Scalable methods for Bayesian selective inference

Snigdha Panigrahi[1] and Jonathan Taylor[2]

[1]*Department of Statistics*
University of Michigan
e-mail: psnigdha@umich.edu

[2]*Department of Statistics*
Stanford University
e-mail: jonathan.taylor@stanford.edu

Abstract: Modeled along the truncated approach in [20], selection-adjusted inference in a Bayesian regime is based on a *selective posterior*. Such a posterior is determined together by a generative model imposed on data and the selection event that enforces a truncation on the assumed law. The effective difference between the *selective posterior* and the usual Bayesian framework is reflected in the use of a truncated likelihood. The normalizer of the truncated law in the adjusted framework is the probability of the selection event; this typically lacks a closed form expression leading to the computational bottleneck in sampling from such a posterior. The current work provides an optimization problem that approximates the otherwise intractable *selective posterior* and leads to scalable methods that give valid post-selective Bayesian inference. The selection procedures are posed as data-queries that solve a randomized version of a convex learning program which have the advantage of preserving more left-over information for inference.

We propose a randomization scheme under which the approximating optimization has *separable* constraints that result in a partially *separable* objective in *lower* dimensions for many commonly used selective queries. We show that the proposed optimization gives a valid exponential rate of decay for the selection probability on a large deviation scale under a Gaussian randomization scheme. On the implementation side, we offer a primal-dual method to solve the optimization problem leading to an approximate posterior; this allows us to exploit the usual merits of a Bayesian machinery in both low and high dimensional regimes when the underlying signal is effectively sparse. We show that the adjusted estimates empirically demonstrate better frequentist properties in comparison to the unadjusted estimates based on the usual posterior, when applied to a wide range of constrained, convex data queries.

Keywords and phrases: Approximate posterior, Bayesian inference, randomized queries, selective posterior, truncated likelihood.

Received September 2017.

Contents

1. Introduction

A line of works [13, 9, 28, 14, 29, 25] has established methodology for exact and asymptotic selection-adjusted inference that provide frequentist coverage guarantees in the regression framework. The driving motivation to adjust for selection is that analysts commonly conduct queries on a database in order to select inferential questions of interest about the population parameters. Inference after such interactions with the data lacks frequentist properties like target coverage when the same data set is used later for answering these very same questions. A Bayesian perspective on modeling the post-selective problem as a truncation is advocated in [30] and extensions of the former work to the more general set-up of linear models are proposed in [20]. These works propose the use of a fixed parameter view where the truncation is applied to the data exclusively conditional on the parameter. This alters the posterior distribution after selection unlike the usual Bayesian variable selection framework in [17, 10] where the posterior is known to display inadapativity to selection.

More precisely, the truncated view point on inference is based on a *selective posterior*, formed by a truncated likelihood in conjunction with a prior that allows an analyst to inject a priori information on parameters in a model after selection. Such an approach has the additional flexibility in allowing the analyst

to fix a model based on a parametrization that can be guided by a selection procedure. Motivated by the conditional approach of modeling Bayesian inference, the current work focuses on developing concrete, scalable methods that will allow the analyst to exploit the full potent of a Bayesian machinery post a wide range of constrained-convex learning programs. The Bayesian problem is by no means a trivial extension of the existing frequentist methods as it requires a closed form expression for the normalizer of the truncated likelihood. We describe the computational difficulties in providing Bayesian inference in the truncated framework and the contributions of this work more formally after introducing the *selective posterior*.

1.1. Selective posterior

A selective posterior modeled along the conditional approach has two components – a truncated likelihood and a prior distribution on the parameters in the likelihood. The truncation is imposed by selection as the analyst is interested in providing inference for a target parameter only if he observed the associated selection event. A generative model that the analyst is willing to impose on data post selection, together with the truncation to all realized values that lead to an observed selection event determine the truncated likelihood. The prior allows him to inject information on the target from his existing knowledge.

Formally, variable selection is based on an observed data vector S and the selection event of observing an active set of variables $\hat{E}(S) = E$ can be described as $\{s : \hat{E}(s) = E\}$, the set of realizations of S that lead to E. It is only after selection that a model is defined, in this case, a generative Bayesian model with a likelihood parametrized by β^*, denoted as $f(s|\beta^*)$ and a prior $\beta^* \sim \pi$. The goal is to provide inference for a target determined by selection event E; that is, we infer about the target only if we observe E. This truncates the generative law of the data conditional on the parameter, resulting in the selection adjusted likelihood

$$f(s|\beta^*)1_{\{\hat{E}(s)=E\}}/\mathbb{P}(\hat{E}(S) = E|\beta^*).$$

In conjunction with the prior $\pi(\beta^*)$, selective Bayesian inference about the data adaptive target is possible by sampling from

$$\pi_E(\beta^*|s) \propto \pi(\beta^*) \cdot \frac{f(s|\beta^*)}{\mathbb{P}(\hat{E}(S) = E|\beta^*)} \tag{1.1}$$

namely, the selective posterior.

As is evident from above, the normalizer of the truncated likelihood is the probability of the selection event, computed as a function of the parameters in the generative density. While sampling from the truncated likelihood in a frequentist regime does not require knowledge of the normalizer (treated as

a constant), the normalizer that typically lacks a closed form expression does contribute to the selective posterior in a Bayesian paradigm. Implementing a sampling scheme then becomes impossible in the absence of an expression for the normalizer to the truncated law. [20] identifies this technical hurdle and proposes an approximation to general affine selection probabilities which gives rise to a pseudo *selective posterior*. Sampling from the *selective posterior* necessitates computing the approximation, cast as an optimization problem for each draw of the sampler. The efficiency of any standard sampler thereby, hinges on the computational cost of solving the optimization objective associated with this approximation. In most cases, this can be very expensive and hard to scale with larger sample size and regression dimensions.

We propose in this work a randomization scheme for commonly used selection queries and offer an approximating optimization under the same to facilitate sampling from an approximate selective posterior. The three major gains in the current work associated with the proposed optimization are

- an objective with simpler constraints on the optimizing variables, as opposed to polyhedral constraints in [13]
- a partially separable objective function with separability in the selective constraints
- a reduction in dimensions of the optimization objective (an objective with smaller number of optimizing variables).

Typically, for popular constrained queries like marginal screening, Lasso, forward stepwise etc., the optimization solves an objective in $\min(d + |E|, p)$ dimensions with d as the size of the observed data vector S, p as the dimension of regression and $|E|$ as the size of active set. The key idea behind these reductions is an upper bound to the normalizer that capitalizes on the structure of an inversion map associated with the randomized selective query and a change of measure induced by the same, discussed in Section 2.

The problem of analytically getting an approximation for the normalizer is similar to the goals of variational Bayesian approaches in [16, 11] that use a known parametric distribution to obtain an approximation to an intractable posterior based on the KL-divergence between the two posteriors. We adopt a different approach here by approximating an intractable integral, the normalizer of the truncated law as a function of the parameters in the model; we show that this approximation gives an asymptotic large deviation behavior of the exact normalizer under Gaussian randomization schemes.

These contributions allow a wider scope of applications of the truncated Bayesian approach to different generative models, randomization schemes and constrained selective queries. Such reductions become very useful in not just higher dimensions, but also, in providing inference after multiple selective queries. Below, we describe methods that demonstrate scalability of truncated Bayesian inference in both high and low dimensional regimes of inference. The effectiveness of the proposed methods is corroborated through Bayesian effect estimates with superior frequentist properties for data-mined variants in a real data set that investigates gene associations with local variants.

1.2. A motivating example

Before introducing our methods, we give an example that motivates the readers towards the inferential gains of the selective posterior over the more common in-adaptive Bayesian approach. Consider data $Y \in \mathbb{R}^n$ and a fixed predictor matrix $X \in \mathbb{R}^{n \times p}$ with columns scaled by $1/\sqrt{n}$ such that the response is generated as $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ given a $\beta \in \mathbb{R}^p$ and $\sigma^2 = 1$. An analyst decides to run Lasso on data (Y, X) in order to choose E, a set of selected predictors. Not having access to the actual generative model, he assumes the screened model from Lasso as a plausible model on his data, that is $Y \sim \mathcal{N}(X_E \beta_E, I_n)$ and a non-informative prior π on the parameters β_E in the selected model to offer Bayesian inference on β_E. Ignoring selection, he uses the unadjusted posterior on β_E

$$\pi(\beta_E | Y = y) \propto \pi(\beta_E) \cdot \exp\left(-\|y - X_E \beta_E\|_2^2 / 2\sigma^2\right) \qquad (1.2)$$

to report credible intervals and the posterior mean as inference for target β_E.

We compare the estimates from the above approach of the analyst to truncated inference post a randomized version of the Lasso query. We give inference on β_E using the same selected model and non-informative prior as the analyst where E is the output from

$$\text{minimize}_\beta \frac{1}{2}\|y - X\beta\|_2^2 - \omega^T \beta + \lambda\|\beta\|_1 + \frac{\epsilon}{2}\|\beta\|_2^2. \qquad (1.3)$$

Randomization enters the objective as $\omega^T \beta$, perturbing selection that is otherwise based only on y; the above randomized version of Lasso has been proposed in [26]. The objective has a small added ridge penalty $\epsilon = 1/\sqrt{n}$ for existence of a solution and tuning parameter is set as $\lambda = \mathbb{E}[\|X^T \psi\|_\infty]$ as proposed in [18] where $\Psi \sim \mathcal{N}(0, I)$. On a high level, our method of providing estimates in the truncated regime involves approximating the intractable posterior truncated to the realizations (y, ω) that lead to the same selection event. We finally use a Langevin walk-based sampler to provide adjusted Bayesian inference based on the approximate posterior.

To compare our methods against the traditional Bayesian inference, we conduct the below experiment with two different generative mechanisms, Model I is a frequentist model with no signal and Model II is a Bayesian model. Let $X \in \mathbb{R}^{n \times p}$, $n = 200$, $p = 1000$ be a design matrix with independent Gaussian entries normalized to have column norm 1.

- Model I: Draw in each trial $Y \sim \mathcal{N}(0, I_n)$; $n = 200$.
- Model II: Draw in each trial $\beta \sim \pi^{\text{TRUE}}$ and $Y|X, \beta$ as follows

  (1). $\beta \sim \pi^{\text{TRUE}}(\beta) = \prod_{j=1}^{1000} \pi_j^{\text{TRUE}}(\beta_j)$ with

  $$\pi_j^{\text{TRUE}}(\beta_j) = 0.90 \cdot \frac{1}{2b_1} \exp\left(-|\beta_j|/b_1\right) + 0.10 \cdot \frac{1}{2b_2} \exp\left(-|\beta_j|/b_2\right);$$

  b_1 and b_2 represent the variance parameters of the Laplace densities

in the mixture prior. In the below experiment, we set $b_1 = 0.1$ and $b_2 = 1.0$ to generate an effectively sparse vector $\beta \in \mathbb{R}^p$, $p = 1000$.

(2). $Y|X, \beta = X\beta + \epsilon$; $\epsilon \sim \mathcal{N}(0, I_n)$; $n = 200$.

In each trial, set target and model of inference based on observed E post Lasso query in (1.3) as described above; we compare estimates based on (1.2) in the untruncated regime against our method of inference that gives adjusted estimates. To conduct the randomized query (1.3), we draw ω as an instance of $\Omega \sim \mathcal{N}(0, \tau^2 I_p)$ in every trial. The below table gives a comparison of coverage of the credible intervals and risk in the frequentist model and of Bayesian FCR and Bayes risk of the posterior mean in the Bayesian model after 50 trials. The target coverage for the intervals is set at 90%. Bayesian FCR in [30] is defined as $\mathbb{E}_{\beta,Y}(V/\max(R, 1))$, where V is the number of non-covering credible intervals and $R = |E|$ is the number of intervals constructed after selection. [31] propose a Bayesian FCR for a class of priors Π; [12, 32] provide intervals that control the Bayes FCR at a desired level α uniformly over a proposed class of priors. Consistent with coverage, we report the proportion of $|E|$ intervals covering the target in the Bayesian model in Table 2 and call it CR. Unlike the non-randomized intervals in [13] that are known to grow very wide, the power inherited from randomization is reflected in shorter lengths of the adjusted intervals. The results clearly highlight the superior frequentist properties of our methods, both in terms of coverage of credible intervals and risk of posterior mean.

TABLE 1
Model I: Coverage, Risk and length of intervals

Method	Coverage	Risk	Lengths
Truncated inference	89.70%	1.81	4.41
Unadjusted inference	51.38%	3.38	3.34

TABLE 2
Model II: Bayesian CR, Bayes risk and length of intervals

Method	CR	Bayes risk	Lengths
Truncated inference	90.99%	1.49	4.49
Unadjusted inference	34.86%	4.28	3.34

1.3. Organization of paper

Section 2 outlines the truncated framework, giving the recipe for adjusted Bayesian inference using a selective posterior. Section 3 lays out the backbone of the paper, the approximating optimization problem that we solve to sample from a tractable version of the selective posterior and provides a sampler that targets the approximate posterior. Section 4 shows asymptotic validity of the finite sample bounds in Section 3 to the otherwise unavailable normalizer for non-local sequences of parameter. Section 5 lays out the optimization-based approach for popular selection queries. Section 6 includes simulations that demonstrate the inferential gains associated with the truncated Bayesian methods in the current work over the unadjusted analog. Section 6 concludes with an application of

our methods to provide adjusted Bayesian effect size estimates for local genetic variants (GTEx gene association data set) data-mined as the strongest effects.

2. A formal background

2.1. A randomized query and an inversion map

Selection events can be broadly viewed as outputs from queries on a data-base. In the context of variable selection, we are typically interested in the active set of coefficients obtained upon solving convex optimization problems. As a follow-up on the recent work on randomized inference in [25, 26], these queries are randomized versions of learning problems with a convex loss $\ell(S(X, y), .)$ and a convex penalty $\mathcal{P}_\lambda(.)$ with tuning parameter λ. Sharing similarity with the concept of reusable hold-out introduced in the field of differential privacy [8], these forms of randomized inference come with the merit of higher statistical power during inference. For the Bayesian problem, randomization results in empirical improvements in the frequentist properties associated with the selective posterior, see [20] for examples illustrating robustness of the randomized-credible intervals. The empirical results in Section 6 of the current work corroborate these merits of a randomized Bayesian procedure, reflected in the coverage properties and shorter lengths of intervals. To add to these advantages, we leverage randomization to obtain significant computational reductions in solving an approximating optimization to sample from the selective posterior in the current work.

A randomized selective query taking a convex loss $\ell(S(X, y), .)$ and a convex penalty $\mathcal{P}_\lambda(.)$ as inputs, assumes the canonical form

$$\hat{\beta}(s, \omega) = \underset{\beta}{\operatorname{argmin}} \, \ell(s; \beta) + \mathcal{P}_\lambda(\beta) - \omega^T \beta \qquad (2.1)$$

with data realization s and randomization instance ω, where $S \sim F$ independent of randomization $\Omega \sim G$. The above algorithm has a linear term in randomization ω, drawn from a distribution G with a density g, fully supported on $\mathbb{R}^p$. This can be viewed as selection with a perturbed version of data, hence the term "randomized" program. Queries of the above form are termed as objective perturbation in the privacy literature, see [4, 5]. Some randomized programs like the Lasso have an additional ℓ_2 penalty term $\frac{\epsilon}{2}\|\beta\|_2^2$ as in ridge regression in [33] to enforce existence of a solution. The analyst has access to the output E, a function of (s, ω) from such a query in the inferential stage, typically the set of active coefficients along with their signs: see [24, 27, 13]. Selective inference seeks to overcome bias from having known the output of query prior to inference through the conditional approach.

[9] presents a more natural analog of classical data-splitting in the form of data-carving, which advocates a random split of the data for selection, but allows the analyst to use the entire data for inference. A data-carved query that is performed on a randomly chosen split of the data is given by

$$\hat{\beta}(s^{(1)}, \omega) = \underset{\beta}{\operatorname{argmin}} \, \frac{1}{r}\ell(s^{(1)}; \beta) + \mathcal{P}_\lambda(\beta). \qquad (2.2)$$

with r as the fraction of data-samples used in selection and $s^{(1)}$ as a random split of the data-vector s. [15] shows that the above selection can be cast as a randomized query of the form (2.1). This can allow an analyst to collect new data and view prior selection on existing data as a split on the updated database. Hence, it can facilitate valid inference post already conducted exploratory analyses on existing data-bases while having extra power in comparison to an analysis on the new data set only. We discuss the data-carved version of Bayesian inference in more details later in Section 5.

The starting point of achieving computational reductions in approximating the selective posterior is an inversion map that characterizes the output from randomized queries. Such a map is obtained from the subgradient equation of (2.1). The canonical selection event of observing an active set of coefficients E with signs z_E can be described in terms of the randomization ω and data instance s using the inversion map. Denoting by $\hat{\beta} = (\hat{\beta}_E, 0)$ the solution from a query in (2.1) with $\hat{\beta}_E$ as the active non-zero coefficients, the inversion map is given by

$$\omega = \partial\ell(s; (\hat{\beta}_E, 0)) + \partial\mathcal{P}_\lambda((\hat{\beta}_E, 0)). \tag{2.3}$$

The above equation maps the randomization instance to realized data $S = s$ and optimization variables $O = (\hat{\beta}_E, v_{-E})$ where $\hat{\beta}_E$ denotes the active coefficients and v_{-E} represents the inactive sub-gradient corresponding to the inactive coordinates of $\partial\mathcal{P}_\lambda((\hat{\beta}_E, 0))$. We denote the optimization variables corresponding to the active coordinates as O_E and the ones corresponding to the inactive sub-gradient variables as O_{-E} from now on, referring them as active and inactive optimization variables respectively. Post an affine randomized selection event, the canonical inversion map that is the basis of a new measure takes the form

$$\omega(s, o) = Ds + Po + q \tag{2.4}$$

with s and o representing data and optimization variables respectively and D, P, q are fixed.

