmatrix}, \qquad r_E = \begin{pmatrix} \lambda s^E \\ 0 \end{pmatrix}.$$

Then, minimizing $n^{2\delta} \cdot \inf_{(b,\eta,w)\in\mathcal{H}_n}\{(b-\beta^*)^T\mathbf{Q}(b-\beta^*)/2 + \eta^T\mathbf{N}\eta/2 + w^T\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}w/2\}$ is equivalent to minimizing

$$\begin{aligned} n^{2\delta} \cdot \inf_{\{(b,\eta,o)\in\mathbb{R}^{2p}: o\in\mathcal{O}_n\}} & \{(b-\beta^*)^T\mathbf{Q}(b-\beta^*)/2 + \eta^T\mathbf{N}\eta/2 \\ & + \big(P_E\begin{pmatrix} b \\ \eta \end{pmatrix} + Q_E o + r_E/n^{\delta}\big)^T \boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\big(P_E\begin{pmatrix} b \\ \eta \end{pmatrix} + Q_E o + r_E/n^{\delta}\big)/2\}, \end{aligned}$$

where the constraints in the two objectives are given, respectively, by

$$\begin{aligned} \mathcal{H}_n &= \left\{(b,\eta,w)\in\mathbb{R}^{2p} : \mathbf{A}_E\begin{pmatrix} b \\ \eta \end{pmatrix} + \mathbf{B}_E w \le n^{-\delta} b_E\right\}; \\ \mathcal{O}_n &= \{o\in\mathbb{R}^p : \operatorname{sgn}(n^{\delta}o_E) = s^E, \|n^{\delta}o_{-E}\|_{\infty} \le \lambda\}. \end{aligned}$$

Note that in the new form of the optimization problem, the variables b and η are unconstrained, while o has the simple constraint given above. For the parametrization of $\beta_n^{E'}$ considered, we obtain a more flexible form of the optimization problem as

$$\begin{aligned} &\log\widetilde{\mathbb{P}}(\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E|\beta_n^{E'}) \\ &\quad = -n^{2\delta} \cdot \inf_{(b,\eta,o)\in\mathbb{R}^{2p}} \{(b-\beta^*)^T\mathbf{Q}(b-\beta^*)/2 + \eta^T\mathbf{N}\eta/2 \\ &\qquad + \big(P_E\begin{pmatrix} b & \eta \end{pmatrix}^T + Q_E o + r_E/n^{\delta}\big)^T \boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\big(P_E\begin{pmatrix} b & \eta \end{pmatrix}^T + Q_E o + r_E/n^{\delta}\big)/2 \\ &\qquad + n^{-2\delta}\psi_{n^{-\delta}}(o_E, o_{-E})\}. \end{aligned} \tag{24}$$

In the above formulation, $\psi_s(o) = \psi_s(o_E, o_{-E})$ is some penalty function corresponding to the set $\mathcal{O}_n$, with a scaling factor s. This specializes to the optimization presented in Proposition 5.1 by taking $\psi_s(o_E, o_{-E})$ to be the characteristic function

$$I_{\mathcal{O}}(o_E, o_{-E}) = \begin{cases} 0 & \text{if } o\in\mathcal{O}, \\ \infty & \text{otherwise.} \end{cases}$$

In the next and crucial step, instead of the characteristic function that restricts the optimizing variables to the set $\mathcal{O}_n$, we use a smoother nonnegative penalty function: we replace

$I_{\mathcal{O}_n}(o_E, o_{-E})$ with a suitable "barrier" penalty function ψ_s that reflects preference for values of o farther away from the boundary and inside the constraint region $\mathcal{O}_n$, by taking on smaller values for such o. Specifically, we use $\psi_{n^{-\delta}}$ defined by

$$\begin{aligned}\psi_{n^{-\delta}}(o) &\equiv \psi_{n^{-\delta}}(o_E, o_{-E}) \\ &= \left(\sum_{i=1}^{E} \log\left(1 + \frac{1}{s_{i,E} n^{\delta} o_{i,E}}\right) + \sum_{i=1}^{p-|E|} \log\left(1 + \frac{1}{\lambda_{i,-E} - n^{\delta} o_{i,-E}}\right)\right. \\ &\quad \left. + \log\left(1 + \frac{1}{\lambda_{i,-E} + n^{\delta} o_{i,-E}}\right)\right).\end{aligned} \tag{25}$$

In line with Remark 3.2 for the univariate thresholding example, we remark that δ in the parameterization of $\beta_n^{E'}$ is only a theoretical construct. This ultimately leads to an approximation to the log of the selection-adjusted posterior (7), as

$$\begin{aligned}\log \widetilde{\pi}_S(\beta_n^{E'} | \widehat{\beta}^{E'}) = \log \pi(\beta_n^{E'}) - n(\widehat{\beta}^{E'} - \beta_n^{E'})^T \mathbf{Q} (\widehat{\beta}^{E'} - \beta_n^{E'})/2 \\ - \log \widetilde{\mathbb{P}}(\mathbf{A}_E \sqrt{n} T_n + \mathbf{B}_E \sqrt{n} W_n < b_E | \beta_n^{E'}),\end{aligned} \tag{26}$$

where we use (25) in (24).

6. Point estimation and consistency of the approximate selection-adjusted posterior. In our Bayesian framework a natural point estimate for $\beta_n^{E'}$ is the selection-adjusted maximum a-posteriori (MAP) statistic, the maximizer in $\beta_n^{E'}$ of (7); the selection-adjusted MLE obtains as a special case when we use a constant prior. We cannot implement the exact MAP because it again involves the adjustment factor, but we may consider an approximate MAP by replacing the adjustment factor in the selection-adjusted likelihood with our approximation. Specifically, the previous section prompts using (26) with the choice of penalty (25) in (24). In what follows, we focus on the approximate selection-adjusted MLE and show that it is consistent in a formal, post-selection sense. A closely related phenomenon, concentration of the approximate posterior around the underlying model parameters, in fact follows as a consequence. Finally, on the computational aspect we show that for an extensive class of priors, the MAP problem associated with our approximation is convex and, therefore, amenable for implementation.

