Journal of Machine Learning Research 26 (2025) 1-44 Submitted 3/23; Revised 12/24; Published 1/25

Selective Inference with Distributed Data

Sifan Liu SFLIU@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305-4020, USA

Snigdha Panigrahi PSNIGDHA@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1107, USA

Editor: Po-Ling Loh

Abstract

When data are distributed across multiple sites or machines rather than centralized in one location, researchers face the challenge of extracting meaningful information without directly sharing individual data points. While there are many distributed methods for point estimation using sparse regression, few options are available for estimating uncertainties or conducting hypothesis tests based on the estimated sparsity. In this paper, we introduce a procedure for performing selective inference with distributed data. We consider a scenario where each local machine solves a lasso problem and communicates the selected predictors to a central machine. The central machine then aggregates these selected predictors to form a generalized linear model (GLM).

Our goal is to provide valid inference for the selected GLM while reusing data that have been used in the model selection process. Our proposed procedure only requires low-dimensional summary statistics from local machines, thus keeping communication costs low and preserving the privacy of individual data sets. Furthermore, this procedure can be applied in scenarios where model selection is repeatedly conducted on randomly sub-sampled data sets, addressing the p-value lottery problem linked with model selection. We demonstrate the effectiveness of our approach through simulations and an analysis of a medical data set on ICU admissions.

Keywords: carving, data aggregation, generalized linear models, lasso, post-selection inference, selective inference, selective likelihood

1. Introduction

In recent years, analyzing decentralized data spread across various sites or machines has become increasingly important, particularly for collaborative research. This has created a growing demand for distributed methods that enable researchers to extract meaningful information from such data while respecting its decentralized nature, and to combine this information using summary statistics for preserving privacy within the individual data sets. One of the simplest and most popular approaches in the distributed framework is "divide-and-conquer", which is also known as the "split-and-merge" or the "one-shot" approach; see for example the early work by Mcdonald et al. (2009); Zinkevich et al. (2010); Zhang et al.

(2012). Most of these approaches use only one round of communication. Each local machine estimates the unknown parameter using its subset of the training data, and communicates the estimator to a central machine which merges the local estimators to obtain a global estimator.

In our paper, we focus on a sparse regression setting where only a few of the measured predictors affect the response. The goal in this setup is usually two-fold: (i) select relevant predictors, and model the response by using the estimated sparsity, (ii) provide uncertainties or conduct hypothesis tests for regression parameters in the estimated model, and all this while, respect the distributed nature of data. Substantial progress has been made on the first goal. For example, Lee et al. (2015) average locally computed, debiased lasso estimators, and show that the averaged estimator achieves the same estimation rate as the full-sample lasso, as long as the number of machines is not too large. Chen and Xie (2014) prove that the models aggregated via majority voting, based on variables selected by local machines, are consistent under some conditions. More recently, Battey et al. (2018) provide a debiased approach for hypothesis testing in the distributed setting. However, these methods are limited to models that are fixed before the selection of relevant predictors. As a result, the communication cost in prior work scales with the number of original predictors, which can be unnecessarily large in sparse settings.

We introduce a new procedure for conducting selective inference with distributed data. Selective inference tools ensure valid inference by accounting for the fact that the same data, used to select models, is reused when providing confidence intervals and p-values. Several ingenious methods have been developed to provide selective inference in sparse regression problems; see papers by Benjamini and Yekutieli (2005); Berk et al. (2013); Belloni et al. (2015); Lee et al. (2016); Tian and Taylor (2018); Charkhi and Claeskens (2018); Bachoc et al. (2019); Panigrahi et al. (2021). Our procedure in this paper reuses data from all machines to form an asymptotic "selective likelihood". This likelihood delivers approximately-valid selective inference in a selected generalized linear model (GLM), which is based on the predictors selected across different machines.

The proposed procedure has several important features:

1. It only requires low-dimensional summary statistics from each machine to conduct selective inference. Thus, the techniques developed are applicable to settings when data sets are distributed across different sites due to security, privacy, or ethical concerns, as seen in the areas of differential privacy (Balcan et al., 2012) and federated learning (McMahan et al., 2017). These sites can merge summary information without sharing individual data, enabling inference in a GLM based on the estimated sparsity.

2. The communication cost of our inferential procedure is only linear in the dimension of the selected model, which can be substantially smaller than the initial dimension of the problem.

3. This procedure can be easily adapted to address the "p-value lottery" problem in model selection, which arises when the selection process is repeated on randomly subsampled data sets. We show that our procedure serves as an efficient alternative to multi-splitting (Meinshausen et al., 2009) and multi-carving (Schultheiss et al., 2021) without recourse to Markov chain Monte Carlo (MCMC) sampling.

The remaining paper is structured as follows. We provide a slightly more technical account of our contributions after outlining the problem setup, and review related work in Section 2. In Section 3, we introduce our procedure for selective inference with distributed data. In Section 4, we provide an asymptotic justification for our selective likelihood, which forms the basis of our approach. We explore how our procedure can be modified to address the p-value lottery problem in Section 5. Section 6 reports findings from the application of our method on simulated data sets and a publicly available, medical data set on intensive care unit (ICU) admissions. We conclude the paper with a discussion in Section 7. Proofs of our technical results are collected in the Appendix.

2. Problem setup and background

In this section, we describe the distributed setup and introduce some background on selective inference with a single machine. Other related work is summarized at the end.

2.1 Problem setup

We consider a distributed setup with K local machines, all connected to a central machine, referred to as machine 0. Suppose that we observe n i.i.d. observations

$$(y_i, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1},\ i \in [n],$$

where $[n] = \{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$. Let $\mathcal{C}^{(k)} \subset [n]$ denote the index set of the samples stored at machine k, for $k \in \{0\} \cup [K]$. The index sets $\mathcal{C}^{(k)}$ are disjoint and form a partition of $[n]$. Let $n_k = |\mathcal{C}^{(k)}|$ be the cardinality of $\mathcal{C}^{(k)}$. Let $\rho_k = \dfrac{n_k}{n}$ be the proportion of samples store at machine k. Then we have

$$n = \sum_{k=0}^{K} n_k, \quad 1 = \sum_{k=0}^{K} \rho_k.$$

Let $D^{(k)} = \left(Y^{(k)}, X^{(k)}\right)$ denote the data stored at machine k, where $X^{(k)} \in \mathbb{R}^{n_k \times p}$ represents the feature matrix and $Y^{(k)} \in \mathbb{R}^{n_k}$ represents the response vector.

Each local machine conducts variable selection with a loss function based on a GLM-density. Let this loss function at machine k be denoted by

$$\ell^{(k)}(\beta; D^{(k)}) = \frac{1}{\sqrt{n}\rho_k} \sum_{i \in \mathcal{C}^{(k)}} \left\{ A(x_i^\intercal \beta) - y_i x_i^\intercal \beta \right\},$$

where $A(\cdot)$ is the log-partition function of the GLM density. That is, machine k solves the following lasso problem:

$$\widehat{\beta}^{\Lambda,(k)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\ \ell^{(k)}(\beta; D^{(k)}) + \|\Lambda^{(k)}\beta\|_1, \tag{1}$$

where $\Lambda^{(k)} = \operatorname{diag}\left(\lambda_1^{(k)}, \ldots, \lambda_p^{(k)}\right)$ is the diagonal matrix of regularization parameters. Although our procedure works when the regularization parameters differ across machines, for simplicity, we assume $\Lambda^{(k)} = \Lambda$ for $k \in [K]$.

