Biometrika (2024), **00**, 0, *pp.* 1–19 https://doi.org/10.1093/biomet/asae019
Advance Access publication 3 April 2024

Exact selective inference with randomization

By SNIGDHA PANIGRAHI

Department of Statistics, University of Michigan,
1085 South University Avenue, Ann Arbor, Michigan 48109, U.S.A.
psnigdha@umich.edu

KEVIN FRY and JONATHAN TAYLOR

Department of Statistics, Stanford University,
Sequoia Hall, 390 Jane Stanford Way, Stanford, California 94305, U.S.A.
kfry@stanford.edu jonathan.taylor@stanford.edu

Summary

We introduce a pivot for exact selective inference with randomization. Not only does our pivot lead to exact inference in Gaussian regression models, but it is also available in closed form. We reduce this problem to inference for a bivariate truncated Gaussian variable. By doing so, we give up some power that is achieved with approximate maximum likelihood estimation in Panigrahi & Taylor (2023). Yet our pivot always produces narrower confidence intervals than a closely related data-splitting procedure. We investigate the trade-off between power and exact selective inference on simulated datasets and an HIV drug resistance dataset.

Some key words: Data carving; Data splitting; Exact inference; Pivot; Post-selection inference; Randomization; Selective inference.

1. Introduction

The polyhedral method of Lee et al. (2016) introduced confidence intervals for exact selective inference in Gaussian regression models. This method provides valid inferences for selected parameters by conditioning on the outcome of selection. A pivot is obtained for each selected parameter from a truncated Gaussian distribution, provided the outcome of selection can be described by linear constraints, also known as polyhedral constraints. However, as shown by Kivaranovic & Leeb (2021), confidence intervals based on this pivot can have infinite length in expectation.

Randomizing data at the time of selection and conditioning on the outcome of randomized selection produces narrower confidence intervals than the polyhedral method. Kivaranovic & Leeb (2024) formally established that some of these randomized procedures guarantee intervals with bounded lengths. A stumbling block for subsequent inference, however, is the lack of a pivot in closed form after marginalizing over the added randomization variables. For example, the pivot based on randomized response, as in Tian & Taylor (2018), or on data carving, which involves holding out a random subsample during selection, as in Fithian et al. (2017), cannot be directly computed.

Recent work by Panigrahi & Taylor (2023) bypassed this computational hurdle by proposing an approximate Gaussian pivot through maximum likelihood estimation. The

Table 1. *Coverage probability and average lengths of intervals for the MLE and our proposed method*

	(a) Coverage		(b) Length	
t	MLE (%)	Exact (%)	MLE	Exact
0.5	86.09	89.94	20.83	27.14
0.75	86.69	90.02	20.86	26.94
1	85.56	90.30	21.09	27.44

approximate pivot is obtained by solving a convex optimization problem that yields the selection-adjusted maximum likelihood estimator and observed Fisher information matrix. The term we use for this approach is the MLE method. Although computationally appealing, this pivot may not provide adequate coverage if the approximation is inaccurate. Moreover, it can be difficult to determine the reliability of the approximation in practical settings. Inaccuracies can arise when the dimensions of the problem are significantly larger than the number of available samples. To provide an example, consider the case where $n = 500$ independent and identically distributed samples are generated from a Gaussian linear regression model with $p = 1000$ predictors, of which 25 are true signals with magnitude of $(2t \log p)^{1/2}$ and the rest are noise. We conduct 500 rounds of simulations with t taking values 0.5, 0.75 and 1. In all three scenarios, the coverage probability of the approximate pivot produced by the MLE method is below the target level 0.90, as reported in Table 1(a).

In this paper, we offer a new pivot for selective inference with randomization. We aim at exact selective inference in closed form, without requiring a case-by-case treatment for different models. In exchange, we give up some power that is achieved with the approximate Gaussian pivot in Panigrahi & Taylor (2023). This trade-off between the coverage probabilities and the averaged lengths of the intervals for both methods, MLE and our proposed method Exact, can be seen in Table 1(a) and (b). Despite sacrificing some power, our pivot produces more reliable inferences that roughly attain the target coverage probability 0.90 in all three scenarios.

2. Background

2.1. *Some preliminaries*

We begin by defining notation that is used throughout the paper. Let $[d] = \{1, 2, \ldots, d\}$ for $d \in \mathbb{N}$. The symbol $e_j \in \mathbb{R}^d$ is understood as a vector with 1 in the jth entry and 0 elsewhere. For $\eta \in \mathbb{R}^d$ and $\Theta \in \mathbb{R}^{d\times d}$, $\eta_j = e_j^{\mathrm{T}}\eta$ is the jth entry of η, $\Theta_{j,k} = e_j^{\mathrm{T}}\Theta e_k$ is the (j,k)th entry of Θ and $\Theta_{[j]}$ is the jth row of Θ. For a given set D, the notation $|D|$ represents its cardinality.

We use $\phi(x; \theta, \Theta)$ to denote the density function of a Gaussian variable with the mean vector $\theta \in \mathbb{R}^d$ and covariance matrix $\Theta \in \mathbb{R}^{d\times d}$ at x. In particular, when $d = 1$, $\theta = 0$, $\Theta = 1$, we let $\phi(x)$ be the density of a standard normal variable and let $\Phi(x)$ be its cumulative distribution function. Denote by

$$\mathrm{TP}^{[a,b]}(\theta, \vartheta) = \Phi\left\{\frac{1}{\vartheta}(b-\theta)\right\} - \Phi\left\{\frac{1}{\vartheta}(a-\theta)\right\}$$

the truncation probability that a univariate Gaussian variable with mean θ and variance ϑ^2 lies in the interval $[a, b]$, where a, b take values in the extended real set.

For background on selective inference, we consider the standard setting of the lasso regression with a fixed design matrix. Suppose that we have a vector of outcomes $y \sim \mathcal{N}(\mu, \sigma^2 I_n) \in \mathbb{R}^n$ for an unknown mean parameter μ and a matrix of p fixed features $X \in \mathbb{R}^{n \times p}$. We observe $w \sim \mathcal{N}(0_p, \Omega)$, a p-dimensional randomization variable that is drawn independently of y. Consider solving

$$\hat{b} = \underset{b \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y - Xb\|_2^2 + \frac{\epsilon}{2}\|b\|_2^2 + \lambda\|b\|_1 - w^{\mathrm{T}} b \tag{1}$$

with regularization parameter $\lambda \in \mathbb{R}^+$.

The selection algorithm in (1) gives a noisy version of the lasso, which is called the randomized lasso in Tian et al. (2016). A small, fixed value of $\epsilon \in \mathbb{R}^+$ in the objective of the randomized lasso simply ensures the existence of a solution. The variance of the Gaussian randomization variable is a tuning parameter that is similar to the split proportion in data splitting. It lets us control how much information we use to select a model versus how much we use for inference. As an example, consider $\Omega(\tau^2) = \tau^2 I_p$. If we increase the value of τ^2, it means that we perform a noisier model selection, which reserves more information for inference. Later in the paper, we discuss incorporating a Gaussian randomization scheme that is related to data splitting.

After solving (1), we seek inference for a set of post-selection parameters. Here is a common example. Let

$$E = \{j \in [p] \colon |\mathrm{sign}(\hat{b}_j)| = 1\}.$$

Having observed the selected subset of features $E = \mathcal{E}$, we infer for

$$\beta^{\mathcal{E}} = (X_{\mathcal{E}}^{\mathrm{T}} X_{\mathcal{E}})^{-1} X_{\mathcal{E}}^{\mathrm{T}} \mu \in \mathbb{R}^{|\mathcal{E}|},$$

which is the best linear representation of μ using the selected subset of features $X_{\mathcal{E}}$. For brevity, let $c^j = X_{\mathcal{E}}(X_{\mathcal{E}}^{\mathrm{T}} X_{\mathcal{E}})^{-1} e_j \in \mathbb{R}^n$ for $j \in [|\mathcal{E}|]$. This allows us to write each entry of $\beta^{\mathcal{E}}$ as

$$\beta_j^{\mathcal{E}} = {c^j}^{\mathrm{T}} \mu.$$

Note that $\beta_j^{\mathcal{E}}$ depends on y and w through c^j, which in turn depends on $\mathcal{E}$.