Before proceeding we remark that point estimates can be obtained with the methods of Lee et al. [5], that rely on a particular conditional (but exact) distribution of a univariate projection. However, such estimates might be suboptimal because, stated informally, they do not utilize all of the information about the parameters in the sample. By contrast, the methods suggested below are based on the full (adjusted) likelihood, defined for a fairly flexible class of models.

6.1. *The special univariate case.* We study first the approximate MAP in the univariate gaussian example of Section 3. Again, in that case we can implement the exact MAP. As before, we assume (8) with $\beta = \beta_n$ such that $\sqrt{n}\beta_n = n^{\delta}\beta^*$, $0 < \delta \leq 1/2$. Notice, the selection event $\{\sqrt{n_1}\bar{Y}^{\mathcal{S}} > 0\} \equiv \{\sqrt{n}(\bar{Y}_n + W_n) > 0\}$ is of the form

$$\sqrt{n}(\bar{Y}_n + W_n)/n^{\delta} \in \mathcal{K},$$

where $\mathcal{K}$ is an interval on the real line. The selection-adjusted likelihood is

$$L_S^n(\beta_n) = -n(\bar{y}_n - \beta_n)^2/2 - \log \mathbb{P}(\sqrt{n}(\bar{Y}_n + W_n)/n^{\delta} \in \mathcal{K} | \beta_n), \tag{27}$$

and our approximate version for the selection-adjusted (log-) likelihood is

$$\widetilde{L}_S^n(\beta_n) = -n(\bar{y}_n - \beta_n)^2/2 + n^{2\delta} \cdot \inf_{(z,w)\in\mathbb{R}^2} \left\{ \frac{(z-\beta^*)^2}{2} + \frac{\rho w^2}{2(1-\rho)} + \frac{1}{n^{2\delta}} \psi_{n^{-\delta}}(z+w) \right\}. \tag{28}$$

First, we make a crucial observation about the strong concavity of our selection-adjusted sequence of approximate (log-) likelihoods. As in Theorem 3.1, the statements in this section hold also for $\delta = 1/2$; compare Theorem 6.4.

THEOREM 6.1 (Strong concavity of log-likelihood). *Let $\sqrt{n}\beta_n = n^\delta \beta^*$, $0 < \delta \le 1/2$. The approximate selection-adjusted log-likelihood in (28) equals*

$$\widetilde{L}_S^n(\beta_n) = n\bar{y}_n\beta_n - n^{2\delta} \cdot \widetilde{C}_n(\beta^*) - n\bar{y}_n^2/2,$$

where

$$\widetilde{C}_n(\beta^*) = (1-\rho) \cdot \beta^{*2}/2 + \bar{H}_n^*(\rho\beta^*), \tag{29}$$

and with $\bar{H}_n^*(\cdot)$ denoting the convex conjugate of $\bar{H}_n(\bar{z}) = \rho \cdot \bar{z}^2/2 + n^{-2\delta} \cdot \psi_{n^{-\delta}}(\bar{z})$. Moreover, $\widetilde{C}_n(\cdot)$ is strongly convex with index of convexity lower bounded by $(1-\rho)$.

The following result establishes consistency for our approximate selective MLE.

THEOREM 6.1. *Denote by $\widehat{\beta}_S$ the maximizer of (28). If $0 < \delta \le 1/2$, then as $n \to \infty$,*

$$\mathbb{P}\big(n^{1/2-\delta}|\widehat{\beta}_S - \beta_n| > \epsilon | \sqrt{n}(\bar{Y}_n + W_n)/n^\delta \in \mathcal{K}\big) \longrightarrow 0$$

Figure 1(b) shows the approximate MLE against the exact MLE. To (literally) complete the picture, we prove in Theorem 4.1 (Supplement 4) that if selection uses the entire data, then the maximizer of the approximate truncated likelihood is *not* consistent. This further highlights the importance of holding out some samples at the selection stage and reserve them for inference.

As a consequence of Theorem 6.1, we now show a form of consistency of the selection-adjusted posterior law with respect to a fixed (not depending on n) prior, meaning that the approximate selection-adjusted posterior concentrates around the true parameter β_n asymptotically.

THEOREM 6.2. *Consider a ball of radius δ around β_n,*

$$\mathcal{B}(\beta_n, \delta) := \{b_n : |b_n - \beta_n| \le \delta\},$$

and suppose that π is a prior which assigns nonzero probability to $\mathcal{B}(\beta_n, \delta)$ for any $\delta > 0$. Under the conditions in Theorem 6.1, for any $\epsilon > 0$ we have

$$\mathbb{P}\big(\Pi_S(\mathcal{B}^c(\beta_n, \delta)|\bar{Y}_n) > \epsilon | \sqrt{n}(\bar{Y}_n + W_n)/n^\delta \in \mathcal{K}\big) \to 0$$

as $n \to \infty$, where

$$\Pi_S(\mathcal{B}^c(\beta_n, \delta)|\bar{y}_n) := \left(\int \pi(b_n) \cdot \exp\big(\widetilde{L}_S^n(b_n)\big)\, db_n \right)^{-1} \int_{\mathcal{B}^c(\beta_n,\delta)} \pi(b_n) \cdot \exp\big(\widetilde{L}_S^n(b_n)\big)\, db_n$$

is the posterior probability under the approximate truncated likelihood.

The next result has implications for the computation of the approximate selective MAP.