Let

$$\widehat{E}^{(k)} = \left\{ j \in [p] : \widehat{\beta}_j^{\Lambda,(k)} \neq 0 \right\}.$$

represent the set of predictors with non-zero lasso coefficients selected at machine k for $1 \leq k \leq K$. We will denote the realized value of $\widehat{E}^{(k)}$ with data $D^{(k)}$ as $E^{(k)}$. Let $d^{(k)} = |E^{(k)}|$ denote the number of selected variables at machine k. Note, for a vector $x \in \mathbb{R}^p$ and $E \subset [p]$, x_E represents the subvector of x with entries in the set E, and for a matrix X, X_E is composed of the columns of X present in the set E.

2.2 Selected model

As described above, the K local machines return as output the selected predictors, which are then communicated to the central machine. The central machine aggregates these selected predictors as

$$\widehat{E} = \texttt{Aggregate}\left(\left\{ \widehat{E}^{(k)}, k \in [K] \right\} \right). \tag{2}$$

We let E be the observed value of $\widehat{E}$ and let $d = |E|$. Deferring the discussion on general aggregation rules to Appendix D, now consider the union aggregation rule

$$\widehat{E} = \bigcup_{k \in [K]} \widehat{E}^{(k)}$$

as a concrete example.

Given E, our response y is modeled using a selected GLM based on X_E. The density of y is given by:

$$f(y \mid x_E, \beta_E) = \exp\left(\frac{y x_E^\intercal \beta_E - A(x_E^\intercal \beta_E)}{\sigma^2} \right) \cdot c(y; \sigma), \tag{3}$$

where β_E, the regression coefficient vector, is the parameter that we want to infer about. We assume that the dispersion parameter, σ, is either known or can be consistently estimated.

Some immediate questions arise when we seek inference for β_E:

- Can the central machine reuse data from the local machines to conduct selective inference for β_E? Of course, naive inference, which uses all the data without adjusting for the bias from the model selection process, falls short of coverage guarantees, as illustrated on one of our simulated instances in Figure 1. Section 6 provides the details of this simulation.

- A procedure for making selective inference must respect the distributed nature of data, as done at the time of selection. What information does the central machine require from the local machines, and how many communication exchanges does it involve?

2.3 Some background when $K = 1$

We provide some background in a rather simple setup with $K = 1$. Consider the special case of a Gaussian linear model, i.e., in (3), we have

$$A(x_E^\intercal \beta_E) = \frac{1}{2}(x_E^\intercal \beta_E)^2, \; c(y; \sigma) = -\frac{y^2}{2\sigma^2}.$$

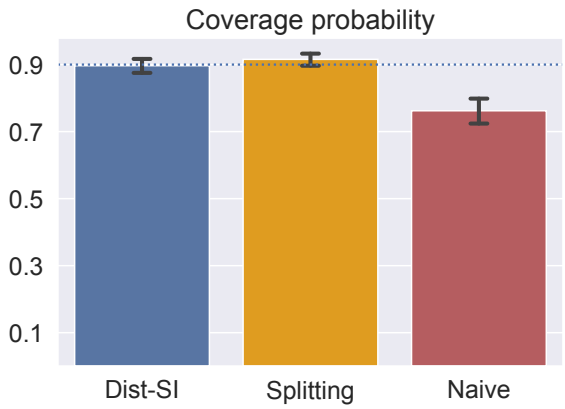


Figure 1: Coverage probabilities of "Dist-SI" (our procedure for selective inference with distributed data), the "Splitting", and "Naïve" methods for a Gaussian linear model, with 2 local machines and a central machine. The 2 local machines and the central machine each have 1000 samples. The pre-specified level of coverage is 0.90, as indicated by the dotted horizontal line at 0.9.

Post selection, we want to infer for β_E.

We begin by noting that the regression problem in (1), at machine 1, can be rewritten as

$$\widehat{\beta}^{\lambda,(1)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{\sqrt{n}} \sum_{i \in [n]} \frac{1}{2}(y_i - x_i^\intercal \beta)^2 + \|\Lambda\beta\|_1 - \sqrt{n}\omega_n^\intercal \beta, \tag{4}$$

where

$$\omega_n = \frac{1}{n} \sum_{i \in [n]} x_i(x_i^\intercal \widehat{\beta}^{\lambda,(1)} - y_i) - \frac{1}{n_k} \sum_{i \in \mathcal{C}^{(1)}} x_i(x_i^\intercal \widehat{\beta}^{\lambda,(1)} - y_i).$$

The term ω_n is what we call a randomization variable. Regression of the form (4), with an added randomization variable ω_n, was termed the randomized lasso in Tian et al. (2016).

Suppose we have n observations in our response vector that are drawn as independent Gaussian variables with the same variance. Further, suppose that we solved the randomized lasso with a Gaussian randomization variable $\sqrt{n}\omega_n \sim \mathcal{N}(0_p, \Sigma_\Omega)$. A recent procedure by Panigrahi and Taylor (2023) constructs a "selective likelihood" by conditioning on a subset of the selection event

$$\left\{\widehat{E} = E\right\}.$$

This approach enables valid selective inference in the selected linear model while allowing for the re-use of data employed during model selection. This procedure centers interval estimates around the maximum likelihood estimator (MLE) of the selective likelihood and estimates its variance using the observed Fisher information matrix. If $\widehat{\beta}_E^{(S)}$ and $\widehat{I}_{E,E}^{(S)}$ denote the selective MLE and the observed Fisher information matrix, using the selective likelihood, a two-sided $100 \cdot (1-\alpha)\%$ confidence interval for $\beta_{E,j}$ is constructed as

$$\widehat{\beta}_{E,j}^{(S)} \pm z_{1-\alpha/2} \cdot \frac{\widehat{\sigma}_j^{(S)}}{\sqrt{n}},$$

where $\widehat{\sigma}_j^{(S)} = \sqrt{\left(\widehat{I}_{E,E}^{(S)}\right)^{-1}_{j,j}}$ is the estimated standard deviation of $\sqrt{n}\widehat{\beta}_{E,j}^{(S)}$, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of a standard normal distribution. This procedure closely resembles classical inference via maximum likelihood, with the exception that the standard estimators are replaced by their selective analogs that adjust for bias from the model selection process.

2.4 Contributions and other related work

The contribution of the proposed method is threefold. First, we identify a simple representation of the selection event by developing a randomization framework. A significant challenge in conditional selective inference is the need for an explicit characterization of the selection event, which is particularly difficult in distributed settings. By leveraging the randomization framework, the selection event can be expressed in a simple closed form, making conditional selective inference feasible for GLMs with selected predictors. Second, the proposed method is efficient in both computation and communication. It requires only low-dimensional summary statistics to be shared from local machines with a central machine. After aggregating these summary statistics, the central machine solves a convex optimization problem. Therefore, the proposed approach is suitable for scenarios where direct data sharing between sites is not possible or inter-site communication is costly. Third, the proposed method can be easily adapted to address the p-value lottery problem. In this sense, our procedure is related to the multi-carving approach in Schultheiss et al. (2021), which uses conditional techniques known as carving (Fithian et al., 2014; Panigrahi, 2023). However, existing carving and multi-carving implementations often depend on computationally expensive MCMC sampling, while our method provides a more efficient alternative.