The scope of randomized queries is quite broad in nature allowing for even discrete versions of randomizations like carving. In practice, the analyst may use a union of outputs $\bar{E} = E_1 \cup E_2 \cdots \cup E_k$ or the final model E_k when $E_k \subset E_{k-1} \cdots \subset E_1$ (or some reasonable combination) based on a sequence of outputs $(E_1, E_2, ..., E_k)$ from multi stagewise selective algorithms to determine a target and a generative mechanism in the inferential stage. We demonstrate the extension of our approach to multiple data queries using a combination of these model selection methods in 3.4 under Section 3.

2.2. A Bayesian inferential scheme using inversion map

The ingredients for selective Bayesian inference are the same as the usual one, a prior and a likelihood, except that we replace the usual likelihood with a truncated one. To describe our inferential framework, we assume a model $f(.|\beta^*)$, parametrized by β^* post-selection on data $S \in \mathbb{R}^d$ and fix a target denoted as

$\Theta_E(\beta^*)$. In the linear model settings with a fixed design matrix $X \in \mathbb{R}^{n \times p}$, $f(.|\beta^*)$ might correspond to a family of models $\mathcal{F}_{\beta^*} = \{\mathcal{N}(X^*\beta^*, \sigma^2 I) : \beta^* \in \mathbb{R}^k\}$ for a known σ with $X^* = \begin{bmatrix} X_{i_1} & \cdots X_{i_k} \end{bmatrix}$, for a set of indices $\{i_1, i_2, \cdots, i_k\} \subset \{1, 2, \cdots, p\}$. We emphasize here that we do not have an idea about β^* before we run a selection mechanism like the Lasso. There are some settings where the parameterization exists before selection and does not change, example being the saturated model of [13]. Typically, we are running Lasso in order to find something that might be an interesting parameterization.

A common target of inference post selection of an active set E is the usual population coefficient corresponding to ordinary least squares on the selected model E, that is $\Theta_E(\beta^*) = (X_E^T X_E)^{-1} X_E^T \mathbb{E}[Y|\beta^*] = (X_E^T X_E)^{-1} X_E^T X^* \beta^*$. With a random design matrix, the target of inference can be described as $\Theta_E(\beta^*) = (\mathbb{E}[X_E^T X_E])^{-1} \mathbb{E}[X_E^T Y|\beta^*]$, with an interesting family of models specified by $\{f(.|\beta^*) : \mathbb{E}_f[Y|X] = X^*\beta^*\}$.

Remark 1. *Generative models and targets: The selected model described in [9] corresponds to parametrization $\beta^* = \beta_E \in \mathbb{R}^{|E|}$ with E being the observed active set and the saturated model corresponds to a parametrization $\beta^* = \mu \in \mathbb{R}^n$. The corresponding X^*'s in the two models are X_E and the identity matrix I_n respectively. Of course, other models are possible. The analyst can allow selection to guide him to a target and a generative model, though these choices do not necessarily have to agree with the observed selected set E.*

Remark 2. *Prior on variance parameter: The variance σ in the generative likelihood can be modeled in a Bayesian paradigm by putting a joint prior on (β^*, σ). We do not delve into details of incorporating a Bayesian model on the variance in the current draft; hereafter, we stick to a fixed variance setting.*

Using a change of measure based on the inversion map in (2.4), the joint truncated density at (s, o) corresponding to a generative model f on data in $\mathbb{R}^d$ with parameters β^* and randomization density $g \in \mathbb{R}^p$ decouples as

$$h_E(s, o) \propto |J| \cdot f(s|\beta^*) \cdot g(Ds + Po + q) \tag{2.5}$$

with support $\mathcal{R} = \{(s, o) : \hat{E}(s, o) = E, z_{\hat{E}}(s, o) = z_E\} = \mathbb{R}^d \times \mathcal{R}_O$. $|J|$ is a Jacobian reflecting the change of measure, a constant for affine inversion maps as in (2.4). The support is unrestricted on data s and constrained to $\mathcal{R}_O \subset \mathbb{R}^p$, representing constraints on optimization variables, imposed by selection output.

Coming back to a Bayesian setting, the selective posterior for generative parameters β^* given data S when $\beta^* \sim \pi(\beta^*)$ is formed by appending the marginal selective density of S to the prior $\pi(.)$. The truncated marginal of S given parameters β^* is obtained by marginalizing over O in the joint density (2.5). The selective posterior is thus, given by

$$\pi_E(\beta^*|S) \propto \pi(\beta^*) \cdot \frac{f(s|\beta^*)}{\mathbb{P}((S, O) \in \mathcal{R}|\beta^*)}. \tag{2.6}$$

The above posterior is however intractable as the normalizer

$$\mathbb{P}((S, O) \in \mathcal{R}|\beta^*) = \int_{\mathcal{R}} |J| \cdot f(s|\beta^*) g(Ds + Po + q) dods$$

has no exact closed form expression. The problem reduces to computing the normalizer $\mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$; we focus on this in Section 3.

3. An approximating optimization

3.1. Approximate normalizer based on inversion map

Using the inversion map that defines the selection output from a query in (2.4), we derive an approximating optimization with a constrained objective in $d + p$ dimensions that bounds from above the log normalizer. We state below the first theorem of this paper that gives rise to an upper bound on the volume of a convex and compact selection region $\mathcal{R}$ with respect to the joint density of data and optimization variables. It involves computating the log-MGF of the augmented vector of data and optimization variables with respect to a transformed measure induced by the inversion map in (2.4).

Theorem 1. *Denoting* $\Lambda_f^*(.|\beta^*)$ *as the convex conjugate of the log-MGF* $\Lambda_f(.|\beta^*)$ *of data vector* $S \in \mathbb{R}^d$ *and* $\Lambda_g^*(.)$ *as the conjugate of the log-MGF* Λ_g *of randomization* $\Omega \in \mathbb{R}^p$, *a Chernoff upper bound to the exact selection probability* $\log \mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$ *for convex, compact* $\mathcal{R} \subset \mathbb{R}^d \times \mathbb{R}^p$ *under the canonical inversion map in* (2.4) *is given by*

$$- \inf_{s, o \in \mathcal{R}} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_g^*(Ds + Po + q) \right\} \tag{3.1}$$

We prove the above in Appendix A. While the above upper bound does hold for compact selection regions, the canonical selective constraints lead to a selection region of the form $\mathcal{R} = \mathcal{R}_S \times \mathcal{R}_O$ with $\mathcal{R}_S = \mathbb{R}^d$ and $\mathcal{R}_O$ is typically tensor of orthants and cubes; this lacks compactness. The upper bound derived in 1 can still be applied as an approximation as we can work with a sufficiently large compact and convex subset of $\mathcal{R}$ that has an almost 1-measure under prior π. A smooth version of (3.1) is seen to lead to better frequentist properties in [20] in the non-randomized settings; in the current work, we opt for (3.3) to solve a smooth objective in place of a constrained optimization.

The bound-based approximation above is given by

$$- \inf_{s \in \mathbb{R}^d, o \in \mathbb{R}^p} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_g^*(Ds + Po + q) + \chi_{\mathcal{R}_O}(o) \right\} \tag{3.2}$$

with $\chi_{\mathcal{R}_O}(.) = -\log 1_{\mathcal{R}_O}$. In particular, $\chi_{\mathcal{R}_O}(.)$ can be interpreted as a function with a uniformly 0 penalty within the selection region. Smoothing the discrete penalty $\chi_{\mathcal{R}_O}(.)$ in the bound with a barrier penalty $b_{\mathcal{R}_O}(.)$ (which imposes a continuously decaying penalty as distance from the selective boundary increases) leads to a smooth, unconstrained version of (3.1) to approximate $\log \mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$ and is given by

$$- \inf_{s \in \mathbb{R}^d, \, o \in \mathbb{R}^p} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_g^*(Ds + Po + q) + b_{\mathcal{R}_O}(o) \right\}. \tag{3.3}$$

Remark 3. *We note that both the approximations* (3.2) *and* (3.3) *approximate the true selection probability as a function of the parameter on a large deviation scale for Gaussian data and randomization, as sample size increases to infinity. Details of the asymptotic guarantees associated with the approximations are discussed in Section 4. From a purely implementation perspective, the smooth convex objective in* (3.3) *is easier to solve and yields better empirical frequentist properties than* (3.2) *for finite sample sizes.*

The gain with (3.3) in comparison to the prior work is a much easier objective function as the canonical constraints on the optimization variables simplify to sign and cube constraints as in [26] instead of the complicated affine constraints as in [13]. We can further benefit from separability and achieve more reductions from such an approximation under certain randomizations, as seen later in (3.6).

The unconstrained optimization given by (3.3) in $d + p$ dimensions can be used to approximate selection probabilities under any randomization with a log-MGF Λ_g, that is independent of the data vector. In particular, we can use the optimization for inference post data carved queries of the form (2.2). Randomization in such queries takes the form of the gradient of difference of losses

$$\omega = \partial\ell(s; (\hat{\beta}_E, 0)) - \frac{1}{r}\partial\ell(s^{(1)}; (\hat{\beta}_E, 0)).$$

and is asymptotically independent of the data vector for a Gaussian generative model and marginally an asymptotic centered Gaussian with a covariance Σ_g. Using the conjugate of the log-MGF of a Gaussian density, we obtain a tractable pseudo posterior. We illustrate inference based on the approximate selective posterior post selection on a random fraction of the data in Section 5.

3.2. Reduction in optimization

Under randomizations with a density supported on $\mathbb{R}^p$ that are independent in all p-component coordinates, we present an approximation that is based on smoothing a modified upper bound. For most common queries, it involves an optimization objective in $d + |E|$ dimensions, where $|E| \leq p$ is the size of the active set from the selective query. Note that the optimization in (3.3) involves $d + p$ optimizing variables. With the reduction in dimensions of the optimization, we make a significant improvement in scalability of our methods in high dimensional sparse problems, when $|E| \ll p$. Such a reduction is possible due to

- decoupling of randomization density under independence
- the structure of the canonical inversion map in (3.4) that allows an exact and easy calculation of the volume of the inactive selection region with respect to the density of O_{-E}.

Before proceeding further, consider a break-up of the canonical randomization map into E active and $p - |E|$ inactive coordinates. Such a decomposition takes the form

$$\omega(s, o) = Ds + Po + q = \begin{pmatrix} D_E s + P_E o_E + q_E \\ D_{-E} s + P_{-E} o_E + o_{-E} + q_{-E} \end{pmatrix} \qquad (3.4)$$

where o_E denotes the active coefficients and o_{-E} represents the inactive subgradient. The inversion map has such a structure in most commonly used queries like the Lasso, forward stepwise, thresholding etc. as we see later in Section 5. The density g under a component-wise independent randomization scheme decouples into the active and inactive coordinates as

$$g(\omega) = g_E(\omega_E) \cdot g_{-E}(\omega_{-E}) = g_E(\omega_E) \cdot \Pi_j g_{j,-E}(\omega_{j,-E}).$$

The constraints on (o_E, o_{-E}) for the canonical map are also separable and particularly, the inactive constraints are separable in each coordinate. The selection region induced by the selective constraints can thus, be denoted by

$$\mathcal{R}_O = \mathcal{R}_E \times \mathcal{R}_{-E} = \mathcal{R}_E \times \prod_j \mathcal{R}_{j,-E} \tag{3.5}$$

where $\mathcal{R}_E$ represents the active constraint region, $\mathcal{R}_{-E}$ the inactive region and $\mathcal{R}_{j,-E}$, each component inactive constraint. The below theorem uses this separability in constraints and independence to obtain an upper bound on the logarithm of the normalizer of the truncated law. It involves computing the exact probability of the inactive subgradient variables lying in the selection region $\mathcal{R}_{-E} = \prod_j \mathcal{R}_{j,-E}$ as a function of realizations of the active optimization variable o_E and data s.

Theorem 2. *Under a randomization scheme composed of p independent components $\Omega = (\Omega_1, \cdots, \Omega_p)$ and a selective query of the form* (3.4) *yielding a compact and convex selection region*

$$\mathcal{R} = \mathcal{R}_S \times \mathcal{R}_O; \ \ where \ \mathcal{R}_S \subset \mathbb{R}^d, \ \mathcal{R}_O \subset \mathbb{R}^p$$

and $\mathcal{R}_O$ takes the form (3.5)*, an upper bound for $\log \hat{\mathbb{P}}((S,O) \in \mathcal{R}|\beta^*)$ for a compact, convex selection region $\mathcal{R}$ is given by*

$$-\inf_{s \in \mathcal{R}_S, o_E \in \mathcal{R}_E} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_{g_E}^*(D_E s + P_E o_E + q_E) - \log \mathcal{B}(o_E; s) \right\} \ with$$

$$\mathcal{B}(o_E;s) = \prod_{j=1}^{p-|E|} \int_{\mathcal{R}_{j,-E}} g_{j,-E}(o_{j,-E} + D_{j,-E}s + P_{j,-E}o_E + q_{j,-E}) do_{j,-E}$$

where $D_{j,-E}$, $P_{j,-E}$ and $q_{j,-E}$ denote the j-th rows of the matrices D_{-E}, P_{-E} and j-th component of vector q_{-E} in (3.4) *respectively.*

A proof of the bound is done in Appendix A. A heuristic[1] minimax argument together with smoothing of constraints by a barrier penalty yields a reduced analog of (3.3) for canonical selective queries in the paper. An approximating optimization with a barrier penalty on the active constraints denoted as $b_{\mathcal{R}_E}(.)$ can be written as

---

[1]Exact is possible if selection region were compact; we still can apply the approximation with a large enough compact subset of the selection region with almost mass 1 under the prior.

$$\log \hat{\mathbb{P}}((S,O) \in \mathcal{R}|\beta^*) = - \inf_{s \in \mathbb{R}^d,\, o_E \in \mathbb{R}^{|E|}} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_{g_E}^*(D_E s + P_E o_E + q_E) \right.$$
$$\left. - \log \mathcal{B}(o_E; s) + b_{\mathcal{R}_E}(o_E) \right\} \qquad (3.6)$$

with $\mathcal{B}(o_E; s)$ as defined in Theorem 2.

Expression (3.6) yields an approximating optimization in $\mathbb{R}^{d+|E|}$ with a barrier function on the sign constraints of the active optimization variables in $\mathbb{R}^{|E|}$. We use the fact that the volume of the inactive selection region $\mathcal{B}(o_E; s)$ can be calculated exactly and easily as $p-|E|$ simple, univariate integrals over intervals $\mathcal{R}_{j,-E} \subset \mathbb{R}$. For example, for a centered Gaussian randomization with covariance matrix $\tau^2 I_p$ and the canonical cube constraints on the inactive subgradient variables O_{-E} taking the form $\mathcal{R}_{j,-E} = \{o_{j,-E} : |o_{j,-E}| \leq \lambda\}$, a closed form expression for the logarithm of the volume of the inactive cube region is

$$\log \mathcal{B}(o_E; s) = \sum_{j=1}^{p-|E|} \log \left\{ \Phi\left(\frac{\lambda + \alpha(o_E; s)_j}{\tau} \right) - \Phi\left(\frac{-\lambda + \alpha(o_E; s)_j}{\tau} \right) \right\}.$$

Here $\alpha(o_E; s)_j$ denotes the j-th coordinate of $\alpha(o_E; s) \in \mathbb{R}^{p-|E|}$, the Gaussian mean of $O_{j,-E}$ given $O_E = o_E, S = s$. Marginalizing over the inactive optimization variables results in a significant reduction in dimensions of optimization from the objective in (3.3).

3.3. Dual problem: low dimensional regime

While solving the pseudo selective posterior using the above optimization as a surrogate to the normalizer is scalable for high dimensional problems, when $p \gg d+|E|$, it is not very ideal in the low dimensional regime with a large sample size, when $d + |E| \gg p$. Further, the optimization in (3.6) requires knowledge of the conjugates of the log-MGFs of the densities of the data and randomization. The dual problem yields an optimization objective in $\mathbb{R}^p$ and hence, renders a scalable version of the optimization in the low dimensional paradigm. The other distinction from the optimization posed in the primal form is that the dual is based on simply the log-MGFs corresponding to the distributions of data and randomization. In the low dimensional situation or when we do not have closed forms for the conjugates of the log-MGFs of the generative model, we can solve for the dual of the optimization problem instead.

Theorem 3. *Denoting* $\Lambda_f(.|\beta^*)$ *as the log-MGF of data generative density* f *and* $\Lambda_g(.)$ *as the log-MGF of randomization* Ω*, the dual to the optimization approximating the selection probability* $\log \mathbb{P}((S,O) \in \mathcal{R}|\beta^*)$ *in* (3.3) *is given by*

$$\inf_{u \in \mathbb{R}^p} \left\{ \Lambda_f(D^T u|\beta^*) + \Lambda_g(-u) + b_{\mathcal{R}_O}^*(P^T u) + u^T q \right\} \qquad (3.7)$$

where $b_{\mathcal{R}_O}^*$ *is conjugate of the barrier function* $b_{\mathcal{R}_O}(.)$ *and* D, P, q *are coefficients of linear terms of map* (2.4).

See proof in Appendix A. A point to note is that dual formulation of the approximating optimization involves computing the conjugate of the barrier penalty function on the optimization variables. Since the constraints on the active and inactive optimization problems are separable, this involves solving conjugates of $|E|$ and $p - |E|$ univariate functions that correspond to the active and inactive constraints respectively. That is the conjugate barrier takes the additive form

$$b^*_{\mathcal{R}_O}(P^T u) = b^*_{\mathcal{R}_E}(P^T_E u) + b^*_{\mathcal{R}_{-E}}(P^T_{-E} u).$$

Details of the computation of the conjugates of the barrier functions used in our implementations are given in Appendix B.