THEOREM 6.3 (Convexity of approximate MAP). *Consider a log-concave prior $\pi(\cdot)$. Then minimizing the negative of the approximate log-posterior*

$$-\log \pi(\beta_n) - \widetilde{L}_S^n(\beta_n)$$

is a convex optimization problem in β_n for any $n \in \mathbb{N}$.

6.2. *The general case.* Having established posterior consistency for the univariate example, we now move on to the general case. Thus, consider the MAP estimator, given as the maximizer of (26). For a flat prior, this reduces to the approximate maximum-likelihood estimate,

$$\begin{aligned}\widehat{\beta}_S^{E'} = \underset{\beta_n^{E'}}{\operatorname{argmin}}\{&n(\widehat{\beta}^{E'} - \beta_n^{E'})'\mathbf{Q}(\widehat{\beta}^{E'} - \beta_n^{E'})/2 \\ &+ \log \widetilde{\mathbb{P}}(\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E|\beta_n^{E'})\},\end{aligned} \tag{30}$$

where in (24) we use (25). Recall the quantitites P_E, Q_E, r_E defined in Proposition 5.1. Let $P_E^{E'}$ denote the matrix that consists of the columns in P_E corresponding to the coordinates in E'. Similarly, $P_E^{-E'}$ denotes the matrix consisting of the remaining $p - |E'|$ columns. We begin by showing that our selection-adjusted sequence of approximate (log-) likelihoods is strongly concave, which follows by proving strong convexity of the corresponding log-partition functions. Notably, this claim holds in finite samples and is therefore true for any $n \in \mathbb{N}$.

THEOREM 6.2 (Strong concavity of log-likelihood). *Under the parameterization $\sqrt{n}\beta_n^{E'} = n^\delta\beta^*$ with $\delta \in (0, 1/2)$ and when $\mathbf{X}_{E'}$ is of full column rank, the approximate log-partition function $n^{2\delta}\widetilde{C}_n(\mathbf{Q}\beta^*)$, corresponding to the approximate negative log-likelihood,*

$$n(\widehat{\beta}^{E'} - \beta_n^{E'})^T\mathbf{Q}(\widehat{\beta}^{E'} - \beta_n^{E'})/2 + \log\widetilde{\mathbb{P}}(\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E|\beta_n^{E'}),$$

is strongly convex. Furthermore, the index of strong convexity for $\widetilde{C}_n(\mathbf{Q}\beta^*)$ is bounded below by $\lambda_{\min}$, the smallest eigenvalue of

$$(\mathbf{Q} + {P_E^{E'}}^T\mathbf{\Sigma}_{\mathcal{G}}^{-1}P_E^{E'})^{-1}.$$

Using Proposition 6.2, we are now able to prove consistency for the approximate selective MLE. We observe here again that randomization is crucial for the theorem below to hold, as we saw already for the univariate example.

THEOREM 6.4. *Consider the conditions in Theorem 5.1. Assume*

$$\operatorname{Var}(\sqrt{n}\widehat{\beta}^{E'}|\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E) = O(1).$$

When $\sqrt{n}\beta_n^{E'} = n^\delta\beta^*$ for $\delta \in (0, 1/2)$, the approximate selective MLE given in (30) is $n^{1/2-\delta}$-consistent for $\beta_n^{E'}$ under the selection-adjusted law (5):

$$\mathbb{P}(n^{1/2-\delta}\|\widehat{\beta}_S^{E'} - \beta_n^{E'}\| > \epsilon|\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E) \longrightarrow 0$$

as $n \to \infty$.

As a consequence, the posterior distribution with respect to a fixed (not depending on n) prior concentrates around the true parameters.

THEOREM 6.5. *Assume the conditions in Theorem 6.4. Consider a ball of radius δ around $\beta_n^{E'}$,*

$$\mathcal{B}(\beta_n^{E'}, \delta) \equiv \{b_n : \|b_n - \beta_n^{E'}\| \le \delta\},$$

and suppose that π is a prior that assigns nonzero probability to $\mathcal{B}(\beta_n^{E'}, \delta)$ for any $\delta > 0$. Then

$$\mathbb{P}(\Pi_S(\mathcal{B}^c(\beta_n^{E'}, \delta)|\widehat{\beta}^{E'}) > \epsilon|\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E) \to 0$$

as $n \to \infty$ for any $\epsilon > 0$, where $\Pi_S(\cdot)$ denotes the posterior probability under the approximate truncated likelihood,

$$\Pi_S\big(\mathcal{B}^c(\beta_n^{E'}, \delta)|\widehat{\beta}^{E'}\big) := \bigg(\int \pi(b_n) \cdot \exp\big(\widetilde{L}_S^n(b_n)\big)\, db_n\bigg)^{-1} \int_{\mathcal{B}^c(\beta_n^{E'}, \delta)} \pi(b_n) \cdot \exp\big(\widetilde{L}_S^n(b_n)\big)\, db_n,$$

for $\widetilde{L}_S^n(b_n) = \sqrt{n}\widehat{\beta}^{E'}\mathbf{Q}\sqrt{n}b_n - n^{2\delta}\widetilde{C}_n(n^{1/2-\delta}\mathbf{Q}b_n) - n\widehat{\beta}^{E'}\mathbf{Q}\widehat{\beta}^{E'}/2$.

Finally, concerning implementation of the approximate MAP we offer the next result.

THEOREM 6.6 (Convexity of approximate MAP). *Let $\pi(\cdot)$ be a log-concave prior. For any barrier function $\psi_s(\cdot)$, minimizing the negative of the approximate log-posterior based on the approximation in* (24),

$$\log \pi\big(\beta_n^{E'}\big) + n\big(\widehat{\beta}^{E'} - \beta_n^{E'}\big)^T\mathbf{Q}\big(\widehat{\beta}^{E'} - \beta_n^{E'}\big)/2 + \log \widetilde{\mathbb{P}}\big(\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E|\beta_n^{E'}\big),$$

in $\beta_n^{E'}$, is a convex optimization problem for any $n \in \mathbb{N}$.