We conclude this section with some more related work. Within the distributed setting, much work has been devoted to the averaged M-estimator (Mcdonald et al., 2009; Zinkevich et al., 2010; Zhang et al., 2012; Rosenblatt and Nadler, 2016). Rosenblatt and Nadler (2016) show that the averaged M-estimator is first-order equivalent to the centralized M-estimator in the fixed-dimension setting. Dobriban and Sheng (2021) study the efficiency of an estimator based on weighted average in the linear regression setting, as dimensions grow with sample sizes. Some methods have taken a likelihood-centric approach, e.g., Jordan et al. (2019) propose a surrogate likelihood where higher-order derivatives in a Taylor-series expansion of the full log-likelihood are replaced by local approximations. Lin and Xi (2011) propose an aggregated estimator for generalized linear models (GLM), where the locally computed MLE and Hessian of the likelihood are merged for efficiency gains. In work by Huang and Gelman (2005); Neiswanger et al. (2014); Wang and Dunson (2013); Scott et al. (2016); Minsker et al. (2017); Srivastava et al. (2018), distributed MCMC algorithms combine local posterior samples to obtain a global posterior distribution.

In the post-selection inference literature, a selective likelihood was appended to priors for Bayesian inference post selection in Panigrahi and Taylor (2018); Panigrahi et al. (2021). The focus in these settings was on a category of variable selection rules that can be written as a set of polyhedral constraints on data. For the same category of selection rules, a different body of work (Lee et al., 2016; Hyun et al., 2018; Le Duy and Takeuchi, 2022) considers a truncated Gaussian distribution for inference from normal data. Moving beyond normal data, Taylor and Tibshirani (2018) outline an asymptotic scheme to provide selective inference in GLMs. For blackbox selection, Liu et al. (2022) obtain the selective likelihood

by learning the selection event from bootstrapped data. We take a different approach in the distributed setup by casting the problem into a randomized framework, and provide an asymptotic likelihood function for the selected GLM. The use of a randomized framework for selective inference was explored in Tian et al. (2016); Panigrahi et al. (2017) to improve power. In Rasines and Young (2023); Panigrahi et al. (2022), a randomized framework for selective inference was investigated for more efficient uses of data than sample splitting. The connection between a randomized framework for selective inference and algorithmic stability was explored in Zrnic and Jordan (2023).

3. Selective inference with distributed data

In this section, we first provide a characterization of the selection event with distributed data. Then we introduce a selective likelihood that takes into account the aggregation of predictors across machines to form the selected GLM.

3.1 The selection event

Fixing some notations, for $z \in \mathbb{R}^n$, let $\nabla A(z)$ and $\nabla^2 A(z)$ denote the vectors in $\mathbb{R}^n$ whose i^{th} coordinates are the first and second derivatives of $A(\cdot)$ at z_i, respectively. We recast the regression problem (1) as a randomized lasso problem similar to (4). To do so, define the randomization variables

$$\omega^{(k)} = \frac{1}{n} X^\intercal (\nabla A(X\widehat{\beta}^{\Lambda,(k)}) - Y) - \frac{1}{n_k} X^{(k)\intercal} (\nabla A(X^{(k)}\widehat{\beta}^{\Lambda,(k)}) - Y^{(k)}) \tag{5}$$

for $k \in [K]$, where $\widehat{\beta}^{\Lambda,(k)}$ is the solution to the lasso problem at machine k. Let $\mathbf{\Omega} = (\omega^{(1),\intercal}, \dots, \omega^{(K),\intercal})^\intercal \in \mathbb{R}^{Kp}$ be the stack of the K randomization variables.

As reviewed earlier, an equivalent expression for the lasso problem on machine k is

$$\widehat{\beta}^{\Lambda,(k)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{\sqrt{n}} \sum_{i \in [n]} (A(x_i^\intercal \beta) - y_i x_i^\intercal \beta) + \|\Lambda\beta\|_1 - \sqrt{n}\omega^{(k)\intercal}\beta. \tag{6}$$

This is because the Karush-Kuhn-Tucker (KKT) conditions of problem (1) and problem (6) are the same. Note that the KKT conditions are given as follows:

$$\begin{aligned} &\frac{1}{\sqrt{n}} X^\intercal \left(\nabla A(X\widehat{\beta}^{\Lambda,(k)}) - Y\right) + \gamma^{(k)} = \sqrt{n}\omega^{(k)}, \\ &\gamma^{(k)}_{E^{(k)}} = \Lambda_{E^{(k)}} s^{(k)}, \quad s^{(k)} = \operatorname{sign}\left(\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}}\right), \\ &\gamma^{(k)}_{-E^{(k)}} = \Lambda_{-E^{(k)}} z^{(k)}, \quad \|z^{(k)}\|_\infty \le 1, \end{aligned} \tag{7}$$

where the vector $\gamma^{(k)}$ represents the subgradient of the penalty $\|\Lambda\beta\|_1$ evaluated at the lasso solution.

To perform inference for the selected GLM (3), we need some additional notations. Let $d_k = |E^{(k)}|$, $d = |E|$, and $\bar{d} = \sum_{k \in [K]} d_k$. Let $b^{(k)} = \widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} \in \mathbb{R}^{d_k}$ represent the lasso solution from machine k with only the active variables $E^{(k)}$, and let $\widehat{B}^{(k)}$ denote the corresponding random vector. Let $\widehat{\mathbf{B}} \in \mathbb{R}^{\bar{d}}$ be the random vector formed by stacking $\widehat{B}^{(k)}$

Variable	Local machine k R.V.	Obs.	Dim	Aggregated version R.V.	Obs.	Dim
Selected variable set	$\widehat{E}^{(k)}$	$E^{(k)}$	$\subseteq [p]$	$\widehat{E}$	E	$\subseteq [p]$
Active variables of lasso solution	$\widehat{B}^{(k)}$	$b^{(k)}$	d_k	$\widehat{\mathbf{B}}$	$\mathbf{b}$	$\bar{d}$
Signs of active variables	$\widehat{S}^{(k)}$	$s^{(k)}$	d_k	$\widehat{\mathbf{S}}$	$\mathbf{s}$	$\bar{d}$
Subgradients of inactive variables	$\widehat{Z}^{(k)}$	$z^{(k)}$	$p - d_k$	$\widehat{\mathbf{Z}}$	$\mathbf{z}$	$pK - \bar{d}$
Subgradients of lasso penalty $\|\Lambda\beta\|_1$	$\widehat{\Gamma}^{(k)}$	$\gamma^{(k)}$	p	$\widehat{\mathbf{\Gamma}}$	$\boldsymbol{\gamma}$	pK
Randomization variables	$\omega^{(k)}$		pK	$\mathbf{\Omega}$		pK

Table 1: Notations for the variables involved in selection. Columns 2-4 list the variables for the lasso problem on local machine k, while columns 5-7 show the aggregated or stacked versions. For the optimization variables (excluding selected variable sets and randomization variables), the corresponding random variables (R.V.) are denoted by uppercase letters with a hat, observed values (Obs.) by lowercase letters, and their dimensions (Dim) are also provided. The relationship among the variables is given in Equation (7).

from all machines, and let $\mathbf{b}$ denote the realized value of $\widehat{\mathbf{B}}$. Following the same scheme, we denote random variables with uppercase letters, while lowercase letters will represent their realized values, and bold letters will represent a vector obtained by stacking these variables from all machines. To make it easier for readers, we have provided Table 1 that contains a list of notations used in our paper.