Remark 4. *The dual of the constrained Chernoff-based optimization in* (3.2) *for the canonical constraint region*

$$\mathcal{R}_O = \prod_{j=1}^{|E|} \{o_{j,E} : diag(s_{j,E})o_{j,E} > 0\} \times \prod_{j=1}^{p-|E|} \{o_{j,-E} : |o_{j,-E}| \leq \lambda\}$$

is given by $\displaystyle \inf_{u \in \mathbb{R}^p : diag(s_E)P^T_E u < 0} \left\{ \Lambda_f(D^T u | \beta^*) + \Lambda_g(-u) + \lambda \sum_{j=1}^{p-|E|} |P^T_{j,-E} u| + \right.$
$\left. u^T q \right\}$ *where* $P^T_{j,-E}$ *denotes the j-th row of matrix transpose of* P_{-E}. *This is by observing that the convex conjugate of the characteristic function* $\chi_{\mathcal{R}_{j,E}}(.)$ *representing the sign constraints on the active optimization variables is*

$$\chi^*_{\mathcal{R}_{j,E}}(P^T_{j,E} u) = \begin{cases} 0 & \text{if } s_{j,E} P^T_{j,E} u < 0 \\ \infty & \text{otherwise.} \end{cases}$$

and that for the cube constraints $\chi_{\mathcal{R}_{j,-E}}(.)$ *on the inactive subgradients is*

$$\chi^*_{\mathcal{R}_{j,-E}}(P^T_{j,-E} u) = \lambda |P^T_{j,-E} u|.$$

3.4. Marginalizing over multiple selections

The optimization problem described above is aimed to approximate the selection probability of an event based on a single randomized data query of the form (2.4). It is however, common practice to apply stages of screening or query the data base multiple times to arrive at a selected set. An example might be laboratory A performing an initial scan of thousands of potential predictors to select a pool that passes a suitably chosen thresholding criterion and laboratory B conducting another screening of predictors. The analyst is interested in combining both screening results to guide her to inference on the same data set that has been analyzed by the two laboratories.

The approximation presented in (3.3), (3.6) and (3.7) can be marginalized over multiple randomizations from multiple stages and hence, be extended to multi-stage selective algorithms. The next Lemma renders an approximation

to the normalizer for a K-stage randomized selection with query in each stage corresponding to an inversion map

$$\omega_k = D_k s + P_k o_k + q_k \text{ for } k = 1, 2, .., K$$

with o_k being the optimization variables for the randomized program in stage k. The selection region, determined by constraints on optimization variables o_K at each stage, separable in the active and inactive coordinates as before, is given by $\mathcal{R}_{(O_1, \cdots, O_k)} = \Pi_{i=1}^K \mathcal{R}_{O_k}$. Again, denote $\mathcal{R} = \mathcal{R}_S \times \mathcal{R}_{(O_1, \cdots, O_k)}$. Typically $\mathcal{R}_S = \mathbb{R}^d$, the unconstrained data-space augmented with the constrained region on optimization variables from each query.

Lemma 1. *Under K randomizations with $\Omega_k \overset{ind}{\sim} g_k(.)$ for $k = 1, 2, ..., K$ and with $\Lambda^*_{g_k}(.)$ as the conjugate of the log-MGF Λ_{g_k} of randomization Ω_k, an upper bound to the logarithm of the exact selection probability $\log \mathbb{P}((S, O_1, \cdots, O_K) \in \mathcal{R}|\beta^*)$ for a convex, compact $\mathcal{R}$ is given by*

$$-\inf_{s \in \mathcal{R}_S, o_k \in \mathcal{R}_{O_k}, k=1,2,...,K} \left\{ \Lambda^*_f(s|\beta^*) + \sum_{k=1}^K \Lambda^*_{g_k}(D_k s + P_k o_k + q_k) \right\}.$$

Proof. The proof is easy to see as with independent randomizations in each stage of selection, $\log \mathbb{P}((S, O_1, \cdots, O_K) \in \mathcal{R}|\beta^*)$ is bounded from above by

$$-\inf_{s \in \mathcal{R}_S, o_k \in \mathcal{R}_{O_k}, k=1,2,..,K} \left\{ \alpha^T s + \sum_{k=1}^K \alpha_k^T o_k \right\} - \log \mathbb{E}(\exp(\alpha^T S + \sum_{k=1}^K \alpha_k^T O_k)|\beta^*).$$

An optimization over $\alpha \in \mathbb{R}$ and $\alpha_k \in \mathbb{R}, \{k = 1, 2, ..., K\}$ and a minimax equality gives the bound

$$-\inf_{s \in \mathcal{R}_S, o_k \in \mathcal{R}_{O_k}, k=1,,..,K} \sup_{\alpha, \alpha_k} \left\{ \alpha^T s + \sum_{k=1}^K \alpha_k^T o_k - \log \mathbb{E}(\exp(\alpha^T S + \sum_{k=1}^K \alpha_k^T O_k)|\beta^*) \right\}.$$

A similar computation of the log-MGF of the augmented vector $(S, O_1, \cdots, O_K)$ as in the proof of Theorem 1 based on the change of variables facilitated by the inversion maps in (3.4) completes the proof. □

The smooth analog of the constrained optimization in Lemma 1 is given by

$$-\inf_{s \in \mathbb{R}^d, o_k \in \mathcal{R}_{O_k}, k=1,2,...,K} \left\{ \Lambda^*_f(s|\beta^*) + \sum_{k=1}^K \Lambda^*_{g_k}(D_k s + P_k o_k + q_k) + b_{\mathcal{R}_{O_k}}(o_k) \right\}.$$

The dual formulation of this approximation, optimizing over dual variables $u_k; \ k = 1, 2, \cdots, K$ is given by

$$\inf_{u_1, .., u_K} \Lambda_f \left(\sum_{k=1}^K D_k^T u_k \right) + \sum_{k=1}^K \left\{ \Lambda_g(-u_k) + b^*_{\mathcal{R}_{O_k}}(P_k^T u_k) + u_k^T q_k \right\}. \qquad (3.8)$$

Remark 5. Cost of optimization: *The optimization in Lemma 1, decomposed into active and separable inactive problems can be solved in its primal form in effectively $d + \sum_{k=1}^{K} \cdot |E_k|$ dimensions, while the dual has an effective cost of solving a Kp dimensional optimization, if selected sizes are of smaller order than p.*

3.5. Sampler: Langevin random walk

We describe below a Langevin random walk to sample from the pseudo posterior

$$\tilde{\pi}_E(\beta^* | S = s) \propto \pi(\beta^*) \cdot f(s|\beta^*) / \hat{\mathbb{P}}((S, O) \in \mathcal{R} | \beta^*)$$

post randomized queries based on the approximate normalizer in (3.3), (3.6) and (3.7). The method of approximating a target distribution using a Langevin diffusion is studied in [23]. Depending on the regime of inference, we require the log-MGFs of the generative density and the randomization density for solving the approximating optimization in its dual form or the convex conjugates of the log-MGFs while solving for the primal. A new update $\beta^{*(K)}$ based on a Langevin random walk with target as the pseudo selective posterior $\tilde{\pi}_E(\beta^* | S)$ is given by

$$\beta^{*(K)} = \beta^{*(K-1)} + \eta \nabla \log \tilde{\pi}_E \left(\beta^{*(K-1)} | S \right) + \sqrt{2\eta} \epsilon^{(K)} \tag{3.9}$$

where η is the step-size and $\epsilon^{(K)} \sim \mathcal{N}(0, I)$. This allows us to provide sample-based effect size estimates in the form of credible intervals and point estimates for any function of the parameter of interest β^* in the generative model.

All that the sampler in (3.9) requires is calculating the gradient of the log-posterior $\tilde{\pi}_E$ as a function of each new draw $\beta^{*(K)}$. For a Gaussian generative model on data vector S with mean parametrized as $\mu(\beta^*)$, the below theorem shows that the gradient of the log-pseudo posterior can be computed in terms of the optimizer to the problem in (3.3).

Theorem 4. *The gradient of the log-pseudo selective posterior $\log \tilde{\pi}_E(. | S = s)$ at $\beta^{*(K)}$ for a Gaussian generative density for data vector S with mean $\mu(\beta^*) : \mathbb{R}^k \to \mathbb{R}^d$ and a variance-covariance matrix Σ_f given by*

$$f(s|\beta^*) = \frac{1}{(2\pi)^{d/2} |\Sigma_f|^{1/2}} \cdot \exp\left(-(s - \mu(\beta^*))^T \Sigma_f^{-1} (s - \mu(\beta^*))/2 \right)$$

with respect to parameter β^* is given by

$$\frac{\partial \log \pi(\beta^*)}{\partial \beta^*} \Big|_{\beta^{*(K)}} + \left(\frac{\partial \mu}{\partial \beta^*} \right)^T \Big|_{\beta^{*(K)}} \cdot \Sigma_f^{-1} \left\{ s - s^* \left(\Sigma_f^{-1} \mu \left(\beta^{*(K)} \right) \right) \right\} \tag{3.10}$$

where $s^*(\Sigma_f^{-1} \mu(\beta^*))$ equals

$$\arg \min_{z \in \mathbb{R}^d} \left(z^T \Sigma_f^{-1} \mu(\beta^*) - \frac{1}{2} z^T \Sigma_f^{-1} z - \inf_{o \in \mathbb{R}^p} \left\{ \Lambda_g^*(Dz + Po + q) + b_{\mathcal{R}_O}(o) \right\} \right).$$

For the dual optimization in (3.7) , the optimizer s^* can be derived from the K.K.T. conditions as $s^* = \nabla \Lambda_f(D^T u^* | \beta^*)$ where u^* is the dual variable that optimizes

$$\arg \min_{u \in \mathbb{R}^p} \left\{ \Lambda_f(D^T u | \beta^*) + \Lambda_g(-u) + b_{\mathcal{R}_O}^*(P^T u) + u^T q \right\}$$

This shows that all we need for inference is the solution to the optimization problem cast as (3.3), (3.6) and (3.7) at each fresh draw. See Appendix A for a proof of Theorem 4.

Remark 6. *Estimating equation for MAP: It is easy to see that equating* (3.10) *to 0 gives rise to an estimating equation for the selective MAP for β^*. It gives rise to a convex objective for the MAP problem for any log-concave prior π on β^* and a generative mean $\mu(.)$ that is linear in β^*. Lemma 2 gives the selective MLE under a non-informative prior $\pi \propto 1$ for a Gaussian density with natural parameter as β^*. A standard gradient descent can be performed on the log-posterior to solve for the MAP in such cases. The pseudo selective MAP in the non-randomized scenario and the simple additive randomized settings is introduced in [20].*

Lemma 2. *Under a Gaussian generative density for data vector S considered in 4 with mean parametrized as $\mu(\beta^*) = \beta^*$ and $\Sigma_f = I$, the approximate selective MLE β^{*MLE} based on the pseudo truncated law $\tilde{\ell}_E(\cdot | \beta^*)$ given by*

$$\log \tilde{\ell}_E(s | \beta^*) \propto -s^T s/2 + \beta^{*T} s - \Gamma(\beta^*) \text{ and}$$

$$\Gamma(\beta^*) = \sup_{z \in \mathbb{R}^d} \left\{ z^T \beta^* - \frac{1}{2} z^T z - \inf_{o \in \mathbb{R}^p} \left\{ \Lambda_g^*(Dz + Po + q) + b_{\mathcal{R}_O}(o) \right\} \right\}$$

satisfies $\nabla \Gamma \left(\beta^{*MLE} \right) = s$.

The proof of this is straight-forward from the estimating equation in (3.10). In the following section, we show that the approximate normalizer in Theorems 1 and 2 give a valid exponential rate to the selection probability on a large deviation scale under a Gaussian randomization and a Gaussian generative density. Under these conditions, the selective MLE obtained by maximizing the pseudo truncated law in Lemma 2 is consistent for β^*.

4. Limiting approximation on large deviation scale

In the implementations in Section 6, the columns of the predictor matrix X are normalized by a factor of $1/\sqrt{n}$. We introduce the suppressed scale and denote the normalized X as $X/\sqrt{n}$ in this section. The optimization for such a query is described in details in 5.1 under Section 5. The set-up is similar to the randomized logistic lasso query considered in [25], except that we use the Lasso query instead of the logistic Lasso query to illustrate the results in this section.

With data $y_i, X_i \sim \mathbb{P}_n(\beta_{E,n})$ for $i = 1, 2, ..., n$ where $\mathbb{P}_n \in \{\mathbb{F}_n : \mathbb{E}_{\mathbb{F}_n}[y_i|X_i] = X_{i,E}\beta_{E,n}, \mathbb{E}_{\mathbb{F}_n}[y_i^2|x_i] - \mathbb{E}_{\mathbb{F}_n}^2[y_i|X_i] = 1\}$, the Lasso query is given by

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta/\sqrt{n}\|_2^2 - \omega^T\beta + \lambda\|\beta\|_1 + \frac{\epsilon}{2}\|\beta\|_2^2.$$

The tuning parameter is set at a theoretical value $\lambda_n = \mathbb{E}\left[X^T\psi/\sqrt{n}\right]$ for $\Psi \in \mathbb{R}^p \sim \mathcal{N}(0, I_p)$ and $\epsilon = 1/\sqrt{n}$. Denote the scaled versions of the data vector and optimization variables as S_n where

$$S_n = \sqrt{n}\left(\begin{array}{c} \left(X_E^T X_E/n\right)^{-1}\dfrac{X_E^T y}{n} \\ \dfrac{X_{-E}^T}{n}\left(y - X_E\left(X_E^T X_E/n\right)^{-1}\dfrac{X_E^T y}{n}\right) \end{array}\right) = \sqrt{n}\bar{S}_n$$

with $\bar{S}_n$ as the mean of data variables $S_{i,n}(y_i, X_i)$, $i = 1, 2, \cdots, n$ such that

$$\mathbb{E}_{\mathbb{P}_n}[S_{i,n}(y_i, X_i)|\beta_E] = \mu_n(\beta_{E,n}) = \left(\begin{array}{c}\beta_{E,n} \\ \mathbb{E}_{\mathbb{P}_n}\left[X_{i,-E}^T(y_i - X_{i,E}\beta_{E,n})\right]\end{array}\right). \quad (4.1)$$

Unlike [25] which assumes local alternatives of the form $\beta_{E,n} = o(n^{-1/2})$, the selection probability is on the scale of a large deviation probability if $\beta_{E,n} = O(1)$. To simplify notations, we denote $\beta_{E,n} = \beta_E$ hereafter.

We assume that the randomization instance ω in (4) is from a Gaussian density, which is used in all the experiments in Section 6. The infinite divisibility property of Gaussian densities allows us to write perturbation $\omega = \sqrt{n}\bar{\omega}_n$ where $\bar{\omega}_n$ is the mean of n i.i.d. Gaussian variables $\omega_i, i = 1, 2, \cdots, n$. The tuning parameter λ_n converges to a constant; thus, we can treat it as a constant and denote it as λ. Also, noting that $X^T X/n$ converges in probability to a constant, we can consider the matrices D_n, P_n, q_n in the inversion map for (4) as fixed. We use notations D, P, q for the affine inversion map. Finally, let

$$\sqrt{n}\bar{O}_n = P^{-1}(\sqrt{n}\bar{\omega}_n - \sqrt{n}D\bar{S}_n - q) \quad (4.2)$$

based on the inversion map, with $\bar{O}_n$ interpreted as the mean of

$$O_{i,n} = P^{-1}(\omega_i - D\bar{S}_n - q/\sqrt{n}), \ i = 1, 2, \cdots, n.$$

Theorem 5 gives the limiting rate of decay of the volume of a compact and convex selection region $\mathcal{R} = \mathcal{R}'_S \times \mathcal{R}'_O$ for $\mathcal{R}'_S \subset \mathbb{R}^p$ and $\mathcal{R}'_O \subset \mathbb{R}^p$ with respect to the probability density of the augmented vector $(\bar{S}_n, \bar{O}_n)$, whenever the data vector mean satisfies a large deviation principle. Define

$$\Lambda_f(\lambda|\beta_E) = \lim_n \frac{1}{n}\Lambda_{\mathbb{P}_n}(n\lambda) = \lim_n \frac{1}{n}\log\mathbb{E}_{\mathbb{P}_n}[\exp(n\lambda^T\bar{S}_n)|\beta_E] \quad (4.3)$$

with $\bar{S}_n$ as the mean of the data vector array $S_{i,n}$, $i = 1, 2, \cdots n$ satisfying (4.1) for $\beta_E \in \mathbb{R}^{|E|}$ and $\mathcal{D} = \{\lambda \in \mathbb{R}^p : \Lambda_f(\lambda|\beta_E) < \infty\}$.