7. Empirical analysis. In this section, we compare our methods to existing alternatives in simulation and real-data examples.

7.1. *Simulation study.* We next present some simulation results that demonstrate the effectiveness of the proposed methods. In each of 50 rounds, we draw a $n \times p$ matrix $\mathbf{X}$ with $n = 500$, $p = 100$ such that the rows $(X_{i1}, \ldots, X_{ip})^T \sim N_p(0, \mathbf{\Sigma})$, $i = 1, 2, \ldots, 500$, and where the (j, k)th entry of $\mathbf{\Sigma}$ equals $0.20^{|j-k|}$. The components of $\beta \in \mathbb{R}^{100}$ are drawn i.i.d. from a normal mixture,

$$0.9 \cdot \mathcal{N}_1(\beta_j; 0, 0.1) + 0.1 \cdot \mathcal{N}_1(\beta_j; 0, V), \tag{31}$$

and we take $V \in \{5, 3.5, 2\}$, roughly corresponding to three different signal regimes. Finally, we draw $\mathrm{y}|\mathbf{X}, \beta \sim \mathcal{N}_n(\mathbf{X}\beta, \mathbf{I})$.

We split the data to form $(\mathrm{y}^{\mathcal{S}}, \mathbf{X}^{\mathcal{S}})$, where $\mathcal{S} \subset \{1, \ldots, n\}$ is a random subset of size $|\mathcal{S}| = n/2 = 250$. We select with Lasso using a theoretical value for the tuning parameter. Specifically, we use $\lambda = \mathbb{E}[\|\mathbf{X}^T\Psi\|_\infty]$, $\Psi \sim \mathcal{N}_n(0, \mathbf{I})$, as proposed in Negahban et al. [7], and denote by $E \subseteq \{1, \ldots, 100\}$ the set corresponding to the nonzero components of the Lasso estimate,

$$\operatorname*{argmin}_\beta \frac{1}{2\rho}\big\|\mathrm{y}^{\mathcal{S}} - \mathbf{X}^{\mathcal{S}}\beta\big\|^2 + \lambda\|\beta\|_1. \tag{32}$$

In the inference stage, we have access to the entire data $(\mathrm{y}, \mathbf{X})$. We give inference assuming that

$$\mathrm{y}|\mathbf{X}, \beta^E \sim \mathcal{N}\big(\mathbf{X}_E\beta^E, \sigma^2\mathbf{I}\big), \tag{33}$$

in other words, we take $E' = E$ in (2). Four different methods for inference are compared:

Unadjusted (Naive). Bayesian inference for β^E using a noninformative prior $\pi(\beta^E) \propto 1$ and the unadjusted likelihood $\mathrm{y}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{X}_E\beta^E, \sigma^2\mathbf{I})$.

Split. Bayesian inference using only the confirmatory (held-out) data: a noninformative prior $\pi(\beta^E) \propto 1$ is prepended to the unadjusted likelihood $\mathrm{y}^{\mathcal{S}^c}|\mathbf{X}^{\mathcal{S}^c} \sim \mathcal{N}_n(\mathbf{X}_E^{\mathcal{S}^c}\beta^E, \sigma^2\mathbf{I})$.

Carving. Bayesian inference for β^E using a noninformative prior $\pi(\beta^E) \propto 1$ and the approximate selection-adjusted likelihood,

$$\widetilde{\pi}_S\big(\beta^E|\widehat{\beta}^E\big) \propto \frac{\exp(-n(\widehat{\beta}^E - \beta^E)^T\mathbf{Q}(\widehat{\beta}^E - \beta^E)/2\sigma^2)}{\widetilde{\mathbb{P}}(\mathbf{A}_E\sqrt{n}T_n + \mathbf{B}_E\sqrt{n}W_n < b_E|\beta^E)}, \tag{34}$$

TABLE 1

Summary of simulation study. Four methods are compared in three signal regimes. For Lee et al., the numbers displayed for length of CI are averages of constructed intervals that had finite length; for $V = 5, 3.5, 2$, Lee et al. produced infinitely long intervals for 1.4%, 1.95% and 4.73% of the selected parameters, on average

	$V = 5$			$V = 3.5$			$V = 2$		
	$1 - \text{FCR}$	Length	Rel. Risk	$1 - \text{FCR}$	Length	Rel. Risk	$1 - \text{FCR}$	Length	Rel. Risk
Carving	0.884	3.92	0.21	0.887	4.04	0.39	0.905	4.22	0.66
Naive	0.753	3.33	0.25	0.728	3.33	0.44	0.598	3.31	0.75
Split	0.9	4.80	0.25	0.915	4.79	0.44	0.902	4.77	0.70
Lee et al	0.913	8.38	0.29	0.939	9.58	0.46	0.863	10.73	0.83

where the denominator on the right-hand side is given by (24) with the penalty defined in (25).

Lee et al. We use the "selectiveInference" package in R [16] that implements the exact methods of Lee et al. [5]. Because there is no implementation for carving, the entire data is used for selection, that is, E is obtained here by setting $\mathcal{S} = \{1, 2, \ldots, n\}$.

The four methods above are compared on the following criteria: (i) we use each method to construct (marginal) 90% interval estimates, and calculate the average (over simulation rounds) proportion of covering intervals (this is reported as "$1 - \text{FCR}$" in Table 1); (ii) lengths of constructed CIs; and (iii) relative prediction risk,

$$\frac{(\widehat{\beta} - \beta)^T (\mathbf{X}^T \mathbf{X})(\widehat{\beta} - \beta)}{\beta^T (\mathbf{X}^T \mathbf{X})\beta},$$

where the "inactive" coordinates $\{j \notin E\}$ of β and $\widehat{\beta}$ are set to zero, and the estimate $\widehat{\beta}$ is the posterior mean for the first three methods, and the plain Lasso estimate for the fourth (Lee et al.).