Our method conditions on

$$\left\{\widehat{\Gamma}^{(k)} = \gamma^{(k)} \; \forall \, k \in [K]\right\}, \tag{8}$$

a proper subset of $\{\widehat{E} = E\}$, where $\widehat{\Gamma}^{(k)}$ is the subgradient of the lasso penalty with realized value $\gamma^{(k)}$. In fact, event (8) is a subset of

$$\{\widehat{E}^{(k)} = E^{(k)} \; \forall \, k \in [K]\},$$

and therefore a subset of the event $\{\widehat{E} = E\}$. Note that the conditional approach for selective inference remains valid even when we work with a proper subset of the selection event, and it is easier to characterize the event (8) than the event $\{\widehat{E} = E\}$, as shown in the next result. In this result, note that $\widehat{\mathbf{S}} \in \mathbb{R}^{\bar{d}}$ and $\widehat{\mathbf{Z}} \in \mathbb{R}^{pK-\bar{d}}$ denote the signs of lasso solutions at active variables and the subgradients at inactive variables, respectively, stacked across K machines, with the realized values $\mathbf{s}$ and $\mathbf{z}$ (see Table 1).

Proposition 1 (Characterization of the selection event) *The conditioning event* (8) *can be characterized as*

$$\left\{\widehat{\Gamma}^{(k)} = \gamma^{(k)} \; \forall \, k \in [K]\right\} = \left\{\text{sign}(\widehat{\mathbf{B}}) = \mathbf{s}, \; \widehat{\mathbf{Z}} = \mathbf{z}\right\}.$$

To avoid any confusion, we will hereafter refer to the conditioning event in (8) as the selection event.

3.2 Marginal distribution of the MLE

One common method to infer for β_E is through a maximum likelihood approach. In a distributed setting, we consider a maximum likelihood estimator (MLE) that is obtained by aggregating local estimators from each of the machines. Assuming that E is a fixed set, we derive the joint marginal distribution of this MLE and the randomization variables $\mathbf{\Omega}$ without considering the model selection process. This is the first step in obtaining our selective likelihood, before we condition the marginal distribution on the selection event.

Let

$$\widehat{\beta}_E^{(k)} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{\sqrt{n}\rho_k} \sum_{i \in \mathcal{C}_k} A(x_{i,E}^\intercal \beta) - y_i x_{i,E}^\intercal \beta$$

be the local MLE at machine k, and let

$$\widehat{\mathcal{I}}_{E,E}^{(k)} = \frac{1}{n_k} (X_E^{(k)})^\intercal \widehat{W}^k X_E^{(k)}, \quad \text{where } \widehat{W}^{(k)} = \operatorname{diag}\left(\nabla^2 A(X_E^{(k)} \widehat{\beta}_E^{(k)})\right),$$

be the corresponding observed Fisher information (obs-FI) matrix. Let $\widehat{\mathcal{I}}_{E,E} = \sum_{k=0}^K \rho_k \widehat{\mathcal{I}}_{E,E}^{(k)}$. Define the aggregated MLE as

$$\widehat{\beta}_E = \widehat{\mathcal{I}}_{E,E}^{-1} \sum_{k=0}^K \rho_k \widehat{\mathcal{I}}_{E,E}^{(k)} \widehat{\beta}_E^{(k)}. \tag{9}$$

Let $\mathcal{I}$ denote the Fisher information $\mathbb{E}\left[\frac{1}{n} X^\intercal \operatorname{diag}(\nabla^2 A(X_E \beta_E)) X\right]$.

In order to derive this marginal distribution of the aggregated MLE $\widehat{\beta}_E$, we state a few conditions.

Assumption 1 *For $k \in [K]$, let $\widetilde{E}^{(k)} = E \setminus E^{(k)}$. For $j \in \widetilde{E}^{(k)}$, either $\beta_{E,j} = O(n^{-1/2})$ or $X_{E^{(k)}} \mathcal{I}_{E^{(k)},E^{(k)}}^{-1} \mathcal{I}_{E^{(k)},j} = X_j$.*

Assumption 2 *The aggregated MLE $\widehat{\beta}_E$ can be written as*

$$\sqrt{n}(\widehat{\beta}_E - \beta_E) = -\mathcal{I}_{E,E}^{-1} \nabla \ell(\beta_E) + o_p(1), \quad n \to \infty.$$

Assumption 1 states conditions on predictors that are present in the selected model but are not selected by machine k. This condition suggests that our asymptotic assertions remain valid as long as these predictors are either weak in strength or have a high partial correlation with a predictor in the selected set $E^{(k)}$. Assumption 2 states that the aggregated MLE has an asymptotically linear representation. This condition is met when the aggregated MLE is asymptotically equivalent to the standard MLE. The latter has been shown to hold under some mild regularity conditions in Lin and Xi (2011).

The following result provides the asymptotic distribution of the randomization variables, with a detailed proof in Appendix A.1.

Proposition 2 *Suppose $n_k/n \to \rho_k$ as $n \to \infty$ for $k \in [K]$ and $\sum_{k=1}^K \rho_k < 1$. Let $U = \operatorname{diag}(\rho_1^{-1}, \ldots, \rho_K^{-1}) - \mathbf{1}_{K \times K}$, where $\mathbf{1}_{K \times K}$ denotes the $K \times K$ matrix with all entries equal to 1. Define*

$$\Sigma_\Omega = U \otimes \mathcal{I},$$

the Kronecker product of U and $\mathcal{I}$. Under Assumptions 1 and 2, we have

$$\sqrt{n}\,\mathbf{\Omega} \overset{d}{\Rightarrow} \mathcal{N}_{pK}\left(\mathbf{0}, \Sigma_{\Omega}\right), \quad n \to \infty.$$

The model selection process is not solely a function of the aggregated MLE and the randomization variables, as it also depends on variables that were not included in the selected GLM. Therefore, we consider an additional statistic, defined as:

$$\widehat{\beta}^{\perp}_{-E} = \frac{1}{n} X^{\intercal}_{-E}(\nabla A(X_E \widehat{\beta}_E) - Y),$$

which involves variables that were not selected by the lasso.

The next theorem derives the marginal distribution of $\widehat{\beta}_E$, $\widehat{\beta}^{\perp}_{-E}$, and $\mathbf{\Omega}$.

Theorem 3 *Under the same assumptions as Proposition 2, it holds that*

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_E - \beta_E \\ \widehat{\beta}^{\perp}_{-E} \\ \mathbf{\Omega} \end{pmatrix} \overset{d}{\Rightarrow} \mathcal{N}_{p(K+1)}\left(\mathbf{0}, \begin{pmatrix} \mathcal{I}^{-1}_{E,E} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}/\mathcal{I}_{E,E} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{\Omega} \end{pmatrix}\right),$$

where $\mathcal{I}/\mathcal{I}_{E,E} = \mathcal{I}_{-E,-E} - \mathcal{I}_{-E,E}\mathcal{I}^{-1}_{E,E}\mathcal{I}_{E,-E}$ is the Schur complement of $\mathcal{I}_{E,E}$ in $\mathcal{I}$.