Theorem 5. *Whenever the limiting log-MGF sequence $\Lambda_f(\lambda|\beta_E) < \infty$ in a neighborhood around 0 in $\mathbb{R}^p$, $\Lambda_f(\lambda|\beta_E)$ is lower semi-continuous and differentiable in $\mathcal{D}^0$ and for any $\lambda \in \partial\mathcal{D}$, $\lim_{\gamma \to \lambda} |\nabla\Lambda_f(\nu|\beta_E)| = \infty$, the following hold for a compact and convex selection region $\mathcal{R} = \mathcal{R}'_S \times \mathcal{R}'_O$.*

(1). Denoting the log-MGF of Gaussian randomization ω_1 as $\Lambda_g(.)$ with conjugate $\Lambda_g^*(.)$ and the conjugate corresponding to $\Lambda_f(.|\beta_E)$ in (4.3) as $\Lambda_f^*(.|\beta_E)$

$$\lim_n \frac{1}{n} \log \mathbb{P}(\bar{S}_n \in \mathcal{R}'_S, \bar{O}_n \in \mathcal{R}'_O|\beta_E) + \inf_{s \in \mathcal{R}'_S, o \in \mathcal{R}'_O} \{\Lambda_f^*(s|\beta_E)$$
$$+ \Lambda_g^*(Ds + Po + q/\sqrt{n})\} = 0.$$

(2). If the Gaussian randomization density supported on $\mathbb{R}^p$ is independent in all p coordinates with the conjugate of the log-MGF corresponding to active coordinates denoted as $\Lambda_{g_E}^*$ and the selective constraints on the optimization variables are separable as in Theorem 2, then

$$\lim_n \frac{1}{n} \log \mathbb{P}(\bar{S}_n \in \mathcal{R}'_S, \bar{O}_n \in \mathcal{R}'_O|\beta_E) + \inf_{s \in \mathcal{R}'_S, o_E \in \mathcal{R}'_E} \left\{\Lambda_f^*(s|\beta_E) + \right.$$
$$\left. \Lambda_{g_E}^*(D_E s + P_E o_E + q_E/\sqrt{n}) - \frac{1}{n} \log \mathcal{B}(o_E; s)\right\} = 0 \ with$$
$$\mathcal{B}(o_E; s) = \prod_{j=1}^{p-|E|} \int_{\mathcal{R}'_{j,-E}} g_{j,-E}(o_{j,-E} + D_{j,-E}s + P_{j,-E}o_E + q_{j,-E}/\sqrt{n})do_{j,-E}.$$

We use in the above theorem the fact that the mean vector $\bar{S}_n$ satisfies a large deviation principle with rate function Λ_f^*

$$\lim_n \frac{1}{n} \log \mathbb{P}(\bar{S}_n \in \mathcal{R}'_S|\beta_E) = -\inf_{s \in \mathcal{R}'_S} \Lambda_f^*(s).$$

Similarly, the conditional probability of $\bar{O}_n$ given $\bar{S}_n$ has a limiting large deviation rate expressed in terms of $\Lambda_g^*(h_n(.))$ composed with the affine inversion map $h_n(.) : \mathbb{R}^p \to \mathbb{R}^p$ given by $h_n(o) = D\bar{S}_n + Po + q/\sqrt{n}$. That is,

$$\lim_n \frac{1}{n} \log \mathbb{P}(\bar{O}_n \in \mathcal{R}'_O|\bar{S}_n = s) + \inf_{o \in \mathcal{R}'_O} \Lambda_g^*(Ds + Po + q/\sqrt{n}) = 0. \qquad (4.4)$$

This is a consequence of the observation that the limiting rate function is the conjugate of $\lim_n \frac{1}{n} \log \mathbb{E}(\exp(n\lambda^T \bar{O}_n)|\bar{S}_n = s)$; the change of measure map yields the following

$$\lim_n \frac{1}{n} \log \mathbb{E}(\exp(n\lambda^T \bar{O}_n)|\bar{S}_n = s) + \lambda^T P^{-1}(Ds + q/\sqrt{n}) - \frac{\lambda^T P^{-1} \Sigma_g P^{-1^T} \lambda}{2} = 0$$

for Gaussian randomization with variance Σ_g. The proof then follows by an application of the below Lemma 3, a modified version of Varadhan's Lemma

(see [7]). The smooth unconstrained optimizations in (3.3) and (3.6) with a continuous barrier penalty function, scaled appropriately also approximate the selection probability accurately as the sample size grows large. Proofs of the above theorem and Lemma 3 are included in Appendix A.

Lemma 3. *For a sequence of functions $H_n(.)$ that uniformly converge to a continuous function H on a compact, convex set $\mathcal{R} \subset \mathbb{R}^d$, the limit*

$$\lim_n \frac{1}{n} \log \mathbb{E}[\exp(nH_n(\bar{Z}_n))1_{\bar{Z}_n \in \mathcal{R}}] = -\inf_{z \in \mathcal{R}}\{\Lambda^*(z) - H(z)\}$$

holds for sequence of variables $\bar{Z}_n \in \mathbb{R}^d$ satisfying a large deviation principle with a rate function $\Lambda^*(.)$.

Remark 7. *Approximation error in file drawer example*: *Below, we use the approximation in Theorem 5 and quantify the approximation error in the simplified example of reporting the mean using data $Y_i \sim \mathcal{N}(\beta, 1)$, $i = 1, 2, \cdots, n$ only when it exceeds a threshold. A randomized version of selection is given by*

$$\sqrt{n}\bar{Y}_n + \omega \geq \lambda; \ \omega \sim \mathcal{N}(0, 1), \lambda > 0.$$

The exact selection probability in this example is $\bar{\Phi}(\{\lambda - \sqrt{n}\beta\}/\sqrt{2})$. The Chernoff-based approximate log-normalizer using inversion map: $\omega = -\sqrt{n}\bar{Y}_n + \sqrt{n}o + \lambda$ is given by

$$-\inf_{s,o>0} n(s-\beta)^2/2 + n(o - s + \lambda/\sqrt{n})^2/2 = -\inf_{o>0} n(o + \lambda/\sqrt{n} - \beta)^2/4.$$

For $\beta < \delta < 0$ (the scenario, when selection is a rare event), the logarithm of the approximate normalizer is $n(\lambda/\sqrt{n} - \beta)^2/4$ for large n. Using the Mill's ratio approximation for the Gaussian tail, the approximation error between the logarithms of the exact normalizer and the above computed approximate normalizer, scaled by $1/n$ is seen to be bounded by

$$\frac{1}{n}|\log \bar{\Phi}(\{\lambda - \sqrt{n}\beta\}/\sqrt{2}) + \inf_{o>0} n(o + \lambda/\sqrt{n} - \beta)^2/4|$$

$$\leq \frac{1}{n}\left|\log \bar{\Phi}(\{\lambda - \sqrt{n}\beta\}/\sqrt{2}) + n(\lambda/\sqrt{n} - \beta)^2/4 + \log|\{\lambda - \sqrt{n}\beta\}/\sqrt{2}|\right|$$

$$+ \frac{1}{n}\left| - n(\lambda/\sqrt{n} - \beta)^2/4 - \log|\{\lambda - \sqrt{n}\beta\}/\sqrt{2}| + n(\lambda/\sqrt{n} - \beta)^2/4\right|$$

$$\leq \frac{1}{n}\log|\{\lambda - \sqrt{n}\beta\}/\sqrt{2}|^3 + \frac{2}{n}\log|\{\lambda - \sqrt{n}\beta\}/\sqrt{2}|.$$

Figure 1 depicts the exact and approximate log-normalizers (scaled by $1/n$) for the selective posterior under a diffuse prior $\pi(\beta) \propto 1$, as β varies with threshold for the randomized mean set at $\lambda = 2$ and sample size $n = 100$.

Under a Gaussian generative density parametrized by β_E, a Gaussian randomization with log-MGF $\Lambda_g(.)$ and a compact and convex selection region
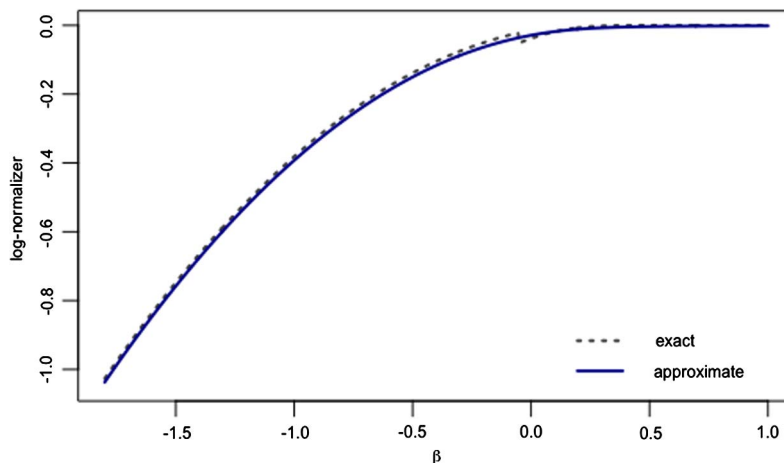
FIG 1. *Logarithms of normalizer in selective posterior in file drawer example*

$\mathcal{R} = \mathcal{R}'_S \times \mathcal{R}'_O$ and the same asymptotic set-up as in Theorem 5, it follows as a consequence that the sequence

$$\Gamma_n(\beta_E) = \sup_{s \in \mathcal{R}'_S} \left\{ s\beta_E - s^T s/2 - \inf_{o \in \mathcal{R}'_O} \left\{ \Lambda^*_g(Ds + Po + q/\sqrt{n}) + b_{\mathcal{R}_O}(o) \right) \right\} \right\}$$

in Lemma 2 approximates the exact log-partition function

$$\Gamma_{\text{exact},n}(\beta_E) = \beta_E^T \beta_E/2 + \frac{1}{n} \log \mathbb{P}(\bar{S}_n \in \mathcal{R}'_S, \bar{O}_n \in \mathcal{R}'_O | \beta_E) \text{ as}$$

$$\Gamma_n(\beta_E) - \Gamma_{\text{exact},n}(\beta_E) \to 0, \; n \to \infty.$$

Denote the sequence of selective MLE obtained by maximizing the sequence of pseudo truncated likelihoods as $\beta_{n,E}^{\text{MLE}}$ that satisfies the estimating equation $\nabla \Gamma \left(\beta_{n,E}^{\text{MLE}} \right) = \bar{S}_n$. Strong convexity of $\Gamma_n(\cdot)$ with a lower bound B on the indices of convexity leads to the identity

$$\|\beta_{n,E}^{\text{MLE}} - \beta_E\|_2 \leq \frac{1}{B} \cdot \|\bar{S}_n - \nabla \Gamma_n(\beta_E)\|_2.$$

Convergence of the approximate log-partition sequence $\Gamma_n(\cdot)$ to the exact one coupled with the identity above prove consistency of the selective MLE $\beta_{n,E}^{\text{MLE}}$ using similar arguments as Theorem 7.6 in [20].

5. Illustrations of truncated Bayesian approach

We illustrate truncated Bayesian approach by revisiting some popular selective queries. These examples are discussed in the context of frequentist inference in [26]. In all the below examples, the generative law on the data vector

is a Gaussian with mean parametrized as β^*. In particular, we assume that $\mathbb{E}_f[Y|X] = X^*\beta^*$. We assume that the columns of the design matrix X are scaled by $1/\sqrt{n}$ and denote the scaled predictor matrix as X, suppressing the scale n. In particular, we assume independent Gaussian entries for X. The randomized queries are conducted using instances of randomization from a Gaussian density supported on $\mathbb{R}^p$ with mean 0 and variance $\tau^2 I_p$.

Remark 8. *Prior information on parameters: We provide inferential results based on the selected model in Section 6 where $X^* = X_E$ and $\beta^* = \beta_E \in \mathbb{R}^{|E|}$ under a non-informative prior. Our methods however, do allow the analyst to elicit a prior from an expert or prior experiments post the selective analysis. Our simulation results show that in the absence of an informative prior, the analyst can still capitalize upon the merits of a Bayesian machinery to provide valid inference post selection.*

The generic recipe for inference using the proposed methods is to compute the inversions maps and selection regions that characterize the output of a query. This is followed by solving the optimization problem for each draw $\beta^{(K)}$ of the sampler. A function of the optimal data vector gives the gradient of the approximate log-posterior at $\beta^{(K)}$ in Theorem 4; thus, we sample from a tractable version of the selective posterior to carry out Bayesian inference. For each of the below examples, we give an explicit approximating optimization based on the inversion map and selective constraints that characterize the randomized query.

5.1. A Lasso query

A randomized version of Lasso with design X based on data vector $S = y \in \mathbb{R}^n$ and randomization instance $\omega \sim \mathcal{N}(0, \tau^2 I_p)$ solves

$$\operatorname{argmin}_\beta \frac{1}{2}\|y - X\beta\|_2^2 - \omega^T\beta + \lambda\|\beta\|_1 + \frac{\epsilon}{2}\|\beta\|_2^2;$$

to give output (E, z_E), the active set with active signs.

The selection region imposed by the ℓ_1-constrained query takes the completely separable form of orthants for active constraints and intervals for inactive constraints; that is $\mathcal{R} = \mathbb{R}^d \times \mathcal{R}_O$ where

$$\mathcal{R}_O = \prod_{j=1}^{|E|} \{o_{j,E} : \operatorname{sign}(o_{j,E}) = z_{j,E}\} \times \prod_{j=1}^{p-|E|} \{o_{j,-E} : |o_{j,-E}| \leq \lambda\}.$$

Inversion map: The inversion map encoding selection output (E, z_E) is

$$\omega(y; o) = -\begin{pmatrix} X_E^T \\ X_{-E}^T \end{pmatrix} y + \begin{bmatrix} X_E^T X_E + \epsilon I & 0 \\ X_{-E}^T X_E & I \end{bmatrix} o + \begin{pmatrix} \lambda z_E \\ 0 \end{pmatrix} = Dy + Po + q$$

$$= \begin{pmatrix} -X_E^T y + (X_E^T X_E + \epsilon I)o_E + \lambda z_E \\ X_{-E}^T y + X_{-E}^T X_E o_E + o_{-E} \end{pmatrix} = \begin{pmatrix} D_E y + P_E o_E + q_E \\ D_{-E} y + P_{-E} o_E + q_{-E} + o_{-E} \end{pmatrix}.$$

Based on the above inversion map, the approximating optimization with $n + |E|$ and p optimizing variables in the primal and dual formulation respectively can be computed as below.

Primal problem: Under the linear model with mean $X^* \beta^*$ and covariance matrix $\sigma^2 I_n$, the approximation to $\log \hat{\mathbb{P}}((S, O) \in \mathcal{R} | \beta^*)$ leading to pseudo posterior $\tilde{\pi}_E(\beta^* | y)$ is

$$- \inf_{s \in \mathbb{R}^n, \, o_E \in \mathbb{R}^{|E|}} \left\{ \frac{\|s - X^* \beta^*\|_2^2}{2\sigma^2} + \frac{\|D_E s + P_E o_E + q_E\|_2^2}{2\tau^2} - \log \mathcal{B}(o_E; s) + b_{\mathcal{R}_E}(o_E) \right\}$$

where the volume of the inactive selection region conditional on $S = s, O_E = o_E$ under the isotropic Gaussian randomization is computed as

$$\log \mathcal{B}(o_E; s) = \sum_{j=1}^{p-|E|} \log \left\{ \Phi \left(\frac{\lambda + \alpha(o_E; s)_j}{\tau} \right) - \Phi \left(\frac{-\lambda + \alpha(o_E; s)_j}{\tau} \right) \right\}$$

with $\alpha(o_E; s) = D_{-E} s + P_{-E} o_E + q_{-E}$ and D_{-E}, P_{-E}, q_{-E} as given in (5.1).

Dual problem: The corresponding dual in terms of the logarithm of the Gaussian MGFs (both for data and randomization) and conjugate of the barrier function is given by

$$\inf_{u \in \mathbb{R}^p} \frac{1}{2} \sigma^2 \|D^T u\|_2^2 + \frac{1}{2} \tau^2 \|u\|_2^2 + b_{\mathcal{R}_O}^*(P^T u) + u^T (D X^* \beta^* + q),$$

where D, P, q are obtained from the map in (5.1). Calculation of the conjugate of the barrier function follows from Appendix B.

A carved query solves the randomized version of lasso described in (5.1) on a random split of the data $S(X^{(1)}, y^{(1)})$ leading to the output (E, z_E). Such a query takes the form $\operatorname{argmin}_\beta \frac{1}{2r} \|y^{(1)} - X^{(1)} \beta\|_2^2 + \lambda \|\beta\|_1 + \frac{\epsilon}{2} \|\beta\|_2^2$ where r is the fraction of the data used in the above selective query.

Inversion map: The randomization inherited from the random split on the data, as described in [15] leads to the below inversion map

$$\omega(s, o) = \partial \ell(s; (\hat{\beta}_E, 0)) - \frac{1}{r} \partial \ell(s^{(1)}; (\hat{\beta}_E, 0)).$$

The randomization described as above is asymptotically Gaussian with mean 0 and covariance Σ_g and is asymptotically independent of the data vector S described for the random X lasso query.

Primal problem: While we can no longer use the reduced optimization in (3.6) (as the randomization inherited from the split is not independent in all p-coordinates), we can use the more general approximation to the normalizer in (3.3). The joint on the data and randomization is an asymptotic Gaussian, with

the data mean parametrized as $\mu(\beta^*)$. The approximating optimization that we solve to sample from the pseudo *selective posterior* is given by

$$-\inf_{s\in\mathbb{R}^p,\, o\in\mathbb{R}^p}\left\{\frac{1}{2}(s-\mu(\beta^*))^T\Sigma_f^{-1}(s-\mu(\beta^*))+\frac{1}{2}(Ds+Po+q)^T\Sigma_g^{-1}(Ds+Po+q)\right.$$
$$\left.+b_{\mathcal{R}_O}(o)\right\}$$

with Σ_g and Σ_f estimated by bootstrap.

5.2. A forward stepwise query

We discuss the approximating optimization that we solve to give truncated Bayesian inference after 2 steps of forward stepwise selection (FS) next. This can be easily generalized to K steps. In Section 6, we give adjusted estimates in a Bayesian model after 1 step of FS. This can also be viewed as a sequential query on the data.