2.2. *Existing work*

We begin by reviewing two existing methods that are closely related to our current proposal. The first method offers an exact pivot for selective inference when solving the standard version of lasso, without randomization. The second method provides an approximate pivot after solving the randomized lasso.

Both pivots are obtained from a conditional distribution of the outcome variable after conditioning on a proper subset of the observed event. Conditioning on $\{E = \mathcal{E}\}$ is ideal if we wanted inference for $\beta^{\mathcal{E}}$. However, the ideal event is usually complicated to describe in terms of y and w, making the conditional distribution of y given $\{E = \mathcal{E}\}$ less amenable to inferences. Therefore, conditioning on a subset of the selection event that has a simpler description is a practical solution, which can ensure valid and feasible selective inference.

First we review the polyhedral method. Consider solving the standard lasso (Tibshirani, 1996), which involves setting $\epsilon = 0$ and $w = 0_p$ in the objective of (1). We denote the set of selected features as E_0. We distinguish E_0 from the selected set E, which is obtained from solving the randomized lasso.

Having observed $E_0 = \mathcal{E}_0$, fix $c_0^j = X_{\mathcal{E}_0}(X_{\mathcal{E}_0}^{\mathrm{T}} X_{\mathcal{E}_0})^{-1} e_j \in \mathbb{R}^n$ that leads to

$$\beta_j^{\mathcal{E}_0} = {c_0^j}^{\mathrm{T}} \mu,$$

our parameters post selection. Let $S_0 \in \mathbb{R}^{|E_0|}$ be the vector of nonzero signs. Let $\hat{\beta}^{\mathcal{E}_0}$ denote the least-squares estimator when we regress y against $X_{\mathcal{E}_0}$ and let

$$\hat{\Gamma}_0^j = \left(I - \frac{c_0^j {c_0^j}^{\mathrm{T}}}{\|c_0^j\|_2^2} \right) y$$

be the projection of y onto the orthogonal complement of the subspace spanned by c_0^j.

Conditional on $E_0 = \mathcal{E}_0$, $S_0 = \mathcal{S}_0$ and the value of $\hat{\Gamma}_0^j$, the polyhedral method of Lee et al. (2016) gives an exact pivot by truncating a univariate Gaussian variable, with mean $\beta_j^{\mathcal{E}_0}$ and variance $\sigma^2 \|c_0^j\|_2^2$, to an interval $[H_-^j, H_+^j]$. The pivot takes the form

$$\mathcal{P}_{\mathrm{Poly}}^j(\beta_j^{\mathcal{E}_0}) = \frac{\int_{-\infty}^{\hat{\beta}_j^{\mathcal{E}_0}} \phi\{(\sigma\|c_0^j\|_2)^{-1}(x - \beta_j^{\mathcal{E}_0})\} \cdot 1_{[H_-^j, H_+^j]}(x)\, \mathrm{d}x}{\int_{-\infty}^{\infty} \phi\{(\sigma\|c_0^j\|_2)^{-1}(x - \beta_j^{\mathcal{E}_0})\} \cdot 1_{[H_-^j, H_+^j]}(x)\, \mathrm{d}x}, \tag{2}$$

where the expressions for H_-^j and H_+^j depend on $\mathcal{E}_0$, $\mathcal{S}_0$ and $\hat{\Gamma}_0^j$.

Next, we turn to selective inference with the randomized lasso. The approximate MLE method of Panigrahi & Taylor (2023) uses the likelihood of y when conditioned on $\{G = \mathcal{G}\}$, where $G = \partial_{\hat{b}} \|b\|_1$ is the subgradient of the ℓ_1 penalty at the randomized lasso solution. Similar to the polyhedral method, the conditioning event is a proper subset of the ideal event $\{E = \mathcal{E}\}$.

Let $\hat{b}^{\mathcal{E}}$ and $\hat{I}^{\mathcal{E}}$ denote the MLE and the observed Fisher information matrix in this conditional likelihood. An approximate Gaussian pivot for $\beta_j^{\mathcal{E}}$ is given by

$$\mathcal{P}_{\mathrm{MLE}}^j(\beta_j^{\mathcal{E}}) = \Phi\left\{ \frac{1}{\sqrt{(\hat{I}^{\mathcal{E}})_{j,j}^{-1}}} (\hat{b}_j^{\mathcal{E}} - \beta_j^{\mathcal{E}}) \right\}.$$

Equivalently, confidence intervals for each component of $\beta^{\mathcal{E}}$ are calculated by centring them around the jth entry of the MLE, with the variance estimated by the corresponding diagonal entry of the observed Fisher information matrix. However, the seemingly simple Gaussian pivot involves computing the exact conditional likelihood function, which cannot be done in closed form, hence making it difficult to compute the two estimators. To overcome this, the approximate MLE method derives approximate values for $\hat{b}^{\mathcal{E}}$ and $\hat{I}^{\mathcal{E}}$, which rely on a consistent approximation to the exact conditional likelihood.

2.3. *Toy example*

We can now informally present the central idea of our paper using a toy example with two features, i.e., $p = 2$. We solve (1) with $\epsilon = 0$ and $w \sim \mathcal{N}(0_2, \Omega) \in \mathbb{R}^2$, where $\Omega = \tau^2 X^{\mathrm{T}} X$.

Say that we select the full model, i.e., $\mathcal{E} = \{1, 2\}$, and that we focus on the first component of the two-dimensional post-selection parameter

$$\beta_1^{\mathcal{E}} = c^{1^{\mathrm{T}}}\mu.$$

Let $\hat{\beta}^{\mathcal{E}}$ be the least-squares estimator when regressing y against $X_{\mathcal{E}}$ and let $\hat{\beta}_1^{\mathcal{E}}$ be its first component.

Introducing some additional notation, let $O \in \mathbb{R}^2$ denote the nonzero randomized lasso solution in this example. Let $S = \mathrm{sign}(O)$ be the corresponding sign vector, and let $\mathcal{S}$ be the observed value of S. Recall that the existing MLE method makes inferences after conditioning on the event $\{G = \mathcal{G}\}$. Since $\mathcal{G} = \mathcal{S}$ in this example, it is easy to see that this conditioning event can be described as

$$\{-\mathrm{diag}(\mathcal{S})O < 0_2\}. \tag{3}$$

While the previously mentioned MLE method obtains an approximate Gaussian pivot with this conditioning event, we can simplify the event by conditioning on some additional information that reduces the conditioning event to an interval on the real line. This is the central idea behind constructing an exact pivot in closed form.

In this specific toy example, we condition on

$$A = O_2 - \frac{e_1^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}e_2}{e_1^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}e_1}O_1,$$

in addition to conditioning on the value of G. Because

$$O = \left\{\frac{1}{e_1^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}e_1}(X^{\mathrm{T}}X)^{-1}e_1\right\}O_1 + \begin{pmatrix}0\\ A\end{pmatrix},$$

the initial conditioning event in (3) simplifies to

$$\{I_-^1 \leqslant O_1 \leqslant I_+^1\}$$

after conditioning on A, where $[I_-^1, I_+^1]$ is a fixed interval. Consequently, we can obtain an exact pivot in closed form by computing the joint bivariate distribution of $\hat{\beta}_1^{\mathcal{E}}$ and O_1 when truncated to the region $\mathbb{R} \times [I_-^1, I_+^1]$. We explain our choice for additional conditioning and derive a bivariate truncated Gaussian distribution in the next section.

To conclude, the difference between our method and the polyhedral method is shown in Fig. 1. For drawing selective inference, the polyhedral method truncates the Gaussian distribution of $\hat{\beta}_1^{\mathcal{E}}$ to the interval $[H_-^1, H_+^1]$, while our method truncates the joint bivariate distribution of $\hat{\beta}_1^{\mathcal{E}}$ and O_1 to $\mathbb{R} \times [I_-^1, I_+^1]$.