The proof of Theorem 3 is deferred to Appendix A.2.

It is worth noting from this result that $\widehat{\beta}^{\perp}_{-E}$ is an ancillary statistic for the parameters in the selected GLM. Furthermore, the randomization variables are independent of the statistics $\widehat{\beta}_E$ and $\widehat{\beta}^{\perp}_{-E}$, which can be seen from the block diagonal covariance matrix in the limiting Gaussian distribution.

3.3 Selective likelihood

In what follows, we base inference on a conditional likelihood in the selected GLM. This likelihood, referred to as the selective likelihood, is stated in Theorem 4. It is derived from the asymptotic distribution of the aggregated MLE, as stated in Theorem 3, after conditioning on the event described in Proposition 1.

Before presenting this likelihood, we specify a technical condition and introduce the matrices involved in this likelihood.

Assumption 3 *Assume that the distribution of*

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_E - \beta_{E,n} \\ \widehat{\beta}^{\perp}_{-E} \\ \mathbf{\Omega} \end{pmatrix}$$

is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{p(K+1)}$. Let $\{p_n\}_{n\geq 1}$ be the corresponding sequence of densities. Assume that $\{p_n\}_{n\geq 1}$ is uniformly equicontinuous and bounded.

The condition in Assumption 3 together with the weak convergence in Theorem 3 implies that the density of $(\widehat{\beta}_E, \widehat{\beta}^{\perp}_{-E}, \mathbf{\Omega})$ converges to the corresponding limiting Gaussian density, uniformly on compact subsets of $\mathbb{R}^{p(K+1)}$.

Define $g_k^{(j)} = \gamma^{(j)}_{E^{(k)}} \in \mathbb{R}^{d_k}$ for $j, k \in [K]$, where $g_k^{(j)}$ collects the coordinates of $\gamma^{(j)}$ belonging to $E^{(k)}$. Define $g^{(j)} = \gamma^{(j)}_E \in \mathbb{R}^d$, where $g^{(j)}$ contains the coordinates of $\gamma^{(j)}$ belonging to E. Let $\rho_0 = 1 - \sum_{k\in[K]} \rho_k$. Let the matrices

$$\Xi \in \mathbb{R}^{\bar{d}\times\bar{d}},\ \Psi \in \mathbb{R}^{\bar{d}\times d},\ \tau \in \mathbb{R}^{\bar{d}},\ \Theta \in \mathbb{R}^{d\times d},\ \Pi \in \mathbb{R}^{d\times d},\ \kappa \in \mathbb{R}^d$$

be defined as follows. The (j,k) block of Ξ^{-1} is a $d^{(j)} \times d^{(k)}$ matrix given by

$$\left\{\Xi^{-1}\right\}_{j,k} = \begin{cases} \left(\rho_k + \dfrac{\rho_k^2}{\rho_0}\right)\mathcal{I}_{E^{(k)},E^{(k)}} & \text{if } j = k, \\ \dfrac{\rho_j\rho_k}{\rho_0}\mathcal{I}_{E^{(j)},E^{(k)}} & \text{if } j \neq k. \end{cases}$$

The $(k,1)$ block of $\Xi^{-1}\Psi$ and $\Xi^{-1}\tau$ are given by

$$\left\{\Xi^{-1}\Psi\right\}_k = \frac{\rho_k}{\rho_0}\mathcal{I}_{E^{(k)},E}; \quad \left\{\Xi^{-1}\tau\right\}_k = -\rho_k g_k^{(k)} - \frac{\rho_k}{\rho_0}\sum_{j=1}^{K}\rho_j g_k^{(j)}.$$

Further, let

$$\Theta^{-1} = \frac{1}{\rho_0}\mathcal{I}_{E,E} - \Psi^\intercal\Xi^{-1}\Psi; \quad \Theta^{-1}\Pi = \mathcal{I}_{E,E}; \quad \Theta^{-1}\kappa = \Psi^\intercal\Xi^{-1}\tau + \sum_{j=1}^{K}\frac{\rho_j}{\rho_0}g^{(j)}.$$

Let $\varphi(\cdot;\mu,\Sigma)$ denote the density function of the normal distribution $\mathcal{N}(\mu,\Sigma)$. Let $\mathcal{O} = \{v \in \mathbb{R}^{\bar{d}} : \text{sign}(v) = \text{sign}(\mathbf{s})\}$ denote the orthant in $\mathbb{R}^{\bar{d}}$ based on the sign vector $\mathbf{s}$.

Theorem 4 (Selective likelihood) *Suppose that the conditions stated in Theorem 3 and Assumption 3 hold. The selective likelihood function, based on the asymptotic distribution of* $\sqrt{n}\widehat{\beta}_E \,\Big|\, \left\{\widehat{\Gamma}^{(k)} = \gamma^{(k)}\ \forall\, k \in [K]\right\}$, *is equal to*

$$f\left(\beta_E; \widehat{\beta}_E, \mathbf{s}, \mathbf{z}\right) = \frac{\varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_E + \kappa, \Theta)}{\mathbb{P}\left[\sqrt{n}\widehat{\mathbf{B}} \in \mathcal{O} \mid \widehat{\mathbf{Z}} = \mathbf{z}\right]},$$

where

$$\mathbb{P}\left[\sqrt{n}\widehat{\mathbf{B}} \in \mathcal{O} \mid \widehat{\mathbf{Z}} = \mathbf{z}\right] = \int \varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_E + \kappa, \Theta)\cdot \varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_E + \tau, \Xi)\cdot \mathbf{1}\left\{\widehat{\mathbf{B}} \in \mathcal{O}\right\}\mathrm{d}\widehat{\beta}_E\mathrm{d}\widehat{\mathbf{B}}.$$

The proof is provided in Appendix A.3.

With this selective likelihood, we proceed to compute the selective MLE and its associated Fisher information to perform inference as outlined in Section 2.3. In Section 4, we offer an approximation to this selective likelihood, which in turn enables an efficient way to compute the selective MLE and the corresponding Fisher information matrix.

3.4 Algorithm

To conclude this section, we identify and specify the information that must be exchanged between the central machine and the local machines in order to carry out inferences in the selected GLM. Three rounds of communication are required for the central machine to perform selective inference:

1. Local machines send $E^{(k)}$ to the central machine.
2. The central machine sends the aggregated model E back to the local machines.
3. Local machines send local MLE $\widehat{\beta}_E^{(k)}$, local Fisher information $\widehat{\mathcal{I}}_{E,E}^{(k)}$, and the subgradient variables $\gamma_E^{(k)}$ to the central machine.

This is summarized in Algorithm 1.

Note that the first two rounds of communication involve only the indices of the selected variables, which lead to the selected model E.