Inversion maps: Denoting $E_1=\{j_1\}$ and $E_2=\{j_1,j_2\}$ and the predictor for second stage as $\tilde{X}=\mathcal{P}_{j_1}^{\perp}X_{-j_1}$ (adjusted for selection of j_1 in the first step), the characterizing inversion maps for the two-stage sequential selection procedure are as below.

$$\omega_1=\begin{pmatrix}-X_{j_1}^T s+o_{j_1}\\-X_{-j_1}^T s+o_{-j_1}\end{pmatrix};\omega_2=\begin{pmatrix}-\tilde{X}_{j_2}^T s+o_{j_2}\\-\tilde{X}_{-j_2}^T s+o_{-\{j_1,j_2\}}\end{pmatrix}.$$

giving selection regions

$$\mathcal{R}_{O_1}=\{(o_{j_1},o_{-j_1})\in\mathbb{R}^p:z_{j_1}o_{j_1}\geq 0,\|o_{-j_1}\|_{\infty}\leq|o_{j_1}|\},$$
$$\mathcal{R}_{O_2}=\{(o_{j_2},o_{-\{j_1,j_2\}})\in\mathbb{R}^{p-1}:z_{j_2}o_{j_2}\geq 0,\|o_{-\{j_1,j_2\}}\|_{\infty}\leq|o_{j_2}|\}$$

where z_{j_1} and z_{j_2} represent the signs of the active variables entering the model in the first and second steps respectively.

Remark 9. *Selection regions in FS: As we can see from above that the selection regions in this example take the form of a cone rather than the usual orthant and cube yielded by the ℓ_1 penalty in the variants of Lasso. We can still write the selection region as $\mathcal{R}_{O_1}=\mathcal{R}_{E_1}\times\prod_{j\neq j_1}\mathcal{R}_{j,-E_1}$ where*

$$\mathcal{R}_{j,-E_1}=\{o_{j,-j_1}:|o_{j,-j_1}|\leq|o_{j_1}|\}\text{ for } j=\{1,2,...,p\}\setminus E_1.$$

That is the inactive selective constraints are all separable in the $p-1$ coordinates, although they are determined by the active optimization variable unlike the example of Lasso. The probability of the inactive optimization variables being constrained to be smaller in magnitude than $|o_{j_1}|$ can be computed exactly as $\mathcal{B}(o_{j_1};s)$ conditional on data s and active variable o_{j_1}. A similar computation goes through for more than 1 step of FS.

Primal problem: Denoting the separable inactive selection regions $\mathcal{R}_{j,-E_1} = \{o_{j,-j_1} : |o_{j,-j_1}| \leq |o_{j_1}|\}$ for $j = \{1,2,...,p\} \setminus E_1$ and $\mathcal{R}_{j,-E_2} = \{o_{j,-\{j_1,j_2\}} : |o_{j,-\{j_1,j_2\}}| \leq |o_{j_2}|\}$ for $j = \{1,2,...,p\} \setminus E_2$, solve $2p-3$ univariate Gaussian probabilities as $\mathcal{B}(o_{j_1};s)$ and $\mathcal{B}(o_{j_2};s)$ where

$$\log \mathcal{B}(o_{j_1};s) = \sum_{j \neq j_1} \left\{ \log \Phi\left((o_{j_1} - X_{j,-j_1}^T s)/\tau\right) - \log \Phi\left((-o_{j_1} - X_{j,-j_1}^T s)/\tau\right) \right\}$$

$$\log \mathcal{B}(o_{j_2};s) = \sum_{j \neq j_1,j_2} \left\{ \log \Phi\left((o_{j_2} - \tilde{X}_{j,-j_2}^T s)/\tau\right) - \log \Phi\left((-o_{j_2} - \tilde{X}_{j,-j_2}^T s)/\tau\right) \right\}.$$

Solve an optimization over (s, o_{j_1}, o_{j_2}) where $s \in \mathbb{R}^n, o_{j_1} \in \mathbb{R}, o_{j_2} \in \mathbb{R}$ with sign barriers $b_{\mathcal{R}_{j_1}}(.)$ and $b_{\mathcal{R}_{j_2}}(.)$ on o_{j_1} and o_{j_2}, that is:

$$\inf_{s,o_{j_1},o_{j_2}} \left\{ \frac{\|s - X^*\beta^*\|^2}{2\sigma^2} + \frac{\|-X_{j_1}^T s + o_{j_1}\|^2}{2\tau^2} - \log \mathcal{B}(o_{j_1};s) + b_{\mathcal{R}_{E_1}}(o_{j_1}) \right.$$
$$\left. + \frac{\|\tilde{X}_{j_2}^T s + o_{j_2}\|^2}{2\tau^2} - \log \mathcal{B}(o_{j_2};s) + b_{\mathcal{R}_{E_2}}(o_{j_2}) \right\}.$$

Dual problem: Solve a dual optimization over $u_1 \in \mathbb{R}^p$ and $u_2 \in \mathbb{R}^{p-1}$, as stated below:

$$\inf_{u_1,u_2} \left\{ \frac{1}{2}\sigma^2 \|Xu_1 + \tilde{X}u_2\|^2 - (Xu_1 + \tilde{X}u_2)^T X^*\beta^* + \frac{1}{2}\tau^2 u_1^2 + \frac{1}{2}\tau^2 u_2^2 \right.$$
$$\left. + b_{\mathcal{R}_{O_1}}^*(u_1) + b_{\mathcal{R}_{O_2}}^*(u_2) \right\}.$$

5.3. A 2-stage query: thresholding followed by Lasso

We present an example of a two-stage screening method in the linear regression setting with a fixed design matrix X with normalized columns; we derive the approximating optimization problem to provide inference in a Bayesian model with prior π on β^* and $Y|\beta^* \sim \mathcal{N}(X^*\beta^*, \sigma^2 I)$. The selective analysis comprises of two stages of screening based on realizations ω_1, ω_2 from independent Gaussian distributions, each with all i.i.d. mean 0 components and variance $\tau^2 I$. The first query is a randomized marginal screening across the Z-statistics at a nominal threshold vector α, that solves

$$\min_\beta \frac{1}{2}\|\beta - X^T y/\hat{\sigma}\|_2^2 - \omega^T \beta + I_{\ell_\infty}^\alpha(\beta); \text{ with}$$

$$I_{\ell_\infty}^\alpha(\beta) = \begin{cases} 0 & \|\beta\|_\infty \leq \alpha \\ \infty & \text{otherwise.} \end{cases}$$

This results in output (E_1, z_{E_1}), the active set of marginally most correlated predictors with active signs from Stage-I screening. Denoting $\tilde{X} = X_{E_1} \in \mathbb{R}^n \times \mathbb{R}^{E_1}$, the predictor matrix with selected predictors from the first round of screening,

the second query is a randomized lasso query that solves (5.1) with design matrix $\tilde{X}$ to yield active set E_2 with signs z_{E_2}.

The first step describes the inversion maps and selective constraints encoding the two selective queries where data vector $S = Y$.

Inversion maps:

$$\text{Map I:} \quad \omega_1 = \begin{pmatrix} \alpha z_{E_1} - X_E^T s/\hat{\sigma} + o_{E_1} \\ -X_{-E}^T s/\hat{\sigma} + o_{-E_1} \end{pmatrix}$$

$$\text{Map II:} \quad \omega_2 = \begin{pmatrix} -\tilde{X}_{E_2}^T s + (\tilde{X}_{E_2}^T \tilde{X}_{E_2} + \epsilon I) o_{E_2} + \lambda z_{E_2} \\ -\tilde{X}_{-E_2}^T s + \tilde{\tilde{X}}_{-E_2}^T \tilde{X}_{E_2} o_{E_2} + o_{-E_2} \end{pmatrix}$$

inducing respective selection regions $\mathcal{R}_{O_i} = \mathcal{R}_{E_i} \times \Pi_j \mathcal{R}_{j,-E_i}$; $i = 1,2$ as

$$\mathcal{R}_{O_1} = \prod_{j=1}^{|E_1|} \{\text{sign}(o_{j,E_1}) = z_{j,E_1}\} \times \prod_{j=1}^{p-|E_1|} \{|o_{j,-E_1}| \leq \alpha\}$$

$$\mathcal{R}_{O_2} = \prod_{j=1}^{|E_2|} \{\text{sign}(o_{j,E_2}) = z_{j,E_2}\} \times \prod_{j=1}^{|E_1|-|E_2|} \{|o_{j,-E_2}| \leq \lambda\}.$$

Using the facts that the convex conjugate of a Gaussian log-MGF with mean μ and variance $\gamma^2 I_k$ at vector x is $\|x - \mu\|_2^2/2\gamma^2$ and the log-MGF is $\mu^T x + \gamma^2 \|x\|^2/2$ we derive the primal and dual optimization problems to sample from the approximate posterior.

Primal problem: The primal marginalizes over the inactive sub-gradients followed by the optimization over active variables and data in $n + |E_1| + |E_2|$ dimensions. Computing the exact log-Gaussian probabilities over intervals $[-\alpha, \alpha]$ and $[-\lambda, \lambda]$ as

$$\log \mathcal{B}(o_{E_1}; s) = \sum_{j=1}^{p-|E_1|} \log \left\{ \Phi \left(\{\alpha - X_{j,-E_1}^T s/\hat{\sigma} + o_{j,E_1}\}/\tau \right) \right.$$
$$\left. - \Phi \left(\{-\alpha - X_{j,-E_1}^T s/\hat{\sigma} + o_{j,E_1}\}/\tau \right) \right\}$$

$$\log \mathcal{B}(o_{E_2}; s) = \sum_{j=1}^{|E_1|-|E_2|} \log \left\{ \Phi \left(\{\lambda + \tilde{X}_{j,-E_2}^T \tilde{X}_{E_2} o_{E_2} - \tilde{X}_{j,-E_2}^T s\}/\tau \right) \right.$$
$$\left. - \Phi \left(\{-\lambda + \tilde{X}_{j,-E_2}^T \tilde{X}_{E_2} o_{E_2} - \tilde{X}_{j,-E_2}^T s\}/\tau \right) \right\}$$

we have the optimization in the primal form as

$$-\inf_{s, o_{E_1}, o_{E_2}} \left\{ \frac{\|s - X^* \beta^*\|^2}{2\sigma^2} + \frac{\|\alpha z_{E_1} - X_{E_1}^T s/\hat{\sigma} + o_{E_1}\|^2}{2\tau^2} - \log \mathcal{B}(o_{E_1}; s) \right.$$

$$\left. + b_{\mathcal{R}_{E_1}}(o_{E_1}) + \frac{\|\tilde{X}_{E_2}^T \tilde{X}_{E_2} o_{E_2} - \tilde{X}_{E_2}^T s + \lambda z_{E_2}\|^2}{2\tau^2} - \log \mathcal{B}(o_{E_2}; s) + b_{\mathcal{R}_{E_2}}(o_{E_2}) \right\}.$$

<div align="center">

TABLE 3
Expt 1- Model I

</div>

Query	Coverage		Risk		Lengths		
	adjusted	unadjusted	adjusted	unadjusted	adjusted	unadjusted	n, p
Lasso (Fixed X) Primal	86.20%	22.72%	1.85	5.43	4.55	3.32	200 1000
Lasso (Fixed X) Dual	89.70%	51.38%	1.81	3.38	4.41	3.31	1000 200
Lasso (Random X)	85.42%	43.44%	1.87	3.74	4.41	3.31	1000 100
Carved Lasso (Random X)	87.30%	23.16%	4.54	4.68	6.05	3.32	1000 100
FS	85.15%	14.85%	3.46	7.11	4.62	3.30	200 1000

Dual problem: With P_1 and P_2 identified respectively as I_p and $\begin{bmatrix} \tilde{X}_{E_2}^T \tilde{X}_{E_2} + \epsilon I & 0 \\ \tilde{X}_{-E_2}^T \tilde{X}_{E_2} & I \end{bmatrix}$ from the randomization maps, solve an optimization over $u_1 \in \mathbb{R}^p$ and $u_2 \in \mathbb{R}^{|E_1|}$ as below:

$$\inf_{u_1, u_2} \left\{ \frac{1}{2} \sigma^2 \| X^T u_1 / \hat{\sigma} + \tilde{X}^T u_2 \|^2 - (X^T u_1 / \hat{\sigma} + \tilde{X}^T u_2)^T X^* \beta^* + u_1^T \begin{pmatrix} \lambda z_{E_2} \\ 0 \end{pmatrix} \right.$$
$$\left. + u_2^T \begin{pmatrix} \alpha z_{E_1} \\ 0 \end{pmatrix} + \frac{1}{2} \tau^2 u_1^2 + \frac{1}{2} \tau^2 u_2^2 + b_{\mathcal{R}_{O_1}}^* (P_1^T u_1) + b_{\mathcal{R}_{O_2}}^* (P_2^T u_2) \right\}.$$

6. Experiments

6.1. Simulated models

We conduct different experiments to show the coverage and risk properties of estimates, obtained using our methods in comparison to those based on untruncated approach. We vary generative models across our experiments: this shows that our methods show good performance even under misspecified models.

In the first experiment, we use Model I in Section 1.2 to generate our data. The ground truth is the null model $Y \sim \mathcal{N}(0, I)$. For a fixed design X, we draw $Y \in \mathbb{R}^n$ for every repetition using the same X. For a random design X, we randomly draw $X \in \mathbb{R}^{n \times p}$ with Gaussian entries and draw Y conditional on X and the underlying parameter β in each repetition of the experiment. The columns of design X are scaled by $1/\sqrt{n}$ in all cases. The second experiment uses Model II in Section 1.2 as a generative mechanism. This is a Bayesian model with ground truth $\Theta_E(\beta) = (X_E^T X_E)^{-1} X_E^T X \beta$, determined by E in each trial.

In table 3, we compare the empirical coverage of the credible intervals, the risk of the posterior mean and the length of intervals between the approximating method that aims at the pseudo selective posterior and the usual Bayesian posterior inference. For the Bayesian model, we report the empirical (Bayesian) FCR, Bayes risk of the posterior mean and the length of intervals in table 4.

TABLE 4
Expt 2- Model II

Query	Bayesian CR		Bayes risk		Lengths		
	adjusted	unadjusted	adjusted	unadjusted	adjusted	unadjusted	n, p
Lasso (Fixed X) Primal	90.99%	33.86%	1.49	4.28	4.49	3.34	200 1000
Lasso (Fixed X) Dual	87.12%	61.99%	1.71	2.71	4.22	3.31	1000 200
Lasso (Random X)	88.26%	55.04%	1.77	3.01	4.23	3.31	1000 100
Carved Lasso (Random X)	82.86%	38.30%	5.98	6.75	5.92	3.31	1000 100
FS	72%	51.40%	3.19	3.99	4.07	3.29	200 1000

The queries as described in Examples under 5.1 are conducted under centered Gaussian randomization with variance $\tau^2 I_p$; with the exception of the carved query which inherits randomization from a randomly chosen split of the data. For inference, we use the selected model $\mathcal{N}(X_E\beta_E, I)$ and a non-informative prior on β_E, where E is the active set from the selective query. In both experiments, we use a misspecified likelihood and prior. Note that despite the fact that the model for inference is a mis-specified one under the true generative models, our methods display superiority in terms of coverage and risk properties in comparison to the unadjusted estimates. The first column states the query- the Lasso with a fixed and random design, a carved Lasso with a random design and 1 step of forward stepwise (FS) and the last column gives the regression dimensions n and p.

Remark 10. *In the case of forward stepwise, the model used for subsequent inference is parametrized by mean $X_E\beta_E$, where E is a singleton set that includes the most correlated variable with the response. This is a very misspecified model under the Bayesian generative mechanism, characterized by Model II. Unsurprisingly, the Bayesian CR under such a misspecified model dips to 72%. We however, use this example to highlight that adjusted inference is more robust than the unadjusted counterpart.*

In table 5, we show a comparison of data carving with data splitting in Models I and II where naive splitting uses a fraction of the data for selection and the remaining fraction for inference. The carved Lasso, on the other hand, uses the truncated likelihood based on the entire dataset for inference; the truncating event being a function of the fraction of data employed in lasso query. We see an advantage of carving over splitting in terms of inferential power, measured as lengths of credible intervals and in estimation risk of the posterior means.