2.4. *Connections with other work*

Several papers have demonstrated the effectiveness of the conditional approach for selective inference across various problems, as evidenced by Lee & Taylor (2014), Yang et al. (2016), Suzumura et al. (2017), Charkhi & Claeskens (2018), Hyun et al. (2018), Zhao & Panigrahi (2019), Chen & Bien (2020), Duy et al. (2020), Tanizaki et al. (2020) and Gao et al. (2024). A significant focus of the current research in this field is on enhancing the

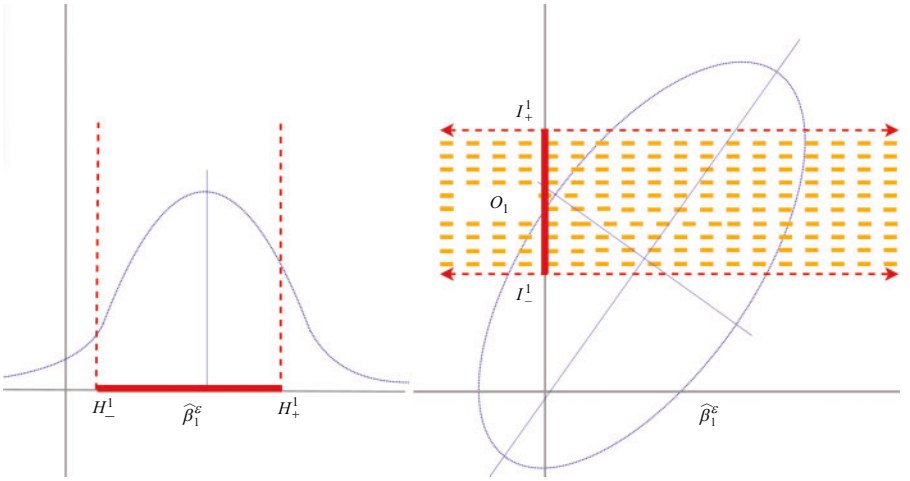

Fig. 1. Comparison with the polyhedral method.

power of earlier approaches. Before we discuss these improvements, it is worth noting that two other approaches to selective inference have been studied in parallel.

The first is the simultaneous inference approach, which has been investigated by Berk et al. (2013) and Bachoc et al. (2020). This approach is not customized to a particular selection method, but the downside is that the confidence intervals are relatively long and may not permit easy calculations in some instances. The second approach is data splitting. This method allows for valid selective inference when the available data can be split into two independent sets. One set is used as training data for the selection process, while the other set is held out as validation data for selective inference. Combined with the bootstrap in regression models, Rinaldo et al. (2019) conducted selective inference by splitting the sample space. Recently, new forms of data splitting have been introduced by Leiner et al. (2023), Neufeld et al. (2023) and Rasines & Young (2023), which split each observation into two parts to construct a training set for selection and a validation set for selective inference. However, these variants of data splitting lose power by discarding data used in selection. In our simulations, we confirm that inverting our pivot results in narrower confidence intervals than two such forms of data splitting.

There are two main branches of the conditional approach that have improved power and overcome the limitations of the polyhedral approach. The first branch of work involves choosing a minimal conditioning set that can be achieved in some special settings. For example, Liu et al. (2018) conditioned on strictly less information than the polyhedral method when inference is based on a full linear model $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. In the saturated model $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, Le Duy & Takeuchi (2022) applied parametric programming to avoid conditioning on the signs of the lasso coefficients and Carrington & Fearnhead (2024) conditioned on less information to provide inference for detected changepoints. The second branch of work utilizes randomization variables at the time of selection to remedy a loss in power. Some of these randomized procedures can be viewed as a more efficient alternative to data splitting and appear as data carving in existing literature (Fithian et al., 2017; Schultheiss et al., 2021; Panigrahi, 2023). Randomization variables have been used to deliver powerful Bayesian inference after model selection in papers by Panigrahi et al. (2021, 2023a,b). Our work falls into the latter category, where we provide a principled approach

to choose a conditioning event and construct a pivot thereof that can work with different Gaussian regression models after selection.

It is not a new idea to find a conditioning event that can lead to a bivariate truncated distribution. In Kivaranovic & Leeb (2024), one such construction is noted, where noise is added to a Gaussian response as proposed by Tian & Taylor (2018). This work achieved an exact pivot by conditioning on the projection of the noisy response onto the orthogonal complement of the subspace spanned by the direction vector of interest. However, in our paper, we have employed a different randomization scheme, which involves adding noise to the optimization objective. As demonstrated by Huang et al. (2023), this scheme has the potential to be applied to a broad range of M-estimation problems, not just the least-squares estimation problem. For example, while adding Gaussian noise to a binary response in logistic regression might not be meaningful, adding noise to the loglikelihood case would create a noisy estimation problem. Although our primary focus in this paper is on exact selective inference, our pivot is likely to generalize and provide asymptotic inference in a more comprehensive context. The concluding discussion in our paper includes a remark on this.

3. Exact selective inference with the lasso

3.1. *Conditioning event*

We continue using the randomized lasso to explain our approach in the general Gaussian regression setting.

Defining some notation, we denote by $O \in \mathbb{R}^{|\mathcal{E}|}$ the active nonzero lasso solution, and by $S = \text{sign}(O)$ the associated sign vector. Throughout, we assume that the active components of the lasso solution are stacked before its inactive components. The p-dimensional subgradient of the ℓ_1 penalty at the randomized lasso solution is denoted by $G = \binom{S}{U}$, where $U \in \mathbb{R}^{p-|\mathcal{E}|}$ collects the components of the subgradient subvector in $\mathcal{E}^c$. To represent the realized values of the variables O, S and U, we use the symbols $\mathcal{O}$, $\mathcal{S}$ and $\mathcal{U}$, respectively.

At the randomized lasso solution, observe that

$$w = Py + Q\mathcal{O} + R\mathcal{U} + T,$$

where

$$P = -\begin{bmatrix} X_{\mathcal{E}}^{\mathrm{T}} \\ X_{\mathcal{E}^c}^{\mathrm{T}} \end{bmatrix}, \qquad Q = \begin{bmatrix} X_{\mathcal{E}}^{\mathrm{T}} X_{\mathcal{E}} + \epsilon I_{|\mathcal{E}|} \\ X_{\mathcal{E}^c}^{\mathrm{T}} X_{\mathcal{E}} \end{bmatrix}, \qquad R = \begin{bmatrix} 0_{|\mathcal{E}|,p-|\mathcal{E}|} \\ \lambda I_{p-|\mathcal{E}|} \end{bmatrix}, \qquad T = \begin{pmatrix} \lambda \mathcal{S} \\ 0_{p-|\mathcal{E}|} \end{pmatrix}, \quad (4)$$

and we have assumed that the active components are stacked before the inactive ones in our matrices.

As demonstrated in the previous section, we first identify a conditioning event that will guide us to a pivot for exact selective inference. Extending the method by Panigrahi & Taylor (2023), we condition on $\{G = \mathcal{G}\}$, which can be described as $\{LO < M,\ U = \mathcal{U}\}$ for $L = -\text{diag}(\mathcal{S})$, $M = 0_{|\mathcal{E}|}$.

To reduce our conditioning event to an interval and obtain a closed-form pivot, we condition on some more information. Proposition 1 below states this event, which is equivalent to truncating a linear combination of O to a fixed interval. To present this result, we introduce a few matrices that rely on the covariance of the randomization variables and the matrices

defined in (4). Let

$$\Theta = (Q^{\mathrm{T}}\Omega^{-1}Q)^{-1}, \qquad P^j = \frac{1}{\|c^j\|_2^2}Pc^j,$$
$$r^j = Q^{\mathrm{T}}\Omega^{-1}P^j \in \mathbb{R}^{|\mathcal{E}|}, \qquad Q^j = \frac{1}{r^{j\mathrm{T}}\Theta r^j}\Theta r^j,$$

for $j \in [|\mathcal{E}|]$.