It is obvious that the local MLE $\widehat{\beta}_E^{(k)}$ and local Fisher information $\widehat{\mathcal{I}}_{E,E}^{(k)}$ are needed at the central machine to compute the aggregated MLE and aggregated Fisher information, as defined in Equation (9). Besides these statistics, the selective likelihood derived in Theorem 4 involves the subgradients $\gamma_{E^{(k)}}^{(j)}$ and $\gamma_E^{(j)}$ for $j, k \in [K]$ in order to correct for the bias from model selection. Hence, in order to calculate the selective likelihood, the central machine needs the subgradients $\gamma^{(j)}$ with indices from $E \cup \bigcup_{k\in[K]} E^{(k)}$ sent from all local machines. Since we assumed for simplicity that $E = \bigcup_{k\in[K]} E^{(k)}$, machine j simply needs to send $\gamma_E^{(j)}$ to the central machine.

Note that $\widehat{\mathcal{I}}_{E,E}^{(k)}$ has a size of d^2, resulting in a communication cost of $O(d^2)$ for each local machine, which does not depend on the original feature dimension p. Therefore, as long as a sparse model is selected, the communication cost for making inferences in the selected GLM remains relatively low. Extensions for more general aggregation rules are given in Appendix D.

4. Approximate selective MLE

In the previous section, we obtained the selective likelihood function. As per Theorem 4, the selective log-likelihood is equal to

$$\log \varphi\left(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_E + \kappa, \Theta\right) - \log \mathbb{P}\left[\sqrt{n}\widehat{\mathbf{B}} \in \mathcal{O} \;\middle|\; \widehat{\mathbf{Z}} = \mathbf{z}\right]. \tag{10}$$

The main focus of this section is to provide an approximation for the second term (in the log-likelihood) by using a large-deviation limit for the log-probability. This approximation is made under the following moment and regularity conditions, which are commonly assumed to ensure the existence of the limit.

Consider a real-valued sequence a_n that goes to infinity as $n \to \infty$, and $a_n = o(n^{1/2})$. Assume $\beta_E = \beta_{E,n}$ satisfies that $\sqrt{n}\beta_{E,n} = a_n\beta_E^* \in \mathbb{R}^{|E|}$, where β_E^* does not depend on n.

Algorithm 1: Communication of information

STEP 1: Variable selection at local machines

Machine k solves Problem (1) and sends $E^{(k)} = \text{Support}(\widehat{\beta}^{\Lambda,(k)})$ to the central machine.

STEP 2: Model aggregation at the central machine

The central machine forms $E = \bigcup_{k\in[K]} E^{(k)}$ and sends E back to the local machines.

STEP 3: Communication of summary statistics

Local machines communicate the following statistics to the central machine:

$$\text{local MLE and Fisher information: } \widehat{\beta}_E^{(j)},\ \widehat{\mathcal{I}}_{E,E}^{(j)}; \qquad \text{subgradient: } \gamma_E^{(j)}$$

From the proof of Theorem 3, we have

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_E - \beta_{E,n}\\ \widehat{\beta}^{\perp}_{-E}\\ \mathbf{\Omega}\end{pmatrix} = \sqrt{n}\bar{E}_n + R_n, \tag{11}$$

where $\bar{E}_n = \frac{1}{n}\sum_{i=1}^n e_{i,n}$ is the average of n i.i.d. observations, and $R_n = o_p(1)$.

Assumption 4 (Moment condition and convergence of remainder) *Assume that*

$$\mathbb{E}\left[\exp(\lambda\|e_{1,n}\|_2)\right] < \infty$$

for some $\lambda \in \mathbb{R}^+$*, and that*

$$\lim_{n\to\infty}\frac{1}{a_n^2}\log\mathbb{P}\left[\frac{1}{a_n}\|R_n\|_2 > \epsilon\right] = -\infty$$

for any $\epsilon > 0$*, where* $e_{i,n}$ *and* R_n *are from the asymptotic linear representation in* (11)*.*

Assumption 4 assumes the existence of an exponential moment in a neighborhood of zero, which is needed to justify a Laplace-type approximation for probabilities involving the mean of n i.i.d. variables, which in our case is $\bar{E}_n$. The second condition in this assumption helps extend this justification to $\bar{E}_n$ with an added $o_p(1)$ error term, R_n.

Assumption 5 *For a fixed convex set* $\mathcal{R}_0 \subseteq \mathbb{R}^{p(K+1)}$*, and for* $O = O_p(1)$*, we impose the condition that*

$$\lim_{n\to\infty}\frac{1}{a_n^2}\left\{\log\mathbb{P}\left[\frac{1}{a_n}\begin{pmatrix}\sqrt{n}\widehat{\beta}_E\\ \sqrt{n}\widehat{\beta}^{\perp}_{-E}\\ \sqrt{n}\mathbf{\Omega}\end{pmatrix}\in\mathcal{R}_0\right] - \log\mathbb{P}\left[\frac{1}{a_n}\begin{pmatrix}\sqrt{n}\widehat{\beta}_E\\ \sqrt{n}\widehat{\beta}^{\perp}_{-E}\\ \sqrt{n}\mathbf{\Omega}\end{pmatrix} + \frac{1}{a_n}O\in\mathcal{R}_0\right]\right\} = 0.$$

In terms of $\widehat{\beta}_E$, $\widehat{\beta}^{\perp}_{-E}$, and $\mathbf{\Omega}$, the probability of the selection event takes the form

$$\mathbb{P}\left[\frac{1}{a_n}\begin{pmatrix}\sqrt{n}\widehat{\beta}_E \\ \sqrt{n}\widehat{\beta}^{\perp}_{-E} \\ \sqrt{n}\mathbf{\Omega}\end{pmatrix} + \frac{1}{a_n}O \in \mathcal{R}_0\right],$$

(see Appendix B.1). Therefore, Assumption 5 allows us to approximate the probability of the selection event up to an $o_p(1)$ remainder.

Theorem 5 *Suppose that the conditions in Assumption 4 and Assumption 5 are met. Define*

$$L_n = \inf_{b,B}\left\{\frac{1}{2}\left(b - \Pi\beta^*_E - \frac{1}{a_n}\kappa\right)^{\intercal}\Theta^{-1}\left(b - \Pi\beta^*_E - \frac{1}{a_n}\kappa\right)\right.$$
$$\left. + \frac{1}{2}\left(B - \Psi b - \frac{1}{a_n}\tau\right)^{\intercal}\Xi^{-1}\left(B - \Psi b - \frac{1}{a_n}\tau\right) + \frac{1}{a_n^2}\mathrm{Barr}_{\mathcal{O}}(a_nB)\right\},$$

where $\mathrm{Barr}_{\mathcal{O}}(x) = \sum_i \log(1 + \frac{1}{\mathbf{s}_i x_i})$. *Then, we have*

$$\lim_{n\to\infty}\frac{1}{a_n^2}\log\mathbb{P}\left[\sqrt{n}\widehat{\mathbf{B}} \in \mathcal{O} \;\middle|\; \widehat{\mathbf{Z}} = \mathbf{z}\right] + L_n = C_0,$$

where C_0 *is a constant that does not depend on* β^*_E.