The generative mechanism in the fourth experiment is a frequentist model that deviates from the all noise model considered in Experiment 1. It gives an assessment of estimates based on the output from a Lasso query with a fixed X design using the primal and dual problems by varying the sparsity

TABLE 5
Expt 3- Comparison of data-carving and data-splitting

$n = 1000, p = 100$	Coverage/Bayesian CR		Risk/Bayes risk		Lengths	
Split, Model	carved	split	carved	split	carved	split
50%; Model I	86.17%	89.72%	1.50	2.20	4.27	4.73
50%; Model II	87.30%	88.28%	1.70	2.45	4.17	4.74
80%; Model I	87.30%	90.8%	4.54	5.23	6.05	7.56
80%; Model II	82.86%	88.15%	5.98	6.54	5.92	7.54

TABLE 6
Expt 4- Deviation from noise model

$n = 500, p = 3000$	Coverage		Risk		Lengths	
Sparsity	adjusted	unadjusted	adjusted	unadjusted	adjusted	unadjusted
0	85.25%	23.22%	2.09	5.56	4.52	3.31
5	87.27%	54.99%	1.73	3.35	4.06	3.35
10	86.05%	63.98%	1.80	2.76	4.05	3.36
20	83.51%	69.49%	1.86	3.73	4.10	3.43

TABLE 7
Expt 4- Deviation from noise model

$n = 3000, p = 500$	Coverage		Risk		Lengths	
Sparsity	adjusted	unadjusted	adjusted	unadjusted	adjusted	unadjusted
0	85.98%	39.94%	1.96	3.93	4.36	3.30
5	88.23%	64.68%	1.44	2.56	3.76	3.30
10	85.06%	75.4%	1.48	1.98	3.68	3.30
20	87.1%	79.4%	1.45	1.60	3.71	3.31

TABLE 8
Expt 5- A 2-stage screening

Model	Coverage/Bayesian CR		Risk/Bayes risk		Lengths		
	adjusted	unadjusted	adjusted	unadjusted	adjusted	unadjusted	n, p
I	86.52%	39.20%	3.22	4.08	5.74	3.40	200, 1000
I	89.26%	29.64%	2.06	4.27	4.89	3.30	1000, 200
II	85.00%	53.88%	3.71	3.27	5.66	3.43	200, 1000
II	91.86%	32.39%	2.01	4.10	4.87	3.31	1000, 200

levels in the true generative mechanism. Based on a fixed predictor matrix, we simulate $Y \in \mathbb{R}^n$ in each draw as below for a sparse vector $\beta_{\mathcal{S}}$ with true support $\mathcal{S} \subset \{1, 2, \cdots, p\}$ $Y|X, \beta_{\mathcal{S}} = X_{\mathcal{S}}\beta_{\mathcal{S}} + \epsilon, \ \epsilon \sim \mathcal{N}(0, I_n)$. We use the primal and dual formulation of the optimization discussed in 5.1 for providing estimates in a high dimensional sparse problem $n = 500, \ p = 3000$ and in the low dimensional regime $n = 3000, \ p = 500$ respectively. We vary the sparsity levels as $|\mathcal{S}| = 0, 5, 10, 20$ signals, each with magnitude 7. Tables 6 and 7 show that the adjusted estimates have superior risk and coverage properties as compared to the unadjusted estimates, using a selected model appended to a diffuse prior.

The final experiment gives the performance of the estimates post the 2-stage screening query with a fixed X design, described in Section 5.3. The target

and model are determined by the final screened model E_2 that combines the output from the two screenings. The coverage and risk comparisons for the above screening procedure are given in table 8. The first column gives the Model generating the data and the last column gives the dimensions of the simulation.

6.2. Data analysis: inference on causal variants

To illustrate the inferential gains with the truncated Bayesian method, we provide adjusted effect size estimates for SNPs (Single nucleotide polymorphisms) that have been data-mined as the strongest associations with gene expression. An analyst will be confronted in defending the strength of these associations if she does not overcome the selective bias encountered in estimation post data-snooping. With gene expression data as the outcome, we give adjusted effect size estimates of SNPs that have been selected as the set of probable causal genetic variants. We highlight the differences between the adjusted Bayesian approach and the unadjusted counterpart (that is inadaptive to selection); we also depict the higher statistical power associated with the adjusted Bayesian estimates post a randomized selection as opposed to the estimates based on [13] post a non-randomized selection.

The data analyzed in this work involves gene expression data $Y \in \mathbb{R}^{97}$ for a gene collected from the human tissue -Liver for a sample of 97 densely genotyped individuals. More details on this data-set are included in Appendix B.1. The exceedingly small sample size in this analysis does not allow the analyst to reserve a hold-out data set for inference. The goal here, is to quantify the effect sizes of variants that have been selected from a set of 5233 of potential predictors, namely $X \in \mathbb{R}^{97 \times 5233}$ as predictors that best explain the variance in expression levels of the gene under study. More specifically, the columns of X represent local genetic variants measured as SNPs that lie within 1MB of the transcription start site of the gene. This data has been investigated as a part of a genome-wide association study conducted in [6, 19] with focus on identifying the significant associations between gene expression and genetic variants across different human tissues. The afore-mentioned works aimed at recognizing genes with at least one causal variant, called eGenes. A more recent work [1] performs a secondary analysis on the eGenes to further identify variants that regulate the expression for these genes. This involves a search over the local variants around the genes. In this work, we employ one such selection procedure, the commonly used Lasso to pick promising predictors and apply our method to give estimates for effect sizes of these selected SNPs based on the truncated posterior.

Below, we outline the analysis that leads to the selection of SNPs. To aid interpretability and recovery of a meaningful set of effects, we a perform hierarchical clustering with a minimax linkage on the set of 5233 SNPs, see [3]. The distance measure between SNPs X_i and X_j is defined as $d(X_i, X_j) = 1 - \rho(X_i, X_j)$ where $\rho(X_i, X_j)$ is the empirical correlation between two SNPs X_i, $X_j \in \mathbb{R}^{97}$. This algorithm introduced in [2] clusters the SNPs and gives a prototype for each cluster. The number of clusters is chosen so that each of the 5233 SNPs has a

correlation of at least 0.5 with at least one of the prototypes. Applying a typical selection procedure like the Lasso on the set of local variants without pruning it to prototypes is not ideal in this analysis as the local variants share substantial empirical correlation; the Lasso will typically suffer from an inability to recover the true set of signals. [22] identifies this shortcoming of the Lasso and proposes inference on effect sizes post a Lasso on prototypes of clusters in such scenarios. While the prototypes in [22] are determined in a greedy fashion; the cluster representative being the most associated with the response, we adapt a completely unsupervised approach here in order to determine the clusters and prototypes with no data-snooping. Using the described hierarchical clustering, we obtain 320 prototype SNPs, each of which has a correlation of at least 0.5 with the SNPs in its cluster. We finally run a randomized Lasso query given by (5.1) on the prototype SNPs with Gaussian randomization. With the tuning parameter is set at the theoretical $\lambda = \hat{\sigma} \cdot \mathbb{E}(X^T \epsilon)$, the randomized Lasso query selects a set of 21 potential regulatory variants; $\hat{\sigma}$ is estimated as 0.4. The ratio of the randomization to noise scale in the data is set at 0.5.

We provide inference for the population least squares coefficients that correspond to the selected set E of SNPs; the adaptive target $(X_E^T X_E)^{-1} X_E^T \mathbb{E}(Y \mid X, \beta)$ is used as a quantification of the effect sizes of the selected SNPs. We assume model $Y \sim \mathcal{N}(X_E \beta_E, \hat{\sigma}^2 I)$ and a non-informative prior on β_E (similar to the simulations in Section 6.1). Figure 2 gives a comparison of the effect sizes of selected SNPs using the proposed truncated approach with the unadjusted Bayesian estimates. Under the diffuse prior, the unadjusted estimates will be centered around the OLS estimator $(X_E^T X_E)^{-1} X_E^T y$ with variance given by the diagonal entries of $(X_E^T X_E)^{-1}$. The optimizations that we solve to obtain the truncated Bayesian estimates are laid out in Section 5. We note the differences in effect sizes based on the adaptive posterior and the unadjusted posterior; we can see that the selected SNPs at positions **492**, **606**, **2960**, **3509**, **3574** will be reported as significantly associated with the gene expression if the analyst did not account for selection. The adjusted inference however, shows that the effect sizes of these SNPs are significantly biased by selection; the adjusted Bayesian intervals for these reportedly significant SNPs cover 0.

We supplement the above randomized effect size estimates with non randomized frequentist inference of [13] post the usual Lasso query (without the randomization term in (5.1)). Figure 3 plots the exact frequentist intervals post a Lasso selection compared against the unadjusted intervals. The non-randomized selection includes 18 SNPs as opposed to 21 SNPs picked up by the randomized Lasso; the common SNPs picked by both queries occur at positions **158**, **492**, **606**, **1830**, **2259**, **2786**, **2876**, **2926**, **2960**, **3155**, **3509**, **3574**. The exact frequentist intervals adjusted for selection again show that the SNPs at **492**, **606**, **2960**, **3509**, **3574** are no longer statistically significant effects, as opposed to the unadjusted estimates. The comparison with the randomized intervals in Figure 2 shows that the estimates post a randomized version of Lasso have more statistical power, highlighted in the shorter lengths of the randomized intervals when compared against the exact frequentist intervals of Figure 3.

To validate the inferential guarantees of our estimates in the above analysis,
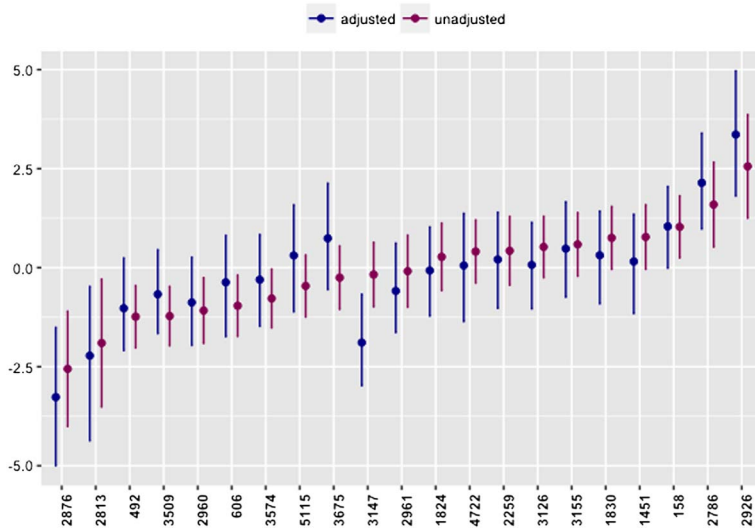
FIG 2. *Effect size estimates: posterior mean and credible intervals are based on the truncated and unadjusted Bayesian posterior. The adjusted intervals have an average length of 2.31, the unadjusted intervals have an average length of 1.86.*
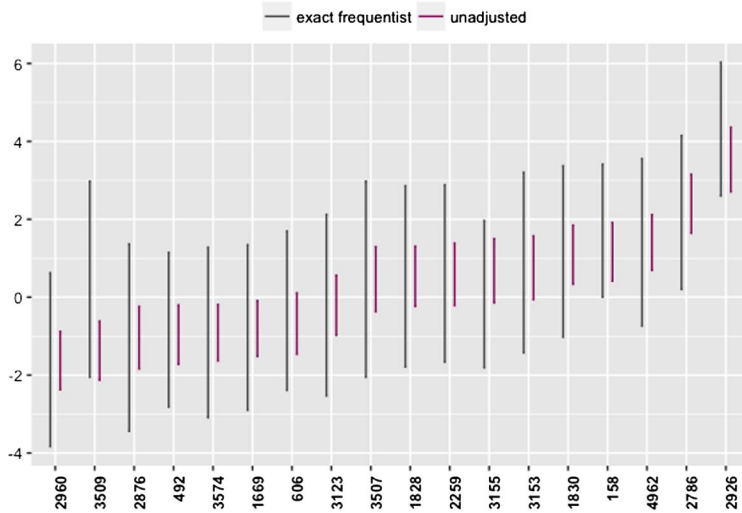

FIG 3. *Effect size estimates: adjusted intervals are the exact frequentist intervals constructed by conditioning on the polyhedral selection region of Lasso. Unadjusted intervals are centered around the OLS estimator post Lasso, $(X_E^T X_E)^{-1} X_E^T y$ with variance of j-th coefficient given by $(X_E^T X_E)^{-1}_{j,j}$. The adjusted intervals have an average length of 4.25, the unadjusted intervals have an average length of 1.56.*

we conclude with a simulation design based on the predictor matrix of SNPs X as considered above. We consider a sparse regime varying the number of signals $|S| \in \{0, 1, 2, 3\}$. In this sparse regime, we simulate the response Y from a model based on the 5233 predictors with $|S|$ true signals as follows:

- subsample $|S|$ clusters from the 320 clusters of SNPs (obtained by hierarchical clustering with a minimax linkage, described above),
- subsample one SNP further from each of the $|S|$ subsampled cluster as the positions of the true signals; the set of true signals is called $\mathcal{S}$ with cardinality $|S|$,
- draw response $y \in \mathbb{R}^{97}$ as $y = X_{\mathcal{S}}\beta_{\mathcal{S}} + \epsilon$; $\epsilon \sim \mathcal{N}(0, I)$, where $|\beta_{j,\mathcal{S}}|$ are of equal strength for $j \in \mathcal{S}$. We vary the the magnitude of signals over the set $\{10, 5, 2.5\}$ corresponding to roughly $K\sqrt{2 \cdot \log p}$ with $K = 4.5, 2, 1$; $p = 320$ respectively. These signal strengths correspond to three SNR regimes-strong, moderate and weak signal regime.

We evaluate the coverage and risk for the adjusted and unadjusted estimates averaged over the selected SNPs and across repetitions of an experiment with 50 trials in the 3 different signal regimes. In each repetition, we provide inference about the ground truth for the population least squares coefficients given by

$$(X_E^T X_E)^{-1} X_E^T \mathbb{E}(Y|X, \beta) = (X_E^T X_E)^{-1} X_E^T X_{\mathcal{S}} \beta_{\mathcal{S}}$$

under the selected model and non-informative prior as before. Note that the prototypes might not be positions of true signals in our simulation study, thereby, the model we assume for inference might be a misspecified model. We see that even with a misspecified model, the adjusted Bayesian estimates show superior performance than the unadjusted estimates, both in terms of coverage and risk. We also show the higher power of the adjusted Bayesian inference post the randomized Lasso with the non-randomized exact frequentist estimates of [13]. The blue color gives the adjusted Bayesian inference under the diffuse prior, the grey color represents the exact frequentist inference of [13] post a non-randomized Lasso query and the red color denotes the unadjusted Bayesian inference. [13] does not give a selection adjusted point estimate; the grey curve in Figure 5 plots the risk of the OLS estimator $(X_E^T X_E)^{-1} X_E^T y$ where E is the set of SNPs selected by (non-randomized) Lasso. The labels in the x-axis represent the number of true signals, $|S| = 0, 1, 2, 3$. The columns give comparison of estimates in the three signal regimes: from left to right- strong, moderate and weak corresponding to model of equally contributing signals with varying strengths $10, 5, 2.5$ respectively.

7. Concluding remarks

The motivations to adjust for selection through a truncation on the generative model is the same as the frequentist line of works; though the technicalities with imposing a Bayesian model post selection are different. While prior works make progress in formalizing a selective Bayesian methodology, the current work
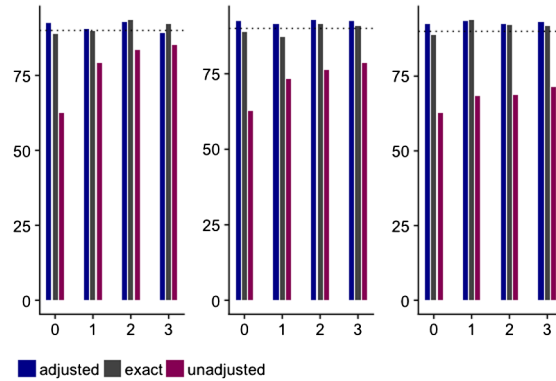
FIG 4. *Comparison of coverages in the strong, moderate and weak signal regime: The black dotted line marks the 90% target coverage.*
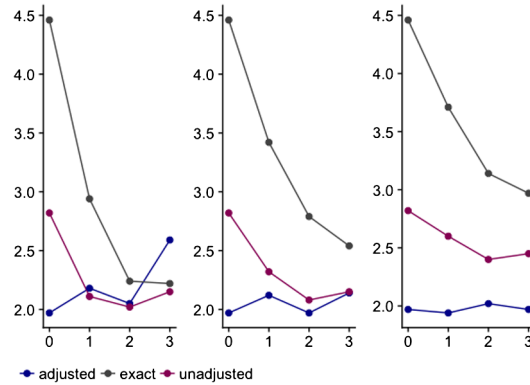

FIG 5. *Comparison of risk in strong, moderate and weak signal regime*

makes significant contributions in proposing a concrete computational recipe to approximate the selective posterior after systematic randomized procedures. The methods extend to multi stage selective queries, marginalizing over randomizations from each selective stage. An attractive property of this approach is scalability in both regimes of inference with empirical demonstration of frequentist coverage with credible intervals and risk of the posterior mean based on the approximate selective posterior.

An interesting future direction includes establishing a Bernstein von Mises result in the selective Bayesian paradigm. We empirically see that the truncated Bayesian methods somewhat recover frequentist coverage under diffuse priors just as they would in the untruncated regime of inference. From a purely application point of view, we see scope of the methodology in this work to be applied to other genome-wide studies where the true model describing the association of phenotypes with variants is assumed to be highly sparse. Inference post iden-
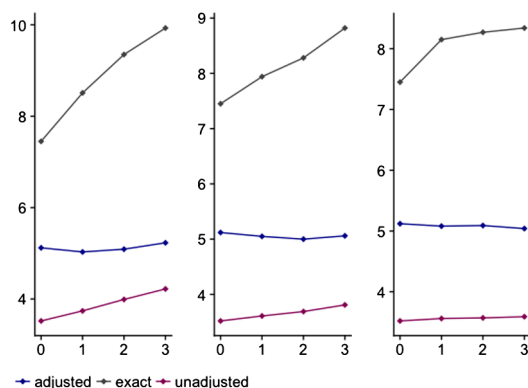
FIG 6. *Comparison of lengths in strong, moderate and weak signal regime*

tification of causal variants is an important goal in these studies. Our methods can provide reliable and reproducible effect size estimates in such settings. The Bayesian model in particular, allows an analyst to leverage information from an objective or subjective prior that can arise from prior experimentation in these studies. Developing tools to sample from an intractable posterior modeled using the truncated framework has been the focus of this paper, this allows the analyst to take full advantage of the Bayesian machinery post selection and provide estimates with better coverage and risk properties than the usual Bayesian estimates.