PROPOSITION 1. *Define the variables*

$$A^{r^j} = (I_{|\mathcal{E}|} - Q^j r^{j\mathrm{T}})O \in \mathbb{R}^{|\mathcal{E}|}.$$

For $j \in [|\mathcal{E}|]$, it holds that

$$\{G = \mathcal{G},\ A^{r^j} = \mathcal{A}^{r^j}\} = \{I_-^j < r^{j\mathrm{T}}O < I_+^j,\ U = \mathcal{U},\ A^{r^j} = \mathcal{A}^{r^j}\},$$

where

$$I_-^j = \max_{k \in S_-^j}\left\{\frac{1}{L_{[k]}^{\mathrm{T}}Q^j}(M_k - L_{[k]}^{\mathrm{T}}\mathcal{A}^{r^j})\right\}, \qquad I_+^j = \min_{k \in S_+^j}\left\{\frac{1}{L_{[k]}^{\mathrm{T}}Q^j}(M_k - L_{[k]}^{\mathrm{T}}\mathcal{A}^{r^j})\right\}$$

and

$$S_-^j = \{k\colon L_{[k]}^{\mathrm{T}}\Theta r^j < 0\}, \qquad S_+^j = \{k\colon L_{[k]}^{\mathrm{T}}\Theta r^j > 0\}.$$

From the previous result, we observe that the conditioning event involves extra information in the form of A^{r^j}, representing linear combinations of the active lasso coefficients. We motivate our choice of conditioning event later. In the next section, we obtain a pivot for $\beta_j^{\mathcal{E}}$ by conditioning on the event in Proposition 1.

3.2. *Pivot*

Let $\hat{\beta}^{\mathcal{E}}$ be the least-squares estimator obtained by regressing y on $X_{\mathcal{E}}$. Specifically, let $\hat{\beta}_j^{\mathcal{E}} = (c^j)^{\mathrm{T}}y$ denote the jth entry of $\hat{\beta}^{\mathcal{E}}$. Define $\hat{\Gamma}^j$ as the projection of y onto the orthogonal complement of the subspace spanned by c^j.

Note that

$$\mu = \frac{c^j}{\|c^j\|_2^2}c^{j\mathrm{T}}\mu + \mathcal{P}_{c^j}^{\perp}\mu.$$

When inferring for $c^{j\mathrm{T}}\mu$, the projection $\mathcal{P}_{c^j}^{\perp}\mu$ includes nuisance parameters. To eliminate these parameters, we follow a similar approach as Lee et al. (2016) and condition on $\hat{\Gamma}^j$. This allows us to obtain a conditional density that involves only our parameter of interest, $\beta_j^{\mathcal{E}}$. We can then use its cumulative distribution function to obtain a pivot.

To state our main result, we introduce the functions

$$\Lambda(y,\mathcal{U}) = -P^{j\mathrm{T}}\Omega^{-1}(Py + R\mathcal{U} + T), \qquad \Delta(y,\mathcal{U}) = -\Theta Q^{\mathrm{T}}\Omega^{-1}(Py + R\mathcal{U} + T).$$

THEOREM 1. *Define the random variable*

$$\mathcal{P}^j_{\text{Exact}}(\beta_j^{\mathcal{E}}) = \frac{\int_{-\infty}^{\hat{\beta}_j^{\mathcal{E}}} \phi\{(x - \lambda^j \beta_j^{\mathcal{E}} - \zeta^j)/\sigma^j\} \cdot \text{TP}^{[I^j_-, I^j_+]}\{\theta^j(x), \vartheta^j\}\, \mathrm{d}x}{\int_{-\infty}^{\infty} \phi\{(x - \lambda^j \beta_j^{\mathcal{E}} - \zeta^j)/\sigma^j\} \cdot \text{TP}^{[I^j_-, I^j_+]}\{\theta^j(x), \vartheta^j\}\, \mathrm{d}x},$$

where the constants ϑ^j, σ^j, λ^j, ζ^j *and the univariate function* θ^j *are computed as*

$$(\vartheta^j)^2 = r^{j\mathrm{T}} \Theta r^j, \qquad (\sigma^j)^2 = \left\{ \frac{1}{\sigma^2 \|c^j\|_2^2} + P^{j\mathrm{T}} \Omega^{-1} P^j - (\vartheta^j)^2 \right\}^{-1},$$

$$\lambda^j = \frac{1}{\sigma^2 \|c^j\|_2^2} (\sigma^j)^2, \qquad \zeta^j = (\sigma^j)^2 \cdot \{\Lambda(\hat{\Gamma}^j, \mathcal{U}) - r^{j\mathrm{T}} \Delta(\hat{\Gamma}^j, \mathcal{U})\},$$

$$\theta^j(x) = r^{j\mathrm{T}} \Delta(\hat{\Gamma}^j, \mathcal{U}) - (\vartheta^j)^2 x.$$

Conditioned on the event in Proposition 1, $\mathcal{P}^j_{\text{Exact}}(\beta_j^{\mathcal{E}})$ *is distributed as a* $\text{Un}(0, 1)$ *variable.*

Inverting the pivot in Theorem 1 gives a confidence interval for $\beta_j^{\mathcal{E}}$. At a predetermined significance level α, a two-sided confidence interval for $\beta_j^{\mathcal{E}}$ is equal to

$$(L^j_\alpha, U^j_\alpha) = \left\{ b \in \mathbb{R} \colon \mathcal{P}^j_{\text{Exact}}(b) \in \left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2} \right] \right\}.$$

A few comments are in order here.

Remark 1. The choice of the simple model $y \sim \mathcal{N}(\mu, \sigma^2 I_n) \in \mathbb{R}^n$ was made for ease of presentation. However, our pivot can also be applied to other Gaussian regression models, such as the model of Fithian et al. (2017) where $y \sim \mathcal{N}(X_{\mathcal{E}} \beta_{\mathcal{E}}, \sigma^2 I_n)$ or the full model of Liu et al. (2018) where $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. The only difference would be in the definition of c^j for each model, which depends on the post-selection parameters chosen for inference.

Remark 2. Liu et al. (2018) noted that the ideal conditioning event could vary across different models. In some special situations, such as when inferring for the selected regression parameters in a full model $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, conditioning on less information than the polyhedral method is possible. However, in our method, the conditioning event based on the outcome of the randomized selection algorithm is the same for different regression models. Therefore, the construct of our pivot is consistent regardless of our modelling preferences. In our empirical experiments, we demonstrate the performance of our pivot in the selected and full models.

The construct of our pivot broadly applies to other common examples of selective inference. In the Supplementary Material we show how inference after a marginal screening of correlations, or after selective reporting with bootstrapped data, can be easily carried out with our pivot.

3.3. *Pivot motivated by data carving*

We instantiate our pivot using a Gaussian randomization scheme that can be seen related to the data-carving proposal of Fithian et al. (2017). Data carving is similar to data splitting

in that it involves using a subset of the data for selection, but differs from data splitting in that it uses the entire dataset for inference instead of relying solely on the held-out portion.

Suppose that we apply the lasso method to a subsample of size n_1 drawn from a dataset that contains n independent and identically distributed pairs of observations $(y_i, x_i) \in \mathbb{R}^{p+1}$. Then, the lasso on the subsample is asymptotically equivalent to solving a randomized lasso with

$$w \sim N(0_p, \tau^2 \mathbb{E}[x_1 x_1^{\mathrm{T}}]), \tag{5}$$

where $\tau^2 = \sigma^2(n - n_1)/n_1$. This result is formally stated in Panigrahi et al. (2021). We provide some additional details in the Supplementary Material to offer insights into this connection. This motivates us to solve

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \ \tfrac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1 - w^{\mathrm{T}} b \tag{6}$$

with w drawn from a Gaussian distribution with mean 0_p and covariance $\tau^2 X^{\mathrm{T}} X$, which is the sample analog of the covariance matrix in (5). Recall that this was also the randomization scheme in our toy example.

Using this particular form of Gaussian randomization, we can observe that the value of r^j is directly proportional to $e_j \in \mathbb{R}^{|\mathcal{E}|}$. This means that our conditioning event is equivalent to truncating the jth active lasso coefficient O_j to an interval on the real line, which is depicted in Fig. 1. As a result, our pivot in Theorem 1 simplifies as follows.