The proof is provided in Appendix B.1. As a consequence of Theorem 5, we can substitute the log-probability in (10) by

$$-\inf_{b,B}\left\{\frac{1}{2}\left(a_nb - a_n\Pi\beta^*_E - \kappa\right)^{\intercal}\Theta^{-1}\left(a_nb - a_n\Pi\beta^*_E - \kappa\right)\right.$$
$$\left. + \frac{1}{2}\left(a_nB - a_n\Psi b - \tau\right)^{\intercal}\Xi^{-1}\left(a_nB - a_n\Psi b - \tau\right) + \mathrm{Barr}_{\mathcal{O}}(a_nB)\right\}, \tag{12}$$

ignoring the additive constant in the limit. Finally, we reparameterize $a_nb = \sqrt{n}v$ and $a_nB = \sqrt{n}V$, which gives the following approximation of the selective log-likelihood

$$\log\varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_{E,n} + \kappa, \Theta)$$
$$+\inf_{v,V}\left\{\frac{1}{2}\left(\sqrt{n}v - \sqrt{n}\Pi\beta_{E,n} - \kappa\right)^{\intercal}\Theta^{-1}\left(\sqrt{n}v - \sqrt{n}\Pi\beta_{E,n} - \kappa\right)\right.$$
$$\left. + \frac{1}{2}\left(\sqrt{n}V - \sqrt{n}\Psi v - \tau\right)^{\intercal}\Xi^{-1}\left(\sqrt{n}V - \sqrt{n}\Psi v - \tau\right) + \mathrm{Barr}_{\mathcal{O}}(\sqrt{n}V)\right\},$$

The score and curvature of this selective likelihood yield the selective MLE and the selective obs-FI matrix. The derivation of these two estimators follows the steps in Panigrahi and Taylor (2023) for the standard Gaussian regression problem. In the interest of completeness, the expressions of these estimators are provided below.

Theorem 6 *Consider solving the optimization problem*

$$\widehat{V}^{\star}_{\widehat{\beta}_E} = \underset{V \in \mathbb{R}^{\bar{d}}}{\operatorname{argmin}} \frac{1}{2}(\sqrt{n}V - \Psi\sqrt{n}\widehat{\beta}_E - \tau)^{\intercal}\Xi^{-1}(\sqrt{n}V - \Psi\sqrt{n}\widehat{\beta}_E - \tau) + \mathrm{Barr}_{\mathcal{O}}(\sqrt{n}V). \quad (13)$$

The maximizer of the approximate selective likelihood and the observed Fisher information matrix are equal to

$$\Pi^{-1}\widehat{\beta}_E - \frac{1}{\sqrt{n}}\Pi^{-1}\kappa + \mathcal{I}^{-1}_{E,E}\Psi^{\intercal}\Theta^{-1}\left(\Psi\widehat{\beta}_E + \frac{1}{\sqrt{n}}\tau - \widehat{V}^{\star}_{\widehat{\beta}_E}\right), \quad (14)$$

$$\mathcal{I}_{E,E}\left(\Theta^{-1} + \Psi^{\intercal}\Xi^{-1}\Psi - \Psi^{\intercal}\Xi^{-1}\left(\Xi^{-1} + \nabla^2\mathrm{Barr}_{\mathcal{O}}\left(\sqrt{n}\widehat{V}^{\star}_{\widehat{\beta}_E}\right)\right)^{-1}\Xi^{-1}\Psi\right)^{-1}\mathcal{I}_{E,E}, \quad (15)$$

respectively.

The proof is provided in Appendix B.2. In practice, the matrices $\Pi, \kappa, \Theta, \Psi, \tau, \Xi$ are computed with the observed Fisher information $\widehat{\mathcal{I}}$, which are then used to compute the selective MLE and the selective obs-FI in (14) and (15), respectively. The algorithm for conducting inference based on the approximate selective MLE is summarized in Algorithm 2.

Algorithm 2: Approximate selective MLE-based inference

Compute the aggregated MLE $\widehat{\beta}_E$ defined by Equation (9).

Solve the $\bar{d}$-dimensional convex optimization in (13).

Compute $\widehat{\beta}^{(S)}_E$ and $\widehat{\mathcal{I}}^{(S)}_{E,E}$ as stated in (14) and (15).

Let

$$\widehat{\sigma}^{(S)}_j = \sqrt{\left(\widehat{\mathcal{I}}^{(S)}_{E,E}\right)^{-1}_{j,j}}, \quad j \in [d].$$

Compute the two-sided p-value for $H_0 : \beta_{E,j} = 0$ as

$$2 \cdot \min\left(\Phi\left(\frac{\sqrt{n}}{\widehat{\sigma}^{(S)}_j}\widehat{\beta}^{(S)}_{E,j}\right), \bar{\Phi}\left(\frac{\sqrt{n}}{\widehat{\sigma}^{(S)}_j}\widehat{\beta}^{(S)}_{E,j}\right)\right)$$

where $\Phi = 1 - \bar{\Phi}$ is the CDF of the standard normal distribution.

Compute the two-sided $100 \cdot (1-\alpha)\%$ confidence interval for $\beta_{E,j}$ as

$$\widehat{\beta}^{(S)}_{E,j} \pm z_{1-\alpha/2} \cdot \frac{\widehat{\sigma}^{(S)}_j}{\sqrt{n}}.$$

5. Addressing p-value lotteries

When dealing with sparse regression, it is often feasible to construct p-values after reducing the number of variables to a manageable size. One common method is sample-splitting,

as described in Wasserman and Roeder (2009), where variables are first selected on a subset of the data and then p-values are reported using classical least squares estimation on the remaining samples. Variables that are not selected are assigned a p-value equal to 1. A more powerful alternative to sample-splitting is carving, introduced in Fithian et al. (2014) through conditioning. However, results produced by sample splitting or carving are overly sensitive to the randomness involved in partitioning the data. This issue has been widely reported in literature as the p-value lottery problem. See, for example, the paper by Meinshausen et al. (2009).

To address the p-value lottery problem, Meinshausen et al. (2009) aggregate p-values from repeated splitting, and more recently, Schultheiss et al. (2021) propose to aggregate p-values after repeated carving on random splits of data. We refer to the former procedure as multi-splitting and the latter procedure as multi-carving. With increasing numbers of replicates, the results are expected to be less sensitive to the randomness from the splits.

Suppose one conducts multi-splitting or multi-carving B times, and obtains the p-values $p_j^{(b)}$, for $b \in [B]$, and $j \in [p]$. This is followed by aggregating the B p-values through their empirical quantiles

$$Q_j(\gamma) := q_\gamma \left(\left\{ \frac{1}{\gamma} p_j^{(b)},\, b \in [B] \right\} \right) \wedge 1,$$

where q_γ denotes the γ-th empirical quantile. One can also minimize over γ and use

$$P_j := \left[(1 - \log(\gamma_{\min})) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma) \right] \wedge 1. \tag{16}$$

The aggregation scheme produces valid p-values as long as all p-values are individually valid.

Below, we show that our proposal in the paper can be easily adapted to address the p-value lottery problem without recourse to MCMC sampling. We proceed as multi-splitting and multi-carving, i.e., we use a subsample of size n_1 for variable selection. For inference, we reuse data from selection by conditioning on the event of selection. We repeat this procedure B times and aggregate the p-values as above. Moreover, in each replicate, we can draw K random subsets of size n_1 with replacement. A base model is selected using each subset, and the K base models are aggregated as done in (2). To conduct selective inference, a similar procedure can be used, with a slight modification in the distribution of the randomization variable Ω. Specifically, $\sqrt{n}\Omega$ still converges to a normal distribution but with a different covariance matrix, as stated in Lemma 7.