Acknowledgement

Appendix A: Proofs of results in Sections 3, 4 and 5

Proof of Theorem 1

Proof. To prove this, we derive an upper bound on $\log \mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$ in terms of the log-MGF of the augmented random variable (S, O).

$$\log \mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$$
$$\leq \log \mathbb{E}\left[\exp(\sup_{s, o \in \mathcal{R}}\{-\alpha_1^T s - \alpha_2^T o\})\exp(\alpha_1^T S + \alpha_2^T O)\Big|\beta^*\right]$$

$$= -\inf_{s,o\in\mathcal{R}}\{\alpha_1^T s + \alpha_2^T o\} - \log\mathbb{E}[\exp(\alpha_1^T S + \alpha_2^T O)|\beta^*]$$

Since the above bound holds for any $\alpha_1 \in \mathbb{R}^d$ and $\alpha_2 \in \mathbb{R}^p$, we can optimize over the choices of α_1, α_2 to obtain the upper bound

$$-\sup_{\alpha_1,\alpha_2}\inf_{s,o\in\mathcal{R}}\{\alpha_1^T s + \alpha_2^T o\} - \log\mathbb{E}[\exp(\alpha_1^T S + \alpha_2^T O)|\beta^*].$$

A minimax equality for convex, compact selection region $\mathcal{R}$ justifies the swapping of infimum and supremum to lead to the bound

$$-\inf_{s,o\in\mathcal{R}}\left\{\sup_{\alpha_1,\alpha_2}\alpha_1^T s + \alpha_2^T o - \log\mathbb{E}(\exp(\alpha_1^T S + \alpha_2^T O)|\beta^*)\right\}. \tag{A.1}$$

The main step is computation of the log-MGF $\log\mathbb{E}(\exp(\alpha_1^T S + \alpha_2^T O)|\beta^*)$, which is possible through the change of measure facilitated by the inversion map in (2.4). Using the joint density of the vector (S, O) in (2.5) and writing $o = P^{-1}(\omega - Ds - q)$, we have $\mathbb{E}[\exp(\alpha_1^T S + \alpha_2^T O)|\beta^*]$

$$= \int \exp\left(\alpha_1^T s + \alpha_2^T o\right)|J|f(s|\beta^*)g(Ds + Po + q)ds do$$

$$= \int \exp\left(\alpha_1^T s + \alpha_2^T P^{-1}(\omega - Ds - q)\right)f(s|\beta^*)g(\omega)ds d\omega$$

$$= \exp(-\alpha_2^T P^{-1}q)\int \exp((\alpha_1 - D^T P^{-1^T}\alpha_2)^T s)f(s|\beta^*)ds$$

$$\times \int \exp(\alpha_2^T P^{-1}\omega)g(\omega)d\omega$$

$$= \exp(-\tilde{\alpha}_2^T q)\mathbb{E}[\exp(\tilde{\alpha}_1 S)|\beta^*)] \cdot \mathbb{E}[\exp(\tilde{\alpha}_2^T \Omega)]$$

$$= \exp(-\tilde{\alpha}_2^T q)\exp\Lambda_f(\tilde{\alpha}_1|\beta^*)\exp\Lambda_g(\tilde{\alpha}_2)$$

with $\tilde{\alpha}_1 = \alpha_1 - D^T(P^{-1})^T\alpha_2$ and $\tilde{\alpha}_2 = (P^{-1})^T\alpha_2$. Plugging

$$\log\mathbb{E}[\exp(\alpha_1^T S + \alpha_2^T O|\beta^*)] = -\tilde{\alpha}_2^T q + \Lambda_f(\tilde{\alpha}_1|\beta^*) + \Lambda_g(\tilde{\alpha}_2)$$

into (A.1) gives the upper bound for $\log\mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$ in terms of the log-MGF corresponding to the data $\Lambda_f(.|\beta^*)$ and the randomization $\Lambda_g(.)$ as

$$-\inf_{s,o\in\mathcal{R}}\left\{\sup_{\tilde{\alpha}_1,\tilde{\alpha}_2}\tilde{\alpha}_1^T s + \tilde{\alpha}_2^T(Po + Ds) - \log\mathbb{E}[\exp(\alpha_1^T S + \alpha_2^T O|\beta^*)]\right\}$$

$$= -\inf_{s,o\in\mathcal{R}}\left\{\sup_{\tilde{\alpha}_1,\tilde{\alpha}_2}(\tilde{\alpha}_1^T s + \tilde{\alpha}_2^T(Po + Ds + q) - \Lambda_f(\tilde{\alpha}_1|\beta^*) - \Lambda_g(\tilde{\alpha}_2))\right\}$$

$$= -\inf_{s,o\in\mathcal{R}}\Lambda_f^*(s|\beta^*) + \Lambda_g^*(Ds + Po + q). \qquad \square$$

Proof of Theorem 2

Proof. The volume of the selection region

$$\mathcal{R} = \mathcal{R}_S \times \mathcal{R}_O = \mathcal{R}_S \times \mathcal{R}_E \times \prod_{j=1}^{p-|E|}\mathcal{R}_{j,-E}$$

based on decoupling of the randomization density into active and inactive coordinates is given by

$$\mathbb{P}((S, O) \in \mathcal{R}|\beta^*)$$

$$= |J| \cdot \int_{\mathcal{R}_S} f(s|\beta^*) \cdot \int_{\mathcal{R}_E} g_E(D_E s + P_E o_E + q_E)$$

$$\cdot \prod_{j=1}^{p-|E|} \int_{\mathcal{R}_{j,-E}} g_{j,-E}(o_{j,-E} + D_{j,-E}s + P_{j,-E}o_E + q_{j,-E}) do_{j,-E} do_E ds$$

$$= |J| \cdot \int_{\mathcal{R}_S} f(s|\beta^*) \cdot \int_{\mathcal{R}_E} g_E(D_E s + P_E o_E + q_E) \mathcal{B}(o_E; s) do_E ds$$

$$= \mathbb{E}\left[\mathcal{B}(O_E; S) 1_{(S,O_E) \in \mathcal{R}_S \times \mathcal{R}_E} \big| \beta^* \right]$$

An upper bound on $\log \mathbb{E}\left[\mathcal{B}(O_E; S) 1_{(S,O_E) \in \mathcal{R}_S \times \mathcal{R}_E} \big| \beta^* \right]$ is given by

$$\log \mathbb{E}\left[\mathcal{B}(O_E; S) 1_{(S,O_E) \in \mathcal{R}_S \times \mathcal{R}_E} \big| \beta^* \right]$$

$$= \log \mathbb{E}\left[\exp(\log \mathcal{B}(O_E; S) - \alpha_1^T S - \alpha_2^T O_E) \right.$$

$$\left. \cdot \exp(\alpha_1^T S + \alpha_2^T O_E) \cdot 1_{(S,O_E) \in \mathcal{R}_S \times \mathcal{R}_E} \big| \beta^* \right]$$

$$\leq \sup_{s \in \mathcal{R}_S, o_E \in \mathcal{R}_E} \left\{ \log \mathcal{B}(o_E; s) - \alpha_1^T s - \alpha_2^T o_E \right\}$$

$$+ \log \mathbb{E}\left[\exp(\alpha_1^T S + \alpha_2^T O_E) \big| \beta^* \right]$$

Optimizing over $\alpha_1 \in \mathbb{R}^d$ and $\alpha_2 \in \mathbb{R}^p$, we have

$$\log \mathbb{E}\left[\mathcal{B}(O_E; S) 1_{(S,O_E) \in \mathcal{R}_S \times \mathcal{R}_E} \big| \beta^* \right]$$

$$\leq \inf_{\alpha_1, \alpha_2} \sup_{s \in \mathcal{R}_S, o_E \in \mathcal{R}_E} \left\{ \log \mathcal{B}(o_E; s) - \alpha_1^T s - \alpha_2^T o_E \right\}$$

$$+ \log \mathbb{E}\left[\exp(\alpha_1^T S + \alpha_2^T O_E) \big| \beta^* \right]$$

$$= - \sup_{\alpha_1, \alpha_2} \left\{ \inf_{s \in \mathcal{R}_S, o_E \in \mathcal{R}_E} \left\{ \alpha_1^T s + \alpha_2^T o_E - \log \mathcal{B}(o_E; s) \right\} \right.$$

$$\left. - \log \mathbb{E}\left[\exp(\alpha_1^T S + \alpha_2^T O_E) \big| \beta^* \right] \right\}.$$

By a minimax equality for compact, convex selection region $\mathcal{R}_S \times \mathcal{R}_E$ and the expression for log-MGF using the change of measure derived in the proof of Theorem 1, we have the result. $\square$

Proof of Theorem 3

Proof. With the introduction of variable $v = Ds + Po + q$, the dual of optimization

$$\inf_{s,o} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_g^*(Ds + Po + q) + b_{\mathcal{R}_O}(o) \right\}$$

in terms of dual variable $u \in \mathbb{R}^p$

$$\sup_u \inf_{s,o,v} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_g^*(v) + b_{\mathcal{R}_O}(o) + u^T(v - Ds - Po - q) \right\}$$

Solving Lagrangian $\mathcal{L}(u)$ over variables (s, o, v)

$$\inf_{s,o,v} \left\{ \Lambda_f^*(s|\beta^*) + \Lambda_g^*(v) + b_{\mathcal{R}_O}(o) + u^T(v - Ds - Po - q) \right\}$$

gives the optimizing equations

$$s = \nabla \Lambda_f(D^T u|\beta^*); v = \nabla \Lambda_g(-u), o = \nabla b_{\mathcal{R}_O}^{*}{}^{-1}(P^T u).$$

Thus, $\mathcal{L}(u) = -\Lambda_f(D^T u|\beta^*) - \Lambda_g(-u) - b_{\mathcal{R}_O}^*(P^T u) - u^T q$ and follows (3.7). □

Proof of Theorem 4

Proof. Plugging in the conjugate of the log-Gaussian MGF of data vector S, the logarithm of the pseudo posterior can be written as

$$\log \tilde{\pi}_E(\beta^*|s) = K + \log \pi(\beta^*) - \frac{(s - \mu(\beta^*))^T \Sigma_f^{-1}(s - \mu(\beta^*))}{2}$$

$$- \sup_{z \in \mathbb{R}^d} \left\{ z^T \Sigma_f^{-1} \mu(\beta^*) - \frac{1}{2} z^T \Sigma_f^{-1} z \right.$$

$$\left. - \inf_{o \in \mathbb{R}^p} \left\{ \Lambda_g^*(Dz + Po + q) + b_{\mathcal{R}_O}(o) \right\} \right\}$$

$$= K + \log \pi(\beta^*) - \frac{(s - \mu(\beta^*))^T \Sigma_f^{-1}(s - \mu(\beta^*))}{2}$$

$$- \delta^*(\Sigma_f^{-1} \mu(\beta^*)) + \frac{\mu(\beta^*)^T \Sigma_f^{-1} \mu(\beta^*)}{2}$$

for constant $K = -d \log 2\pi/2 - \log|\Sigma_f|/2$ and δ^* representing the conjugate of

$$\delta(z) = \frac{1}{2} z^T \Sigma_f^{-1} z + \inf_{o \in \mathbb{R}^p} \left\{ \Lambda_g^*(Dz + Po + q) + b_{\mathcal{R}_O}(o) \right\}.$$

The derivative of the log-pseudo posterior is thus given by

$$\frac{\partial \log \tilde{\pi}_E(\beta^*|s)}{\partial \beta^*} = \frac{\partial \log \pi(\beta^*)}{\partial \beta^*} + \left(\frac{\partial \mu}{\partial \beta^*} \right)^T \Sigma_f^{-1} s - \left(\frac{\partial \mu}{\partial \beta^*} \right)^T \Sigma_f^{-1} \nabla \delta^*(\Sigma_f^{-1} \mu)$$

$$= \frac{\partial \log \pi(\beta^*)}{\partial \beta^*} + \left(\frac{\partial \mu}{\partial \beta^*} \right)^T \Sigma_f^{-1}(s - s^*(\Sigma_f^{-1} \mu(\beta^*)))$$

for s^* satisfying

$$\arg\inf_{z\in\mathbb{R}^d}\left(z^T\Sigma_f^{-1}\mu(\beta^*)-\frac{1}{2}z^T\Sigma_f^{-1}\ z-\inf_{o\in\mathbb{R}^p}\left\{\Lambda_g^*(Dz+Po+q)+b_{\mathcal{R}_O}(o)\right\}\right).$$

The last equality follows by noting that $\nabla\delta^*(\Sigma_f^{-1}\mu)\ =\ \nabla\delta^{-1}(\Sigma_f^{-1}\mu)\ =\ s^*(\Sigma_f^{-1}\mu)$. $\square$

Proof of Theorem 5

Proof. (1) We can write $\lim_n\frac{1}{n}\log\mathbb{P}(\bar{S}_n\in\mathcal{R}_S',\bar{O}_n\in\mathcal{R}_O'|\beta_E)$ as

$$\lim_n\frac{1}{n}\log\mathbb{E}\left[\mathbb{P}(\bar{O}_n\in\mathcal{R}_O'|\bar{S}_n=s)1_{\bar{S}_n\in\mathcal{R}_S}|\beta_E\right]$$
$$=\lim_n\frac{1}{n}\log\mathbb{E}\left[\exp\left(n\cdot\frac{1}{n}\log\mathbb{P}(\bar{O}_n\in\mathcal{R}_O'|\bar{S}_n=s)\right)1_{\bar{S}_n\in\mathcal{R}_S'}\Big|\beta_E\right].$$

Note that $H_n(s)=\frac{1}{n}\log\mathbb{P}(\bar{O}_n\in\mathcal{R}_O'|\bar{S}_n=s)$ satisfies the limit (4.4). We also use the observation that $H_n'(s)=-\inf_{o\in\mathcal{R}_O'}\Lambda_g^*(Ds+Po+q/\sqrt{n})$ satisfies $\lim_n H_n'(s)=-\inf_{o\in\mathcal{R}_O'}\Lambda_g^*(Ds+Po)$; this follows as the limiting sequence of objectives, composition of an affine map with the conjugate of log-Gaussian MGF, is convex and converges to a non-monotonic convex objective $\Lambda_g^*(Ds+Po)$. These two facts ensure that $H_n(.)$ and $H_n'(.)$ are two sequences of continuous functions that converge uniformly on the set $\mathcal{R}_S$ to the limit $H(s)=-\inf_{o\in\mathbb{R}_O'}\Lambda_g^*(Ds+Po)$. A direct application of Lemma 3 now leads to the limiting rate

$$\lim_n\frac{1}{n}\log\mathbb{E}\left[\exp(nH_n(\bar{S}_n))1_{\bar{S}_n\in\mathcal{R}_S}|\beta_E\right]$$
$$=\lim_n\frac{1}{n}\log\mathbb{E}\left[\exp(nH_n'(\bar{S}_n))1_{\bar{S}_n\in\mathcal{R}_S}|\beta_E\right]$$
$$=-\inf_{s\in\mathcal{R}_S',o\in\mathcal{R}_O'}\{\Lambda_f^*(s|\beta_E)+\Lambda_g^*(Ds+Po)\}$$
$$=-\lim_n\inf_{s\in\mathcal{R}_S',o\in\mathcal{R}_O'}\{\Lambda_f^*(s|\beta_E)+\Lambda_g^*(Ds+Po+q/\sqrt{n})\}.$$

(2) For the second part, we use the tower property of expectation to have

$$\mathbb{P}\left(\bar{O}_n\in\mathcal{R}_O'|\bar{S}_n=s\right)$$
$$=\mathbb{E}\left[\mathbb{P}(\bar{O}_{-E,n}\in\mathcal{R}_{-E}'|\bar{O}_{E,n}=o_E,\bar{S}_n=s)1_{\bar{O}_{E,n}\in\mathbb{R}_E'}|\bar{S}_n=s\right]$$

where $\bar{O}_{-E,n}$ is the vector of E coordinates of $\bar{O}_n$ and similarly, $\bar{O}_{E,n}$ is defined. Denoting

$$G_n^s(o_E)=\frac{1}{n}\log\mathbb{P}(\bar{O}_{-E,n}\in\mathcal{R}_{-E}'|\bar{O}_{E,n}=o_E,\bar{S}_n=s)=\frac{1}{n}\log\mathcal{B}(o_E,s)$$

we know that $G_n^s(o_E)$ converges to a continuous function $G^s(o_E)$ uniformly on $o_E \in \mathcal{R}'_E$ using a large deviation rate. Applying Lemma 3 gives

$$\lim_n \frac{1}{n} \log \mathbb{P}\left(\bar{O}_n \in \mathcal{R}'_O | \bar{S}_n = s\right)$$

$$= \lim_n \frac{1}{n} \log \mathbb{E}\left[\exp\left(n \cdot \frac{1}{n} \log \mathcal{B}(\bar{O}_E, s)\right) 1_{\bar{O}_E \in \mathbb{R}'_E} | \bar{S}_n = s\right]$$

$$= -\inf_{o_E \in \mathcal{R}'_E} \left\{\Lambda^*_{g_E}(D_E s + P_E o_E + q_E/\sqrt{n}) - G^s(o_E)\right\}.$$

Also, note that $G_n^s(o_E)$ is a sequence of continuous, concave functions in o_E, s converging to a concave function $G^s(o_E)$. The concavity in o_E follows from the fact that convolution of log-concave densities with a log-concave indicator preserves concavity. Thus, we have

$$\inf_{o_E \in \mathcal{R}'_E} \left\{\Lambda^*_{g_E}(D_E s + P_E o_E + q_E/\sqrt{n}) - G^s(o_E)\right\}$$

$$= \lim_n \inf_{o_E \in \mathcal{R}'_E} \left\{\Lambda^*_{g_E}(D_E s + P_E o_E + q_E/\sqrt{n}) - G_n^s(o_E)\right\}$$