COROLLARY 1. *Suppose that* Ω *is defined according to* (5)*. Then,*

$$\mathcal{P}^j_{\mathrm{Exact}}(\beta_j^{\mathcal{E}}) = \frac{\int_{-\infty}^{\hat{\beta}_j^{\mathcal{E}}} \phi\{(\sigma\|c^j\|_2)^{-1}(x - \beta_j^{\mathcal{E}})\} \cdot \mathrm{TP}^{[I^j_-, I^j_+]}\{\theta^j(x), \vartheta^j\}\, \mathrm{d}x}{\int_{-\infty}^{\infty} \phi\{(\sigma\|c^j\|_2)^{-1}(x - \beta_j^{\mathcal{E}})\} \cdot \mathrm{TP}^{[I^j_-, I^j_+]}\{\theta^j(x), \vartheta^j\}\, \mathrm{d}x},$$

where

$$(\vartheta^j)^2 = \frac{1}{\tau^2\|c^j\|_2^2}$$

and $\theta^j : \mathbb{R} \to \mathbb{R}$ *is equal to*

$$\theta^j(x) = \frac{1}{\tau^2\|c^j\|_2^2}\{\lambda e_j^{\mathrm{T}}(X_{\mathcal{E}}^{\mathrm{T}} X_{\mathcal{E}})^{-1}\mathcal{S} - x\}.$$

Upon revisiting our toy example, we recall that the polyhedral method truncates the Gaussian distribution of $\hat{\beta}_j^{\mathcal{E}}$ to the interval $[H^j_-, H^j_+]$ for $j \in \{1, 2\}$. In contrast, our new pivot replaces the indicator function $1_{[H^j_-, H^j_+]}(x)$ with the Gaussian probability

$$\mathrm{TP}^{[I^j_-, I^j_+]}\{\theta^j(x), \vartheta^j\}$$

in the integrand of (2).

Remark 3. Of course, solving (6) is not exactly the same as applying the lasso on a subsample of size n_1. If selection is carried out on a randomly selected subsample then our

pivot would provide asymptotic selective inference rather than exact, due to the asymptotic equivalence between the Gaussian randomization and selection on the subsample. Since our current focus is on providing exact guarantees for selective inference, we defer a formal proof of this to future work.

3.4. *Choice of conditioning*

We return to our conditioning event in Proposition 1.

Denote by

$$(L_\alpha^{j,\mathcal{G}}, U_\alpha^{j,\mathcal{G}}) \tag{7}$$

the confidence interval for $\beta_j^{\mathcal{E}}$ if we had based inference on the conditional distribution of $\hat{\beta}^{\mathcal{E}}$, given the event $\{G = \mathcal{G}\}$ as done by the MLE method. In principle, we can fix any arbitrary vector $\eta \in \mathbb{R}^{|\mathcal{E}|}$ and further condition on

$$A^\eta = \left(I - \frac{1}{\eta^{\mathrm{T}}\Theta\eta}\Theta\eta\eta^{\mathrm{T}}\right)O. \tag{8}$$

By writing

$$O = \frac{\Theta\eta}{\eta^{\mathrm{T}}\Theta\eta}\eta^{\mathrm{T}}O + A^\eta,$$

our conditioning event simplifies to an interval as

$$\begin{aligned}\{G = \mathcal{G},\, A^\eta = \mathcal{A}^\eta\} &= \{LO < M,\, U = \mathcal{U},\, A^\eta = \mathcal{A}^\eta\} \\ &= \{I_-^\eta < \eta^{\mathrm{T}}O < I_+^\eta,\, U = \mathcal{U},\, A^\eta = \mathcal{A}^\eta\},\end{aligned}$$

where I_-^η and I_+^η now depend on L, M, Θ and $\mathcal{A}^\eta$.

If we follow the same steps as before then we can obtain an exact pivot for $\beta_j^{\mathcal{E}}$ by using a truncated distribution that is supported on $\mathbb{R} \times [I_-^\eta, I_+^\eta]$. If we let $\eta = r^j$, it leads to the conditioning event in Proposition 1 and to the proposed pivot.

Now we address choosing η, which determines the additional conditioning information. Consider a situation when selection has no impact, i.e., the truncated distribution is no different from the usual distribution with no further adjustment for selection. Our specific choice $\eta = r^j$ is motivated from the fact that no extra price is paid by conditioning on A^{r^j} in the situation described above. In other words, the confidence intervals produced by our pivot

$$\{(L_\alpha^j, U_\alpha^j) : j \in \mathcal{E}\}$$

narrow down to the intervals in (7) as selection has a diminishing impact. We formalize this fact in the following proposition.

PROPOSITION 2. *Let A^η be defined according to* (8)*. Then, we have*

$$\begin{aligned}\mathrm{var}(\hat{\beta}_j^{\mathcal{E}} \mid U = \mathcal{U},\, A^{r^j} = \mathcal{A}^{r^j},\, \hat{\Gamma}^j = g) &= \mathrm{var}(\hat{\beta}_j^{\mathcal{E}} \mid U = \mathcal{U},\, \hat{\Gamma}^j = g) \\ &= \underset{\eta}{\mathrm{maximum}}\, \mathrm{var}(\hat{\beta}_j^{\mathcal{E}} \mid U = \mathcal{U},\, A^\eta = \mathcal{A}^\eta,\, \hat{\Gamma}^j = g).\end{aligned}$$

There might be other ways to choose the direction η. One such option is to choose η in a way that minimizes the variance of the bivariate truncated distribution that arises when we condition on $\{G = \mathcal{G},\ A^\eta = \mathcal{A}^\eta\}$. While this approach seems ideal, it is not straightforward as the resulting optimization is not convex in η and may not be easily solvable.

Another option is to condition on all active lasso coefficients, except for the jth one, when inferring the effect of the jth selected variable. However, this choice will not generalize well to other models post selection. For example, if we add a new variable X^* to the selected model and fit it using the features $E \cup \{X^*\}$, it is unclear what to condition on when inferring for the effect of X^* in this selected model.

In contrast, our approach to choosing η is simple yet principled, which applies broadly to Gaussian linear models with our form of additive randomization introduced at the selection step.

4. Simulations

4.1. *Settings and modelling strategies*

To evaluate how well our pivot performs, we use data generated from a sparse Gaussian model given by

$$y = X_{E^*}\beta_{E^*} + \epsilon. \tag{9}$$

Here, $\epsilon \in \mathbb{R}^n$ is a vector of independent and identically distributed Gaussian errors with mean 0 and variance σ^2 and $E^* \subset [p]$ is a sparse support set for $\beta \in \mathbb{R}^p$.

We construct the feature matrix X by drawing $n = 500$ samples from a $(p = 200)$-dimensional Gaussian distribution $\mathcal{N}(0_p, \Sigma)$ with $\Sigma_{ij} = 0.9^{|i-j|}$. Then, we simulate y from the model in (9) with noise level $\sigma^2 = 3$ and $|E^*| = 5$.

We design two main settings to study how our method compares with previously proposed procedures in selective inference. In our first setting, we vary the proportion of data used for model selection, referred to as *split proportion* in our findings. We compare methods that use roughly the same amount of information for feature selection as data splitting at a prespecified value of split proportion. We elaborate on this further when we describe the different methods under study. In the second setting, we vary the signal strength of the nonzero entries of β to investigate how different methods compare under varying signal regimes. Specifically, we set the magnitude of the nonzero entries for β as $(2f \log p)^{1/2}$. We vary the fraction f in the set $\{0.50, 1, 1.5, 2, 3\}$, and number the corresponding settings as signal regimes 1–5 in our plots.

In each setting, we consider two common modelling strategies.

1. Full Model: we model our response using the full set of features. In other words, we model our response as

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

We estimate the noise level in our data by using residuals based on a regression of y against all p features.

In each round of simulation, we select a sparse set of features $E = \mathcal{E}$. We then consider inference for the selected coefficients in the full model. To be precise, our

parameters, after selection, are

$$\beta^{\mathcal{E}} = (\beta_j : j \in \mathcal{E})^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{E}|}.$$

The vector $\beta^{\mathcal{E}}$ contains entries of β that are present in the selected set $\mathcal{E}$.