We see that for all methods except the unadjusted, the coverage, as measured by one minus the false coverage rate (FCR), is roughly the nominal level 0.9. In particular, the CIs constructed based on the proposed approximation to the selection-adjusted posterior have good coverage. We emphasize that this is in spite of the fact the the assumed model (33) is (with high probability) misspecified. Meanwhile, the length of the intervals for the proposed method ("Carving" in Table 1), is much smaller than that for Lee et al. intervals, which are sometimes infinitely long. More importantly, the carved intervals based on our method are smaller in length than the intervals for sample splitting. This matches our expectations because, as opposed to carving, sample-splitting exploits only information in the held-out data. Investigations with varying sample size in this example setting are included in Section 5 of the supplement.

7.2. *HIV drug-resistance data.* In this section, we apply our methods to the HIV dataset analyzed in [2, 13]. With an attempt to understand the genetic basis of drug resistance in HIV, [13] used markers of inhibitor mutations to predict susceptibility to 16 antiretroviral drugs. We follow [2] and focus on the protease inhibitor subset of the data, and on one particular drug, Lamivudine (3TC), where the goal is to identify mutations associated with response to 3TC. There are $n = 633$ cases and $p = 91$ different mutations occurring more than 10 times in the sample.

In the selection stage, we applied the Lasso to a 80% split of the data, with the regularization parameter set to the theoretical value proposed in Negahban et al. [7]. This resulted in 17 selected mutations, corresponding to the nonzero Lasso estimates. Figure 2 shows 90%

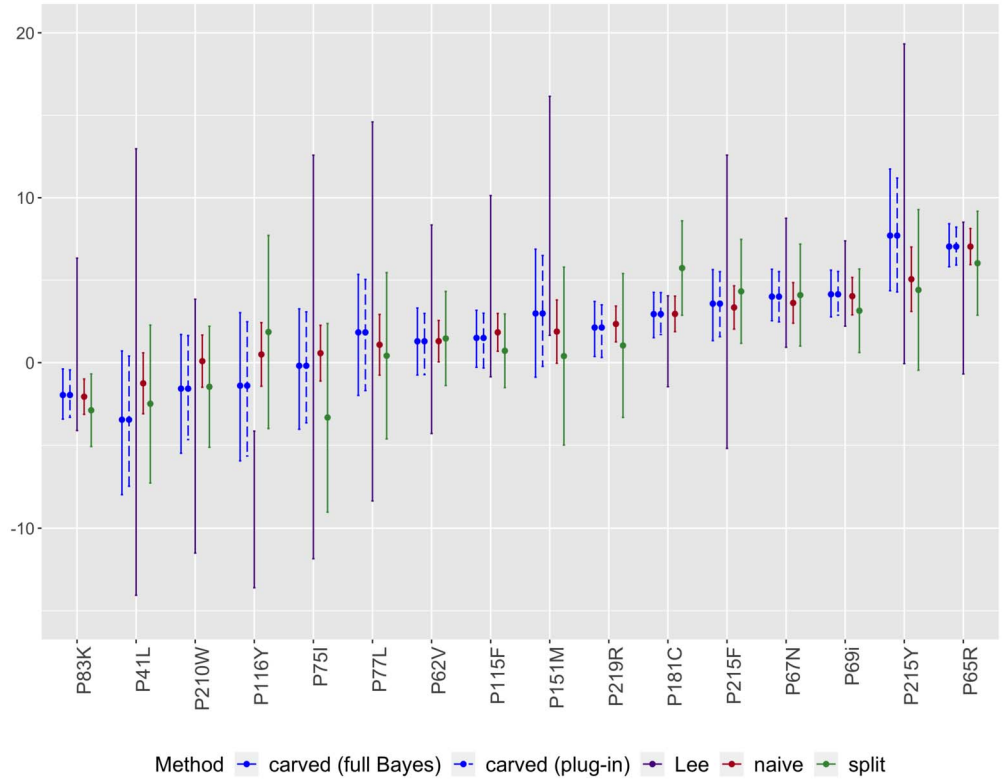


FIG. 2. *Interval and point estimates for selected features. To allow convenient visualization, the figure is not showing mutation 'P184V', which has a different scale from the other variables.*

interval estimates constructed for selected variables (excluding mutation "P184V") according to four different methods: "naive" is the usual, unadjusted intervals using the entire data; "split" uses only the 20% left-out portion of the data to construct the intervals; "Lee" is the adjusted confidence intervals of [5] and based on the original 80% portion used for selection; finally, "carved" are the intervals relying on the methods we propose in the current paper. Of course, we cannot assume knowledge of σ^2 in this analysis, and we offer two different implementations to handle an unknown variance. The first is to estimate σ^2 from a least squares fit against the 91 available predictors, and then simply plug this estimate in (as if it were data independent). The second approach is to add a prior on σ^2 as described in Section 8. Specifically, we model

$$\pi\big(\beta^{E'}|\sigma^2\big) \propto \sigma^{-|E'|} \exp\big(-\|\beta^{E'}\|^2/(2c^2\sigma^2)\big), \qquad \pi\big(\sigma^2\big) \propto \text{Inv-Gamma}(a;b), \tag{35}$$

where we denote $\text{Inv-Gamma}(a;b) := (\sigma^2)^{-a-1}\exp(-b/\sigma^2)$, and where c^2 is a large constant. In this example we used $a = b = 0.1$ for the other hyperparameters. The two implementations are referred to as "plug-in" and "full Bayes" in Figure 2.