Denoting $H_n(s) = -\inf_{o_E \in \mathcal{R}'_E} \left\{\Lambda^*_{g_E}(D_E s + P_E o_E + q_E/\sqrt{n}) - G_n^s(o_E)\right\}$, we finally note that $H_n(.)$ is a sequence of continuous functions with a uniform limit $H(.)$ on $\mathcal{R}_S$. This completes the proof of the second part with similar arguments as the first part of the theorem. $\square$

Proof of Lemma 3

Proof. Using the uniform convergence of $H_n(.)$ to $H(.)$ on compact $\mathcal{R} \in \mathbb{R}^d$, we have for any $x \in \mathcal{R}^0$ such that $\Lambda^*(x) - H(x) < \infty$ and $\delta > 0$ such that $\mathcal{B}(x, \delta) \subset \mathcal{R}^0$

$$\liminf_n \frac{1}{n} \log \mathbb{E}[\exp(n H_n(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{R}}]$$

$$= \liminf_n \frac{1}{n} \log \mathbb{E}\left[\exp(n(H_n(\bar{Z}_n) - H(\bar{Z}_n)) \exp(n H(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{R}}\right]$$

$$\geq -\lim_n \sup_{z \in \mathcal{R}} |H_n(z) - H(z)| + \liminf_n \frac{1}{n} \log \mathbb{E}\left[\exp(n H(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{R}^0}\right]$$

$$\geq \liminf_n \frac{1}{n} \log \mathbb{E}\left[\exp(n H(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{B}(x, \delta)}\right].$$

Let $C > 0$ be chosen such that it satisfies for $x \in \mathcal{R}^0$

$$C > \Lambda^*(x) - H(x) + \liminf_n \frac{1}{n} \log \mathbb{E}\left[\exp(n H(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{B}^c(x, \delta)}\right]. \tag{A.2}$$

where $\mathcal{B}^c(x, \delta) = \{z \in \mathbb{R}^d : z \notin \mathcal{B}(x, \delta)\}$. Defining a continuous bounded function $\Psi(.)$ as $\Psi(y) = C \cdot \min\left(\frac{d(x, y)}{\delta}, 1\right)$

$$\mathbb{E}\left[\exp(-n\Psi(\bar{Z}_n) + n H(\bar{Z}_n))\right] \leq \exp(-nC) \cdot \mathbb{E}\left[\exp(n H(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{B}^c(x, \delta)}\right]$$

$$+ \mathbb{E}\left[\exp(n H(\bar{Z}_n)) 1_{\bar{Z}_n \in \mathcal{B}(x, \delta)}\right].$$

Using Varadhan's limit lemma that states for $\bar{Z}_n$ satisfying a large deviation principle with a rate function $\Lambda^*(.)$

$$\lim_n \frac{1}{n} \mathbb{E}[\exp(-n\Psi(\bar{Z}_n) + nH(\bar{Z}_n))] = -\inf_z \{\Lambda^*(z) + \Psi(z) - H(z)\},$$

and the above bound for $\mathbb{E}\left[\exp(-n\Psi(\bar{Z}_n) + nH(\bar{Z}_n))\right]$, we have

$$\max\left(-C + \frac{1}{n} \liminf_n \mathbb{E}\left[\exp(nH(\bar{Z}_n))1_{\bar{Z}_n \in \mathcal{B}^c(x,\delta)}\right],\right.$$
$$\left.\frac{1}{n} \liminf_n \mathbb{E}\left[\exp(nH(\bar{Z}_n))1_{\bar{Z}_n \in \mathcal{B}(x,\delta)}\right]\right)$$
$$\geq \liminf_n \frac{1}{n} \log \mathbb{E}\left[\exp(-n\Psi(\bar{Z}_n) + nH(\bar{Z}_n))\right]$$
$$= -\inf_z \{\Lambda^*(z) + \Psi(z) - H(z)\} \geq -\Lambda^*(x) + H(x).$$

Due to the choice of C in (A.2), we can complete the proof of the lower bound on the limit of infimums by observing

$$\liminf_n \frac{1}{n} \log \mathbb{E}[\exp(nH_n(\bar{Z}_n))1_{\bar{Z}_n \in \mathcal{R}}] \geq \liminf_n \frac{1}{n} \log \mathbb{E}\left[\exp(nH(\bar{Z}_n))1_{\bar{Z}_n \in \mathcal{B}(x,\delta)}\right]$$
$$\geq -\Lambda^*(x) + H(x)$$
$$\geq -\inf_{z \in \mathcal{R}^0} \{\Lambda^*(z) - H(z)\}$$
$$= -\inf_{z \in \mathcal{R}} \{\Lambda^*(z) - H(z)\}.$$

To prove the upper bound, let $\phi_j = j\min(d(z,\mathcal{R}),1)$ be a sequence of bounded continuous functions increasing to $\chi_{\mathcal{R}}(.)$, the characteristic function of $\mathcal{R}$. Again using the uniform convergence of $H_n(.)$ to $H(.)$ on $\mathcal{R}$ and Varadhan's limit lemma for continuous bounded function $\phi_j(.)$, we have

$$\limsup_n \frac{1}{n} \log \mathbb{E}[\exp(nH_n(\bar{Z}_n))1_{\bar{Z}_n \in \mathcal{R}})]$$
$$\leq \limsup_n \frac{1}{n} \log \mathbb{E}[\exp(n(H_n(\bar{Z}_n) - \phi_j(\bar{Z}_n)))]$$
$$\leq \limsup_n \frac{1}{n} \log \mathbb{E}[\exp(n\sup_{z \in \mathcal{R}} |H_n(z) - H(z)| + nH(\bar{Z}_n) - n\phi_j(\bar{Z}_n))]$$
$$= \limsup_n \sup_{z \in \mathcal{R}} |H_n(z) - H(z)| + \lim_n \frac{1}{n} \log \mathbb{E}[\exp(nH(\bar{Z}_n) - n\phi_j(\bar{Z}_n))]$$
$$= -\inf_z \{\Lambda^*(z) - H(z) + \phi_j(z)\}$$
$$\leq -\liminf_{j\to\infty} \inf_z \{\Lambda^*(z) - H(z) + \phi_j(z)\} \leq -\inf_{z \in \mathcal{R}} \{\Lambda^*(z) - H(z)\}.$$

The penultimate step follows by applying Varadhan's lemma for a continuous, bounded function $H(.) - \phi_j(.)$. The proof is thus complete by showing

$$\liminf_{j\to\infty} \inf_z \{\Lambda^*(z) - H(z) + \phi_j(z)\} \geq \inf_{z \in \mathcal{R}} \{\Lambda^*(z) - H(z)\}.$$

Since $\phi_j(.) = 0$ on $\mathcal{R}$, it suffices to prove

$$\liminf_{j\to\infty} \inf_{z\in\mathcal{R}^c} \{\Lambda^*(z) - H(z) + \phi_j(z)\} \geq \inf_{z\in\mathcal{R}} \{\Lambda^*(z) - H(z)\} = \mathcal{L}.$$

Suppose the above claim is untrue. Then, there exists a subsequence j_k, $z_{j_k} \in \mathcal{R}^c$ and $0 < \epsilon < \mathcal{L}/2$ such that $\Lambda^*(z_{j_k}) - H(z_{j_k}) + \phi_{j_k}(z_{j_k}) \leq \mathcal{L} - \epsilon$. This would in turn imply $d(z_{j_k}, \mathcal{R}) \to 0$ as $k \to \infty$. This also means $\sup_k \Lambda^*(z_{j_k}) - H(z_{j_k}) \leq \mathcal{L} - \epsilon$. There exists a z^* such that $\Lambda^*(z^*) - H(z^*) \leq \mathcal{L} - \epsilon$ and a further subsequence such that $d(z_{j_{k_l}}, z^*) \to 0$. For this subsequence, we know that we can construct a sequence $y_{j_{k_l}} \in \mathcal{R}$ such that $d(z_{j_{k_l}}, y_{j_{k_l}}) \to 0$. This would imply that $d(y_{j_{k_l}}, z^*) \to 0$ and the consequence of this is that $z^* \in \mathcal{R}$ which shall lead to the contradiction $\Lambda^*(z^*) - H(z^*) > \mathcal{L} = \inf_{z\in\mathcal{R}} \{\Lambda^*(z) - H(z)\}$. $\qquad\square$

Appendix B: Conjugates of barrier function

The barrier function for the canonical sign and cube constraints that we use in our implementations are $b_{\mathcal{R}_{j,E}}(o_{j,E}) = \log(1 + \frac{1}{s_{j,E}o_{j,E}})$ and

$$b_{\mathcal{R}_{j,-E}}(o_{j,-E}) = \log\left(1 + \frac{1}{\lambda - o_{j,-E}}\right) + \log\left(1 + \frac{1}{\lambda + o_{j,-E}}\right)$$

respectively. The conjugate for the sign barrier function at $P_{j,E}^T u$ can be computed as $b_{\mathcal{R}_{j,E}}^*(v) = \sup_{z_j} z_j v - \log(1 + \frac{1}{s_{j,E}z_j})$ at $v = P_{j,E}^T u$. The optimal z_j^* maximizing the above optimization problem is given by

$$z_j^* = \begin{cases} -\dfrac{1}{2} + \sqrt{\dfrac{1}{4} - \dfrac{1}{v}} & \text{if } v < 0 \\ \text{no root} & \text{otherwise} \end{cases}$$

whenever $s_{j,E} = 1$ and whenever $s_{j,E} = -1$

$$z_j^* = \begin{cases} \dfrac{1}{2} - \sqrt{\dfrac{1}{4} + \dfrac{1}{v}} & \text{if } v > 0 \\ \text{no root} & \text{otherwise.} \end{cases}$$

The conjugate for the cube barrier that reflect the inactive constraints does not have a closed form expression. Yet, we may employ an easy binary search method to find the roots of $p - |E|$ univariate separable conjugate problems

$$b_{\mathcal{R}_{j,-E}}^*(v) = \max_{z_j \in [-\lambda,\lambda]} v z_j - b_{\mathcal{R}_{j,-E}}(z_j) \text{ at } v = P_{j,-E}^T u.$$

We may alternatively choose a log-barrier on inactive coordinates in which case the barrier function is given by $b_{\mathcal{R}_{j,-E}}(o_{j,-E}) = -\log(\lambda - o_{j,-E}) - \log(\lambda + o_{j,-E})$ whose conjugate has an explicit form of roots within $[-\lambda, \lambda]$. That is, solving the optimal point z_j^* that yields

$$b_{\mathcal{R}_{j,-E}}^*(v) = \max_{z_j \in [-\lambda,\lambda]} v z_j + \log(\lambda - z_j) + \log(\lambda + z_j) \text{ is given by}$$
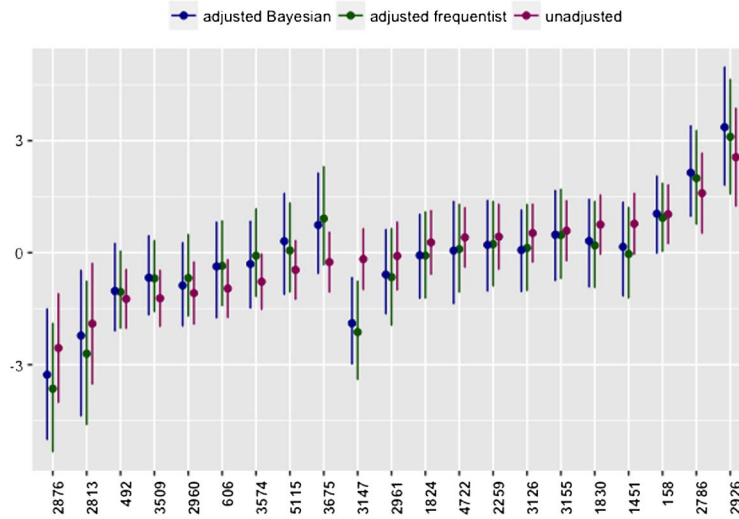
FIG 7. *Effect size estimates: adjusted posterior mean and credible intervals based on the truncated and unadjusted Bayesian posterior; selective MLE and confidence intervals are based on the sampler free approach in [21]. The Bayesian estimates, both posterior mean and intervals mimic the frequentist estimates.*

$$
z_j^* = \begin{cases} -\dfrac{1}{v} + \sqrt{\dfrac{1}{v^2} + \lambda^2} & \text{if } v > 0 \\[3ex] -\dfrac{1}{v} - \sqrt{\dfrac{1}{v^2} + \lambda^2} & \text{if } v < 0. \end{cases}
$$

B.1. Supplementary to real data analysis

Details of data: We give below the details of the gene expression data set analyzed in 6.2 under Section 6. The data set consists of an outcome variable that represents gene expression levels of a gene with ID **"ENSG00000131697.13"**, sampled for 97 individuals from the tissue of Liver. Both the gene expression outcome and the genotypic data consisting of local variants measured within 1 MB up and downstream from the transcription gene site are a part of the GTEx project https://www.gtexportal.org/home/. The DNA genotyping on blood-derived DNA samples of these individuals was performed at the GTEx Laboratory Data Analysis and Coordination Center (LDACC) at the Broad Institute. More details on the sample procurement, gene, variants inclusion and reads of gene-level expression can be found in the papers [6, 1].

Comparison of the Bayesian estimates with frequentist approach: We have seen in Section 6 that the Bayesian estimates under a flat prior and modeled along the conditional approach have good frequentist properties like coverage

and risk. For interested readers, we also compute the adjusted frequentist estimates using the methods in [21]. The mentioned paper uses a sampler free approach to solve for an intractable pivot and provides adjusted intervals and an approximate selective MLE based on the truncated likelihood, conditioned on the selection event. For the discovered SNPs post a randomized Lasso on prototype SNPs as described in 6.2, we plot the frequentist estimates alongside the Bayesian estimates under the diffuse prior. The fact that the Bayesian intervals and posterior mean mimic the frequentist intervals and selective MLE validates that our Bayesian approach displays the Bernstein von Mises phenomenon as the unadjusted Bayesian estimates do.

References

[1] Francois Aguet, Andrew A Brown, Stephane Castel, Joe R Davis, Pejman Mohammadi, Ayellet V Segre, Zachary Zappala, Nathan S Abell, Laure Fresard, Eric R Gamazon, et al. Local genetic effects on gene expression across 44 human tissues. *bioRxiv 074450*, 2016.

[2] Sio Iong Ao, Kevin Yip, Michael Ng, David Cheung, Pui-Yee Fong, Ian Melhado, and Pak C Sham. Clustag: hierarchical clustering and graph methods for selecting tag snps. *Bioinformatics*, 21(8):1735–1736, 2004.

[3] Jacob Bien and Robert Tibshirani. Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, 106 (495):1075–1084, 2011. MR2894765

[4] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pages 289–296, 2009.

[5] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011. MR2786918

[6] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[7] Amir Dembo and Ofer Zeitouni. Large deviations techniques and applications second edition. *Large deviations techniques and applications*, 38, 1998. MR1619036

[8] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 117–126. ACM, 2015. MR3388189

[9] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal Inference After Model Selection. *arXiv preprint arXiv:1410.2597*, October 2014.

[10] Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

[11] Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013. MR3081926

[12] JT Gene Hwang and Zhigen Zhao. Empirical bayes confidence intervals for selected parameters in high-dimensional data. *Journal of the American Statistical Association*, 108(502):607–618, 2013. MR3174645

[13] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference with the lasso. *The Annals of Statistics*, 44(3): 907–927, November 2016. URL http://projecteuclid.org/euclid.aos/1460381681. MR3485948

[14] Joshua R Loftus and Jonathan E Taylor. A significance test for forward stepwise model selection. May 2014. URL http://xxx.tau.ac.il/abs/1405.3920v1.

[15] Jelena Markovic and Jonathan Taylor. Bootstrap inference after using multiple queries for model selection. *arXiv preprint arXiv:1612.07811*, 2016.

[16] Thomas P Minka. *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology, 2001. MR2717007

[17] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404): 1023–1032, 1988. MR0997578

[18] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[19] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2015.

[20] Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*, 2016.

[21] Snigdha Panigrahi, Jelena Markovic, and Jonathan Taylor. An mcmc free approach to post-selective inference. *arXiv preprint arXiv:1703.06154*, 2017.

[22] Stephen Reid and Robert Tibshirani. Sparse regression and marginal testing using cluster prototypes. *Biostatistics*, 17(2):364–376, 2016. MR3516006

[23] Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996. MR1440273

[24] Jonathan Taylor, Joshua Loftus, and Ryan Tibshirani. Tests in adaptive regression via the Kac-Rice formula. *The Annals of Statistics*, 44(2): 743–770, August 2013. URL http://projecteuclid.org/euclid.aos/1458245734. MR3476616

[25] Xiaoying Tian and Jonathan E. Taylor. Selective inference with a randomized response. *arXiv preprint arXiv:1507.06739*, July 2015. MR3782381

[26] Xiaoying Tian, Snigdha Panigrahi, Jelena Markovic, Nan Bi, and Jonathan Taylor. Selective sampling after solving a convex problem. *arXiv preprint arXiv:1609.05609*, 2016.

[27] Ryan Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Post-selection adaptive inference for Least Angle Regression and the

Lasso. *arXiv preprint arXiv:1401.3889*, January 2014. MR2634010

[28] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. MR3538689

[29] Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477, 2016.

[30] Daniel Yekutieli. Adjusted Bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541, 2012. MR2925372

[31] Zhigen Zhao and JT Gene Hwang. Empirical bayes false coverage rate controlling confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(5):871–891, 2012. MR2988910

[32] Zhigen Zhao and Sanat K Sarkar. A Bayesian approach to constructing multiple confidence intervals of selected parameters with sparse signals. *Statistica Sinica*, pages 725–741, 2015. MR3379096

[33] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2): 301–320, 2005. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/abstract. MR2137327