2. Selected Model: we model our response using the selected set of features $\mathcal{E}$. That is, we use the model

$$y \sim \mathcal{N}(X_{\mathcal{E}}\beta_{\mathcal{E}}, \sigma^2 I_n).$$

In this case, we estimate the noise level by using the residuals based on a regression of our response against the selected features.

We infer for the partial regression coefficients in the selected model

$$\beta^{\mathcal{E}} = (X_{\mathcal{E}}^{\mathrm{T}} X_{\mathcal{E}})^{-1} X_{\mathcal{E}}^{\mathrm{T}} X_{E^*} \beta_{E^*} \in \mathbb{R}^{|\mathcal{E}|},$$

which are obtained by projecting the true mean $X_{E^*}\beta_{E^*}$ onto the subspace spanned by the selected features.

In both models, we adopt a plug-in approach to estimate the noise variance. We comment on this approach below.

Remark 4. The work by Tian & Taylor (2018) supports the use of a plug-in estimator for σ as long as it is a consistent estimator of the true noise variance before selection. While we use the parametric form of the fitted model to obtain a plug-in estimator, the plug-in approach can be more general in principle. For example, one can estimate the noise variance by using nonparametric function estimation methods, which separates the task of error estimation from the precise parametric modelling of our response.

Our reported findings are based on 500 rounds of simulations for each pair of setting and modelling strategy.

4.2. *Methods*

We compare the following methods.

1. Exact: our proposed method to conduct exact selective inference with Gaussian randomization after solving (1).
2. MLE: the approximate maximum likelihood method reviewed in § 2; this method conducts selective inference with an approximate Gaussian pivot after selecting features through (1).
3. Polyhedral+: this method, introduced by Liu et al. (2018), applies the standard lasso algorithm for selecting features and then conducts inference for the selected coefficients in the Full Model by conditioning on strictly less information than the polyhedral method of Lee et al. (2016).
4. Split: this method is based on data splitting; we divide the training data into two independent parts, using n_1 samples for model selection with the standard lasso algorithm, which is followed by using the remaining samples for valid selective inference.
5. UV: this method uses the UV decomposition proposed by Rasines & Young (2023), where selection and inference are conducted on two independent datasets; for a randomization variable $\tilde{w} \sim N(0_n, \sigma^2 f I_n)$, selection is conducted with the U-estimator

$Y + \tilde{w}$ as the train response and selective inference is conducted for the selected parameters using the V-estimator $Y - \tilde{w}/f$ as the test response.

The two methods Exact and MLE are constructed under the Gaussian randomization scheme that was discussed in § 3.4. Specifically, we fix the randomization covariance as $\Omega = \tau^2 X^{\mathrm{T}} X$ with

$$\tau^2 = \hat{\sigma}^2 \frac{(n - n_1)}{n_1},$$

where $\hat{\sigma}$ is the estimated noise level in our model. Both these methods are compared to data splitting that uses n_1 samples for feature selection. To implement the UV method, we replace σ by its estimated value under our model and to ensure fair comparisons, we set $f = (n - n_1)/n_1$ in our analysis. We report comparisons of our method with the UV method across different signal regimes.

Remark 5. In our simulations, we choose not to use the polyhedral method from Lee et al. (2016). This is because, on average across 500 simulations, the interval lengths it produces are much longer than the other four methods we are using. In fact, the polyhedral method returns infinitely long interval estimates in every setting, which is consistent with the findings of Kivaranovic & Leeb (2021).

Remark 6. When considering the Full Model, we include a summary of the performance of the Polyhedral+ method, as well as the Exact and MLE methods. The benefits of utilizing the entire dataset rather than dividing it into samples are significantly noticeable when applying the Full Model. Therefore, we exclude the split-based methods from our summary plots since they produce considerably longer intervals, on average, compared to the other methods.

For the Selected Model, we compare the four randomized methods used in our simulations. The Polyhedral+ method is designed to provide selective inference only under the Full Model and does not apply to the Selected Model.

4.3. *Findings*

First, we evaluate the accuracy of feature selection by using

$$\text{F1 score} = \frac{\text{true positives}}{\text{true positives} + (\text{false positives} + \text{false negatives})/2}$$

in our two main settings.

In Fig. 2(a), we vary the split proportion $\rho = n_1/n$ at a fixed strength of signals while keeping the signal strength fixed. We use the randomized lasso method to conduct feature selection with Gaussian randomization that corresponds to the prespecified split proportion ρ; Exact and MLE provide inference for the effects of the features selected with this randomized version of the lasso. The standard implementation of the lasso, which is used by Polyhedral+, applies feature selection on the entire dataset, and is represented in the plot as Standard. The distribution of the F1 score for the Gaussian randomization scheme closely resembles the randomization involved in the related Split procedure. As expected, the accuracy of selection increases with higher values of the split proportion, eventually matching the accuracy attained by the standard method on the full data.

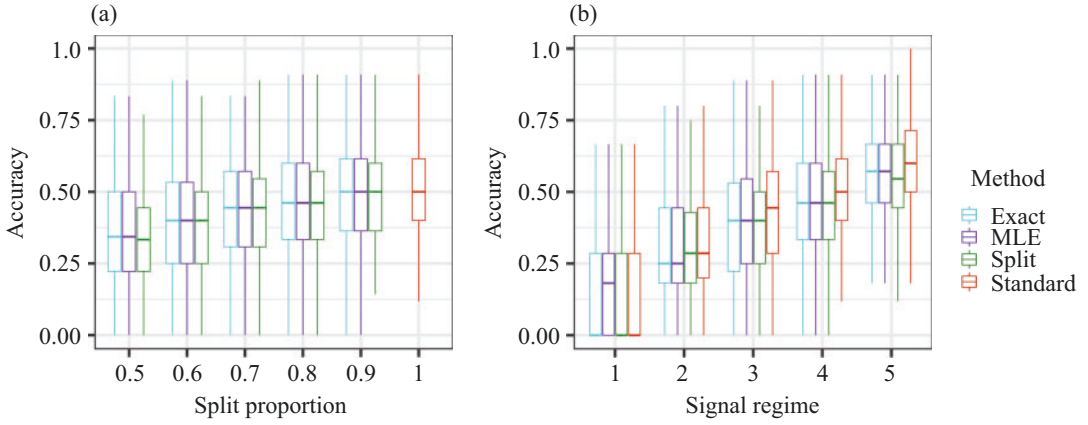

Fig. 2. Accuracy based on the quality of feature selection.

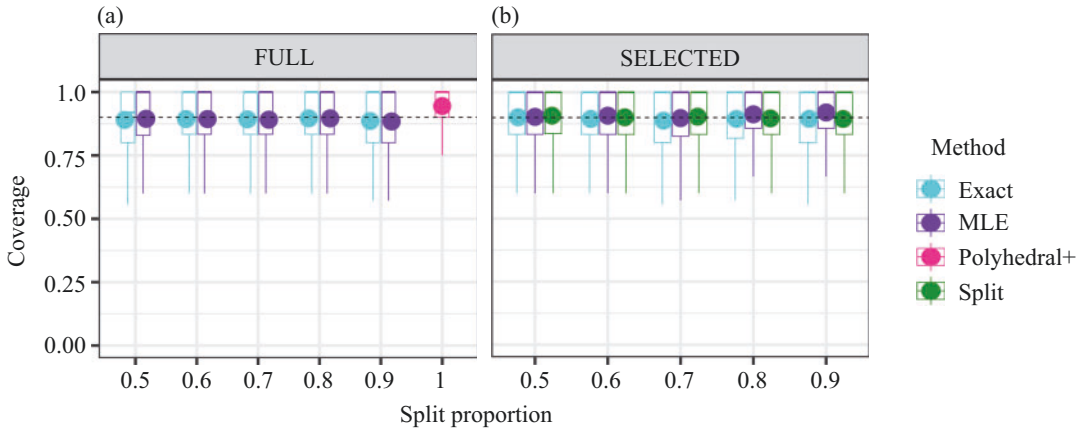

Fig. 3. Coverage rate of confidence intervals under signal regime 3.