All four methods are implemented assuming a linear model consisting of the selected variables. If the model were correctly specified, "split" and "Lee" would both produce valid interval estimates, although the first uses only the held-out data, and the second relies only on the portion of the data used for selection. Our "carved" intervals are approximate, but they utilize information from both splits of the data. The "naive" intervals are invalid. In terms of length, it can be seen that the two implementations of carving ("plug-in" and "full Bayes") produce similar interval estimates, that are shorter than both "split" and "Lee." As in the simulation of the previous subsection, that "Lee" interval estimates are considerably longer than "split" matches our expectations.

The figure also shows point estimates: for "naive" these are just the Lasso (nonzero) estimates, for "split" these are least-squares, and for "carved" these are (approximate) posterior modes. Lacking knowledge of the "true" means, it is hard to compare the different estimates and directions of shrinkage; "split" estimates are unbiased under the assumed model, while "carved" arguably have smaller variance (because they use additional data). Figure 1 in the Supplement depicts the effect-size estimates of all variables in the selected set, including mutation "P184V."

8. Computation. We provide computational details for sampling from the approximate posterior and solving for the approximate MAP (MLE) problem. Note that both of these problems require computing the gradient of the approximate log-posterior, which amounts to solving for the optimizing variables in a certain convex optimization problem. Throughout this section, we assume that π is a log-concave prior and the columns of $\mathbf{X}$ are scaled by $\sqrt{n}$ (but suppress the subscript n).

Treating first the case where σ^2 is known, assume again without loss of generality that $\sigma^2 = 1$. To sample from the log-concave (approximate) posterior incorporating (24), we use a Langevin random walk to obtain a sample of size n_S from the (approximate) selection-adjusted posterior $\widetilde{\pi}_S(\cdot)$. As a function of the previous draw, the $(K+1)$th draw for $\beta^{E'}$ of the Langevin sampler is computed as

$$\beta^{E'}_{(K+1)} = \beta^{E'}_{(K)} + \gamma \cdot \nabla \log \widetilde{\pi}_S\big(\beta^{E'}_{(K)} | \widehat{\beta}^{E'}\big) + \sqrt{2\gamma} \cdot \epsilon_{(K)}, \tag{36}$$

where $\epsilon_{(K)}$ for $K = 1, 2, \ldots, n_S$ are independent draws from a centered Gaussian with unit variance, and γ is a predetermined step size. The sampler takes a noisy step along the gradient of the log-posterior, with no accept–reject step—compare to the usual Metropolis Hastings (MH) algorithm. Hence, at each draw of the sampler, the main computational cost is calculating the gradient of the approximate (log-) selection-adjusted posterior.

Recalling the approximation based on (24), we note that the approximate log-posterior is

$$\begin{aligned} &\log \pi\big(\beta^{E'}_{(K)}\big) - {\beta^{E'}_{(K)}}^T \mathbf{Q} \beta^{E'}_{(K)}/2 + {\beta^{E'}_{(K)}}^T \mathbf{Q} \widehat{\beta}^{E'} - \log \widetilde{\mathbb{P}}\big(\widehat{E} = E, \widehat{S}^{\widehat{E}} = s^E | \beta^{E'}_{(K)}\big), \text{ where} \\ &\log \widetilde{\mathbb{P}}\big(\widehat{E} = E, \widehat{S}^{\widehat{E}} = s^E | \beta^{E'}_{(K)}\big) = -\inf_{b \in \mathbb{R}^{|E'|}} \big\{ \big(b - \beta^{E'}_{(K)}\big)^T \mathbf{Q} \big(b - \beta^{E'}_{(K)}\big)/2 + H_\psi(b) \big\}. \end{aligned} \tag{37}$$

Above, H_ψ is a function of $b = (\eta, o) \in \{(\eta, o) : \eta \in \mathbb{R}^{p-|E'|}, o \in \mathbb{R}^p\}$, and given by

$$\begin{aligned} H_\psi(b) = \inf_{(\eta,o) \in \mathbb{R}^{2p-|E'|}} \Big\{ & \eta^T \mathbf{N} \eta/2 \\ & + \big(P_E \begin{pmatrix} b & \eta \end{pmatrix}^T + Q_E o + r_E\big)^T \mathbf{\Sigma}_{\mathcal{G}}^{-1} \big(P_E \begin{pmatrix} b & \eta \end{pmatrix}^T + Q_E o + r_E\big)/2 \\ & + \sum_{i=1}^{E} \log(1 + 1/s_{i,E} o_{i,E}) + \sum_{i=1}^{p-|E|} \log\big(1 + 1/(\lambda_{i,-E} - o_{i,-E})\big) \\ & + \log\big(1 + 1/(\lambda_{i,-E} + o_{i,-E})\big) \Big\}. \end{aligned}$$

Denoting by $\bar{H}^*(\cdot)$ the conjugate of $\bar{H}(b) = b^T \mathbf{Q} b/2 + H_\psi(b)$, we can write

$$\begin{aligned} \log \widetilde{\mathbb{P}}\big(\widehat{E} = E, \widehat{S}^{\widehat{E}} = s^E | \beta^{E'}_{(K)}\big) &= -{\beta^{E'}_{(K)}}^T \mathbf{Q} \beta^{E'}_{(K)}/2 + \sup_{b \in \mathbb{R}^E} \big\{ b^T \mathbf{Q} \beta^{E'}_{(K)} - \big(b^T \mathbf{Q} b/2 + H_\psi(b)\big) \big\} \\ &= -{\beta^{E'}_{(K)}}^T \mathbf{Q} \beta^{E'}_{(K)}/2 + \bar{H}^*\big(\mathbf{Q} \beta^{E'}_{(K)}\big). \end{aligned}$$