In Fig. 2(b), we fix the split proportion at 0.80 and vary our signal regimes in the set 1–5. Consistent with expectations, the accuracy of feature selection increases as we strengthen the signals. Notably, all the methods used for feature selection perform almost equally well at a split proportion of 0.80, which is consistent with the findings of Fig. 2(a).

Next, we compute the false coverage rate of the confidence intervals for different methods, which is equal to $\mathrm{FCR} = \{j \in \mathcal{E} : \beta_j^{\mathcal{E}} \notin \mathcal{C}_j^{\mathcal{E}}\}/\max(|\mathcal{E}|, 1)$. In Figs. 3 and 4, we plot the coverage rates $1 - \mathrm{FCR}$ for 90% confidence intervals under the Full Model and Selected Model. The averaged coverage rate, over all replications, is highlighted with a filled circle. The horizontal dashed line at 0.90 depicts the target coverage rate for all the methods.

Exact achieves the desired rate of coverage, as do the previous methods of selective inference. This pattern remains consistent even as we change the split proportion or the strength of signals in different signal regimes.

In Figs. 5 and 6, we investigate how the Exact confidence intervals compare in length when we vary the split proportion and the strength of signals.

Under the Full Model, we observe that the interval lengths produced by Exact and MLE are consistently less variable than those of Polyhedral+. This observation is also true if

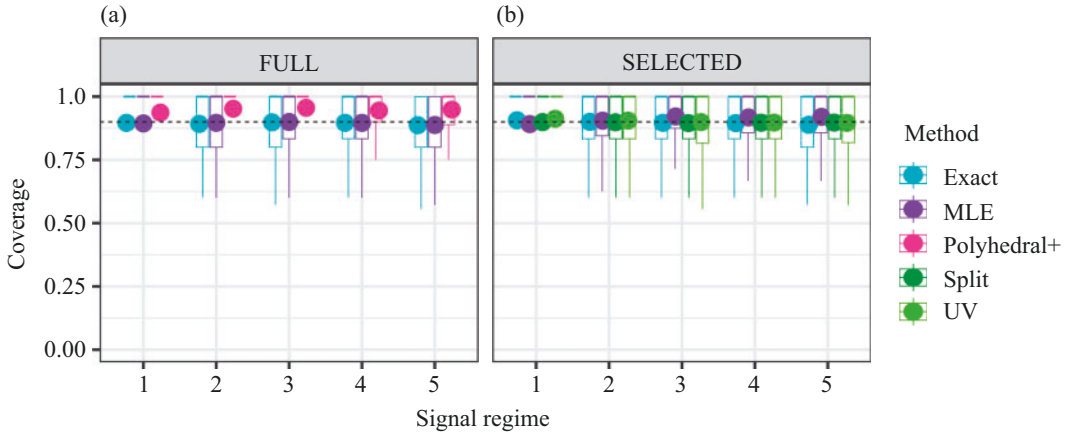

Fig. 4. Coverage rate of confidence intervals at fixed split proportion 0.80.

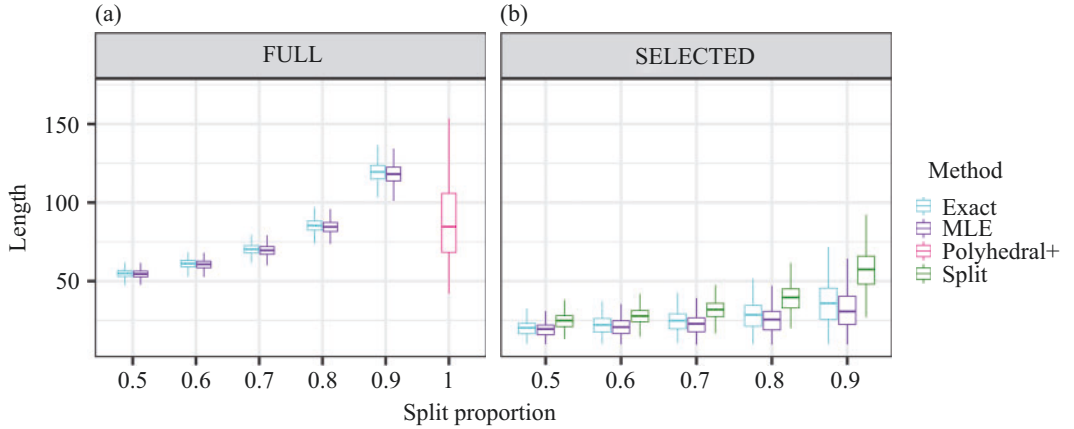

Fig. 5. Length of confidence intervals under signal regime 3.

we focus attention on split proportion $\rho = 0.80$, at which the randomized methods are comparable to Polyhedral+ in terms of the quality of feature selection.

Similar patterns are seen in Fig. 6 as we change the signal strengths under signal regimes 1–5. Under both models, our Exact method yields only nominally longer intervals than MLE, but, consistently gives shorter intervals than the two split-based strategies Split and UV. As previously mentioned, we only display the lengths of split-based methods for the Selected Model, as they are much longer than the other methods when used under the Full Model. The increasing cost of discarding data from the selection stage is evident from Fig. 5(b).

5. Analysis of HIV drug resistance data

We apply our method to the HIV drug resistance data. This dataset, originally analysed by Rhee et al. (2006), is publicly available on the Stanford HIV Database. The goal of the analysis is to find associations between mutations of the HIV virus and drug resistance to antiretroviral drugs. We extract a part of this dataset that focuses on the response to one

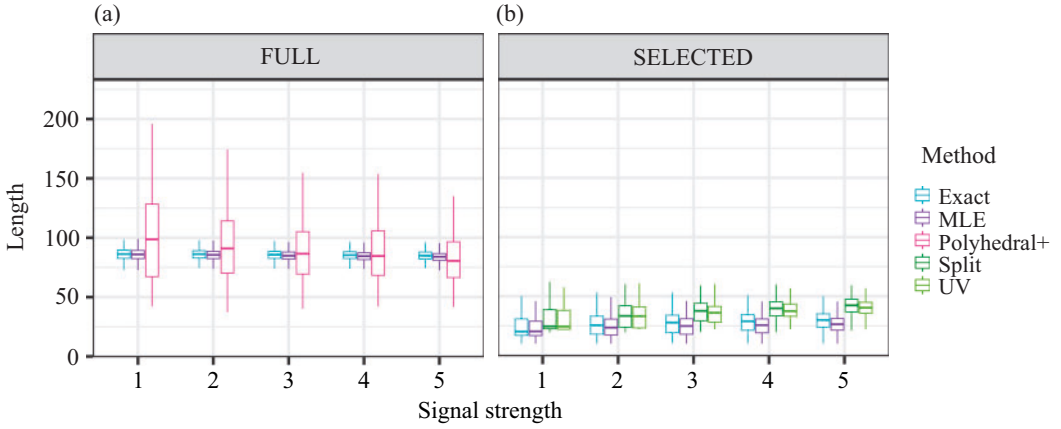

Fig. 6. Length of confidence intervals at fixed split proportion 0.80.

particular drug, lamivudine (3TC), as has been described previously by Bi et al. (2020) and Panigrahi et al. (2021). The predictive features in these data are 91 mutations that appeared more than 10 times in the samples, and the response is a log-transformed value of the measurement for drug resistance. Our dataset contains 633 sample observations for the response and the set of 91 features.

We focus on three randomized procedures for interval estimation. To run our method, we consider drawing a Gaussian randomization variable $w \sim \mathcal{N}(0_p, \Omega)$, where $\rho = n_1/n = 0.8$, and Ω is set as per (5). We implement the randomized lasso with the randomization variable w. The randomized lasso selects a subset of 14 mutations. At the inference stage, we use our exact pivot to construct confidence intervals for the selected regression coefficients. For comparison, we construct approximate confidence intervals using MLE after the same run of the randomized lasso. We also consider the intervals produced by Split based on $\rho = 0.8$. That is, Split uses 80% of the data samples for selecting features, and this resulted in selecting a subset of 17 features. The remaining 20% of the samples were reserved for selective inference.