Letting $\hat{b}(\mathbf{Q}\beta^{E'}_{(K)})$ be the maximizer of $\{b^T\mathbf{Q}\beta^{E'}_{(K)} - (b^T\mathbf{Q}b/2 + H_\psi(b))\}$, we have

$$\nabla \log \widetilde{\mathbb{P}}(\widehat{E} = E, \widehat{S}^{\widehat{E}} = s^E | \beta^{E'}_{(K)}) = -\mathbf{Q}\beta^{E'}_{(K)} + \mathbf{Q}\nabla \bar{H}^{-1}(\mathbf{Q}\beta^{E'}_{(K)}) = -\mathbf{Q}\beta^{E'}_{(K)} + \mathbf{Q}\hat{b}(\mathbf{Q}\beta^{E'}_{(K)}),$$

and, finally, the gradient of the log-posterior in (36) equals

$$\nabla \log \pi(\beta^{E'}_{(K)}) + \mathbf{Q}\widehat{\beta}^{E'} - \mathbf{Q}\hat{b}(\mathbf{Q}\beta^{E'}_{(K)}). \tag{38}$$

For the MAP problem, we note that the approximate MAP minimizes the convex objective

$$\underset{\beta^{E'} \in \mathbb{R}^{|E'|}}{\text{minimize}} - \log \pi(\beta^{E'}) - \beta^{E'^T}\mathbf{Q}\widehat{\beta}^{E'} + \bar{H}^*(\mathbf{Q}\beta^{E'}).$$

Employing a gradient descent algorithm to compute the approximate MAP, we note that the Kth update can be written as

$$\widehat{\beta}^{E'}_{S;(K+1)} = \widehat{\beta}^{E'}_{S;(K)} - \eta^T \cdot (\mathbf{Q}\hat{b}(\mathbf{Q}\widehat{\beta}^{E'}_{S;(K)}) - \nabla \log \widetilde{\pi}(\widehat{\beta}^{E'}_{S;(K)}) - \mathbf{Q}\widehat{\beta}^{E'}), \tag{39}$$

involving again the optimizer $\hat{b}(\mathbf{Q}\beta^{E'})$.

Next, suppose that σ^2 is unknown. In that case, we propose to replace (6) with a joint prior on $\beta^{E'}$ and σ^2. For example, we can take the conjugate prior (35), a large c^2 entailing a diffuse prior for $\beta^{E'}$ conditionally on σ^2. Let $\widetilde{L}_S(\beta^{E'}, \sigma^2)$ denote the approximate (log) selection-adjusted likelihood where we plug in our approximation for the selection probability, given by (23)

$$\begin{aligned} &\log \widetilde{\mathbb{P}}(\widehat{E} = E, \widehat{S}^{\widehat{E}} = s^E | \beta^{E'}_{(K)}, \sigma^2) \\ &\quad = -(\sigma^2)^{-1} \cdot \inf_{b \in \mathbb{R}^{|E'|}} \{(b - \beta^{E'}_{(K)})^T\mathbf{Q}(b - \beta^{E'}_{(K)})/2 + H_\psi(b)\}. \end{aligned} \tag{40}$$

We employ a Gibbs scheme to draw a new sample $(\beta^{E'}_{(K+1)}, \sigma^2_{(K+1)})^T$ from the resulting posterior, whose logarithm equals

$$\log \pi(\beta^{E'}, \sigma^2) + \widetilde{L}_S(\beta^{E'}, \sigma^2)$$

up to a constant, under the prior in (35). Using our previous notation, let us denote the posterior distribution for $\beta^{E'}$ conditional on σ^2 by $\widetilde{\pi}_S(\beta^{E'}|\widehat{\beta}^{E'}, \sigma^2)$. Conditional on the Kth update for $\sigma^2_{(K)}$, we make a fresh draw $\beta^{E'}_{(K+1)}$ through a noisy update along the gradient of the log-posterior, as described previously in (36):

$$(\beta^{E'})\text{-Update:} \quad \beta^{E'}_{(K+1)} \leftarrow \beta^{E'}_{(K)} + \gamma \cdot \nabla \log \widetilde{\pi}_S(\beta^{E'}_{(K)}|\widehat{\beta}^{E'}, \sigma^2_{(K)}) + \sqrt{2\gamma} \cdot \epsilon_{(K)}.$$

Observe that the gradient of the log-posterior which leads to our sample update equals

$$-\beta^{E'}_{(K)}/c^2\sigma^2_{(K)} + \mathbf{Q}\widehat{\beta}^{E'}/\sigma^2_{(K)} - \mathbf{Q}\hat{b}(\mathbf{Q}\beta^{E'}_{(K)})/\sigma^2_{(K)}. \tag{41}$$

We alternate this step with updates for the variance parameter; see, for example, Tong, Morzfeld and Marzouk [17]. Conditional on $\beta^{E'}_{(K+1)}$, we see that the choice of an inverse-gamma prior for the variance parameter serves as a conjugate prior as it does usually in Bayesian linear regression. Specifically, the posterior distribution of $\sigma^2_{(K+1)}$ conditional on $\beta^{E'}_{(K+1)}$ is an inverse-gamma random variable with density proportional to

$$(\sigma^2)^{-|E'|/2-p-a-1} \exp(-(b - \widetilde{L}_S(\beta^{E'}_{(K+1)}, 1) + (\beta^{E'}_{(K+1)})^T\beta^{E'}_{(K+1)}/2c^2))/\sigma^2).$$

That is, we now sample

$$(\sigma^2)\text{-Update:} \quad \sigma^2_{(K+1)} \leftarrow \text{Inv-Gamma}(|E'|/2 + p + a; \widetilde{b}),$$

where the updated hyperparameter $\widetilde{b}(\beta^{E'}_{(K+1)})$ equals

$$b - \widetilde{L}_S(\beta^{E'}_{(K+1)}, 1) + (\beta^{E'}_{(K+1)})^T\beta^{E'}_{(K+1)}/2c^2.$$

9. Discussion. To address the problem of inference after variable selection, we propose methods based on an approximation to the adjustment factor—the denominator in a selection-adjusted likelihood. Our proposal applies to a large class of selection rules, including ones that involve randomization. By working directly with the full truncated likelihood, we obviate the need to differentiate between various models according to choices of the statistician. In Panigrahi, Zhu and Sabatti [10], the approximation proposed in our paper is employed to obtain tractable pivotal quantities, and more recently Panigrahi and Taylor [8] and Panigrahi et al. [11] build on this approximate scheme to explore a maximum-likelihood approach and propose scalable samplers for integrative models in a high dimensional radiogenomic context, respectively.