On average, we observe that the length of interval estimators based on the Exact method is 3.76. This follows our simulated results, which showed that the Exact intervals are longer than the MLE intervals, with an average length of 2.76. However, this longer length is a necessary trade-off to achieve exact selective inference with our pivot. Despite this, our intervals are still shorter than those produced by the Split procedure, which have an average length of 6.58 in this instance. Figure 7 displays box plots for the interval lengths, which clearly demonstrate this pattern.

6. Discussion

We conclude the paper with two remarks. Firstly, while exact selective inference has its benefits, it also comes at a cost. By conditioning on additional information to obtain our pivot, we sacrifice some power in comparison to approximate techniques developed in prior work, such as Panigrahi et al. (2017) and Panigrahi & Taylor (2023). Further research is required to investigate the cost in power for exact inference. Secondly, the pivot generated from the randomization scheme used in the paper can also be applied to more general estimation problems, including the class of M-estimation problems. We believe that the same

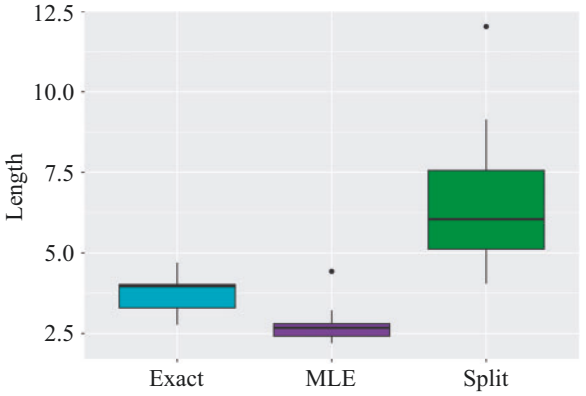

Fig. 7. Lengths of interval estimators.

pivot could be used as long as the selection algorithm permits a linear representation in optimization variables at the solution. For these problems, our pivot would provide asymptotic selective inferences instead of exact selective inferences, which would require a formal theoretical justification and needs to be studied in future work.

ACKNOWLEDGEMENT

Panigrahi's research was supported in part by the National Science Foundation. Fry's research was supported by a National Science Foundation Graduate Research Fellowship. Taylor's research was supported by ARO.

SUPPLEMENTARY MATERIAL

The Supplementary Material contains further examples and proofs of the technical results.

REFERENCES

BACHOC, F., PREINERSTORFER, D. & STEINBERGER, L. (2020). Uniformly valid confidence intervals post-model-selection. *Ann. Statist.* **48**, 440–63.

BERK, R., BROWN, L., BUJA, A., ZHANG, K. & ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41**, 802–37.

BI, N., MARKOVIC, J., XIA, L. & TAYLOR, J. (2020). Inferactive data analysis. *Scand. J. Statist.* **47**, 212–49.

CARRINGTON, R. & FEARNHEAD, P. (2024). Improving power by conditioning on less in post-selection inference for changepoints. *arXiv:* 2301.05636v3.

CHARKHI, A. & CLAESKENS, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* **105**, 645–64.

CHEN, S. & BIEN, J. (2020). Valid inference corrected for outlier removal. *J. Comp. Graph. Statist.* **29**, 323–34.

DUY, V. N. L., TODA, H., SUGIYAMA, R. & TAKEUCHI, I. (2020). Computing valid *p*-value for optimal changepoint by selective inference using dynamic programming. In *Proc. 34th Int. Conf. Neural Info. Proces. Syst.*, pp. 11356–67. Red Hook, NY: Curran Associates.

FITHIAN, W., SUN, D. & TAYLOR, J. (2017). Optimal inference after model selection. *arXiv:* 1410.2597v4.

GAO, L. L., BIEN, J. & WITTEN, D. (2024). Selective inference for hierarchical clustering. *J. Am. Statist. Assoc.* **119**, 332–42.

HUANG, Y., PIRENNE, S., PANIGRAHI, S. & CLAESKENS, G. (2023). Selective inference using randomized group lasso estimators for general models. *arXiv:* 2306.13829v3.

Hyun, S., G'Sell, M. & Tibshirani, R. J. (2018). Exact post-selection inference for the generalized lasso path. *Electron. J. Statist.* **12**, 1053–97.
Kivaranovic, D. & Leeb, H. (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *J. Am. Statist. Assoc.* **116**, 845–57.
Kivaranovic, D. & Leeb, H. (2024). A (tight) upper bound for the length of confidence intervals with conditional coverage. *arXiv:* 2007.12448v3.
Le Duy, V. N. & Takeuchi, I. (2022). More powerful conditional selective inference for generalized lasso by parametric programming. *J. Mach. Learn. Res.* **23**, 1–37.
Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference with the lasso. *Ann. Statist.* **44**, 907–27.
Lee, J. D. & Taylor, J. E. (2014). Exact post model selection inference for marginal screening. In *Proc. 27th Int. Conf. Neural Info. Proces. Syst.*, pp. 136–44. Cambridge, MA: MIT Press.
Leiner, J., Duan, B., Wasserman, L. & Ramdas, A. (2023). Data blurring: sample splitting a single sample. *arXiv:* 2112.11079v9.
Liu, K., Markovic, J. & Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv:* 1801.09037v2.
Neufeld, A., Gao, L. L., Popp, J., Battle, A. & Witten, D. (2023). Inference after latent variable estimation for single-cell RNA sequencing data. *arXiv:* 2207.00554v5.
Panigrahi, S. (2023). Carving model-free inference. *Ann. Statist.* **51**, 2318–41.
Panigrahi, S., MacDonald, P. W. & Kessler, D. (2023a). Approximate post-selective inference for regression with the group lasso. *J. Mach. Learn. Res.* **24**, 1–49.
Panigrahi, S., Markovic, J. & Taylor, J. (2017). An MCMC-free approach to post-selective inference. *arXiv:* 1703.06154v2.
Panigrahi, S., Mohammed, S., Rao, A. & Baladandayuthapani, V. (2023b). Integrative Bayesian models using post-selective inference: a case study in radiogenomics. *Biometrics* **79**, 1801–13.
Panigrahi, S. & Taylor, J. (2023). Approximate selective inference via maximum likelihood. *J. Am. Statist. Assoc.* **118**, 2810–20.
Panigrahi, S., Taylor, J. & Weinstein, A. (2021). Integrative methods for post-selection inference under convex constraints. *Ann. Statist.* **49**, 2803–24.
Rasines, D. G. & Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika* **110**, 597–614.
Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L. & Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Nat. Acad. Sci.* **103**, 17355–60.
Rinaldo, A., Wasserman, L. & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47**, 3438–69.
Schultheiss, C., Renaux, C. & Bühlmann, P. (2021). Multicarving for high-dimensional post-selection inference. *Electron. J. Statist.* **15**, 1695–742.
Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K. & Takeuchi, I. (2017). Selective inference for sparse high-order interaction models. In *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, pp. 3338–47. PMLR.
Tanizaki, K., Hashimoto, N., Inatsu, Y., Hontani, H. & Takeuchi, I. (2020). Computing valid *p*-values for image segmentation by selective inference. In *2020 IEEE/CVF Conf. Comp. Vis. Pat. Recog.*, pp. 9550–9. Los Alamitos, CA: IEEE Computer Society.
Tian, X., Panigrahi, S., Markovic, J., Bi, N. & Taylor, J. (2016). Selective sampling after solving a convex problem. *arXiv:* 1609.05609v1.
Tian, X. & Taylor, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46**, 679–710.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
Yang, F., Barber, R. F., Jain, P. & Lafferty, J. (2016). Selective inference for group-sparse linear models. In *Proc. 30th Int. Conf. Neural Info. Proces. Syst.*, pp. 2477–85. Red Hook, NY: Curran Associates.
Zhao, Q. & Panigrahi, S. (2019). Selective inference for effect modification: an empirical investigation. *Observat. Studies* **5**, 131–40.

[*Received on* 13 *March* 2023. *Editorial decision on* 7 *February* 2024]