There is certainly room for further research and extensions of the current work. On the methodological side, it would be interesting to investigate if a variational Bayes approach can be taken instead of implementing MCMC sampling schemes for posterior updates. From a theoretical point of view, we have shown consistency properties of the posterior that appends a "carved" likelihood to a prior, that is, we proved that such a posterior concentrates around the true underlying parameter with probability converging to one as the sample size increases. Empirical evidence showing that our credible intervals (under a diffuse prior) are similar to the frequentist post-selection intervals, suggests that the frequentist guarantees that we provided can be strengthened, for example, by presenting a Bernstein–von Mises-type of result.

Acknowledgments. The idea of providing Bayesian adjusted inference after variable selection was previously proposed by Daniel Yekutieli and Edward George, and presented at the 2012 Joint Statistical Meetings in San Diego. Our interest re-arose with the recent developments on exact post-selection inference in the linear model. A.W. is thankful to Daniel and Ed for helpful conversations and to Ed for pointing out the paper by Bayarri and DeGroot. S.P. is thankful to Xuming He, Liza Levina, Rina Foygel Barber and Yi Wang for reading an initial version of the draft and offering valuable comments and insights. The authors would also like to sincerely thank and acknowledge Chiara Sabatti for the long discussions and for her suggestions along the way.

Funding. J.T. was supported in part by ARO Grant 70940MA. A.W. was partially supported by ERC Grant 030-8944 and by ISF Grant 039-9325.

SUPPLEMENTARY MATERIAL

Supplement to "Integrative methods for post-selection inference under convex constraints" (DOI: 10.1214/21-AOS2057SUPP; .pdf). Supplementary information.

REFERENCES

[1] ARCONES, M. A. (2002). Moderate deviations for M-estimators. *TEST* **11** 465–500. MR1947607 https://doi.org/10.1007/BF02595717

[2] BI, N., MARKOVIC, J., XIA, L. and TAYLOR, J. (2020). Inferactive data analysis. *Scand. J. Stat.* **47** 212–249. MR4075236 https://doi.org/10.1111/sjos.12425

[3] EICHELSBACHER, P. and LÖWE, M. (2003). Moderate deviations for i.i.d. random variables. *ESAIM Probab. Stat.* **7** 209–218. MR1956079 https://doi.org/10.1051/ps:2003005

[4] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. arXiv preprint, arXiv:1410.2597.

[5] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

[6] MARKOVIC, J. and TAYLOR, J. (2016). Bootstrap inference after using multiple queries for model selection. arXiv preprint, arXiv:1612.07811.

[7] 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* 1348–1356.

[8] PANIGRAHI, S. and TAYLOR, J. (2019). Approximate selective inference via maximum likelihood. arXiv preprint, arXiv:1902.07884.

[9] PANIGRAHI, S., TAYLOR, J. and WEINSTEIN, A. (2021). Supplement to "Integrative methods for post-selection inference under convex constraints." https://doi.org/10.1214/21-AOS2057SUPP

[10] PANIGRAHI, S., ZHU, J. and SABATTI, C. (2021). Selection-adjusted inference: An application to confidence intervals for cis-eQTL effect sizes. *Biostatistics* **22** 181–197. MR4207150 https://doi.org/10.1093/biostatistics/kxz024

[11] 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.

[12] REID, S., TAYLOR, J. and TIBSHIRANI, R. (2017). Post-selection point and interval estimation of signal sizes in Gaussian samples. *Canad. J. Statist.* **45** 128–148. MR3646193 https://doi.org/10.1002/cjs.11320

[13] RHEE, S.-Y., TAYLOR, J., WADHERA, G., BEN-HUR, A., BRUTLAG, D. L. and SHAFER, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. USA* **103** 17355–17360.

[14] TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381 https://doi.org/10.1214/17-AOS1564

[15] TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689 https://doi.org/10.1080/01621459.2015.1108848

[16] TIBSHIRANI, R., TIBSHIRANI, R., TAYLOR, J., LOFTUS, J., REID, S. and MARKOVIC, J. (2019). selectiveInference: Tools for post-selection inference. R package version 1.2.5.

[17] TONG, X. T., MORZFELD, M. and MARZOUK, Y. M. (2020). MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure. *SIAM J. Sci. Comput.* **42** A1765–A1788. MR4111677 https://doi.org/10.1137/19M1284014

[18] WEINSTEIN, A., FITHIAN, W. and BENJAMINI, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *J. Amer. Statist. Assoc.* **108** 165–176. MR3174610 https://doi.org/10.1080/01621459.2012.737740

[19] YEKUTIELI, D. (2012). Adjusted Bayesian inference for selected parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 515–541. MR2925372 https://doi.org/10.1111/j.1467-9868.2011.01016.x

[20] YU, G., BIEN, J. and TIBSHIRANI, R. (2019). Reluctant interaction modeling. arXiv preprint, arXiv:1907.08414.

[21] ZÖLLNER, S. and PRITCHARD, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80** 605–615.