Lemma 7 *If the K subsets $D^{(1)}, \ldots, D^{(K)}$ are independent random samples of the data set D (rather than disjoint partitions) and each subset has size $[\rho n]$, then Theorem 3 holds with*

$$\Sigma_\Omega = \frac{1-\rho}{\rho} I_K \otimes \mathcal{I}.$$

The proof is provided in Appendix C.

Following the same argument as in Theorem 4, the asymptotic selective likelihood function takes the same form as the expression in Theorem 4, with slightly different expressions

Algorithm 3: Multiple carving

for $k = 1, \ldots, K$ **do**
  Perform n_1-out-of-n subsampling to form data set $D^{(k)}$.
  Solve the lasso problem with data $D^{(k)}$ and select variable set $E^{(k)}$.
Construct the aggregated model $E = \texttt{Aggregate}(E^{(1)}, \ldots, E^{(K)})$.
Compute the MLE $\widehat{\beta}_E$ and Fisher information $\widehat{\mathcal{I}}_{E,E}$ in the selected GLM (3) using all the data.
Compute matrices $\Xi, \Psi, \tau, \Theta, \Pi, \kappa$ as given in Equation (17).
Perform selective inference for β_E using the approximate selective-MLE based method in Algorithm 2.

of the matrices $\Xi, \Psi, \tau, \Theta, \Pi, \kappa$ due to the change of the covariance matrix Σ_Ω. There matrices are now calculated as

$$\{\Xi^{-1}\}_{j,k} = \delta_{j,k}\frac{\rho}{1-\rho}\mathcal{I}_{E^{(j)},E^{(j)}};\;\; \{\Xi^{-1}\Psi\}_k = \frac{\rho}{1-\rho}\mathcal{I}_{E^{(k)},E};\;\; \{\Xi^{-1}\tau\}_k = -\frac{\rho}{1-\rho}g_k^{(k)}; \qquad (17)$$

$$\Theta^{-1} = \left(1 + \frac{K\rho}{1-\rho}\right)\mathcal{I}_{E,E} - \Psi^\intercal\Xi^{-1}\Psi;\;\; \Theta^{-1}\Pi = \mathcal{I}_{E,E};\;\; \Theta^{-1}\kappa = \Psi^\intercal\Xi^{-1}\tau + \frac{\rho}{1-\rho}\sum_{j=1}^{K} g^{(j)}.$$

The entire procedure is summarized in Algorithm 3.

When $K = 1$ subset is drawn in each of the B replicates, this procedure closely resembles the multi-carving procedure described in Schultheiss et al. (2021). When $B = 1$ and $K = 1$, it reduces to the data carving of Fithian et al. (2014). However, the multi-carving method of Schultheiss et al. (2021) conditions on the indices of the samples used for selection which results in a polyhedral event similar to the event in Lee et al. (2016). In contrast, our method marginalizes over the randomness involved in subsampling instead of conditioning on it. This is captured in the definition of our randomization variable $\mathbf{\Omega}$. As shown in the simulation results in Section 6.2, our method achieves higher power compared to the method that conditions on more.

6. Experiments

In this section, we demonstrate the numerical performance of our proposed procedure. Our code can be accessed from the GitHub repository `https://github.com/snigdhagit/Distributed-Selectinf`.

6.1 Experiments with distributed data sets

The data in our experiments is simulated as follows. For $i \in [n]$, we draw $x_i \sim \mathcal{N}_p(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.9$, $p = 100$. In the first model, we draw a real-valued response from a linear model as $y_i \sim \mathcal{N}(x_i^\intercal\beta, \sigma^2)$ with $\sigma^2 = 1$. The dispersion parameter σ^2 is estimated by $\hat{\sigma}^2 = \frac{1}{n-d}\sum_{i=1}^{n}(y_i - x_{i,E}^\intercal\widehat{\beta}_E)^2$. In our second model, we draw a binary response from a logistic-linear model as $y_i \sim \text{Bernoulli}(1/(1 + e^{-x_i^\intercal\beta}))$. Observation i is

independent of all the other observations in our data set. There are 5 non-zero coefficients in our model; each non-zero β_j is equal to $\pm\sqrt{2c\log p}$, where the sign is random and c is referred to as the "signal strength." Unless otherwise specified, the signal strength c is 0.7 for the linear model and 1 for the logistic model. The regularization parameter λ is set to $\sqrt{2\log p}$ for linear regression and $\sqrt{0.5\log p}$ for logistic regression. The n observations are partitioned into $K+1$ disjoint subsets $D^{(0)}, D^{(1)}, \ldots, D^{(K)}$. Subsets 1 through K, representing the data stored at K local machines, are used for variable selection. Subset 0, representing the data at the central machine, is used only at the time of selective inference.

We design three different scenarios to investigate the performance of our procedure. We conduct 500 rounds of simulations in each scenario.

(I). In Scenario 1, we vary the number of distributed data sets $K \in \{2, 4, 6, 8\}$. Each local machine has $n_k = [8000/K]$ samples, and the central machine has access to 1000 samples.

(II). In Scenario 2, we vary the signal strength c. For linear regression, we vary c among $\{0.3, 0.5, 0.7, 0.9\}$; while for logistic regression, we vary c among $\{0.5, 1, 1.5, 2\}$. We fix $K = 2$. The central machine has 2000 samples, and each of the two local machines has 4000 samples.

(III). In Scenario 3, we vary the number of samples that are reserved only for selective inference at the central machine. Specially, we vary $n_0 \in \{250, 500, 1000, 2000\}$. We fix $K = 3$ and $n_k = 2000$ for $1 \leq k \leq K$.

We report the coverage proportions and average interval lengths for β_E in Figure 2 and Figure 3, for the linear model and the logistic model, respectively. The target parameter β_E is defined as

$$\beta_E := \underset{\beta_E \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} \mathbb{E}\left[\ell(y_i; x_{i,E}^\intercal \beta_E)\right],$$

where ℓ represents the corresponding GLM loss and the expectation is taken over the true distribution of y_i. Specifically, for linear regression, $\ell(y_i;\theta) = -\frac{1}{2}(y_i - \theta)^2$ and the true distribution is $y \sim \mathcal{N}(x_i^\intercal \beta, \sigma^2)$. For logistic regression, $\ell(y_i;\theta) = -[y_i \log \frac{1}{1+e^{-\theta}} + (1 - y_i)\log(1 - \frac{1}{1+e^{-\theta}})]$ and the true distribution is $y_i \sim \text{Bernoulli}((1 + e^{-x_i^\intercal \beta})^{-1})$. The target coverage of the confidence interval is 90%. The lengths of confidence intervals are indications of the statistical power associated with selective inference for β_E. Our method "Dist-SI" constructs the confidence intervals based on the approximate selective MLE method as introduced before. As a baseline for comparison, we consider the "Splitting" method, which uses only the data at the central machine to construct standard Wald confidence intervals. Error bars represent variations among 500 random replicates.

Observations. Across all scenarios, the confidence intervals produced by "Dist-SI" (approximately) attain the desired coverage probability. The "Splitting" method produces valid confidence intervals, but discards samples used by the local machines. The advantages of reusing data from the local machines are quite evident in the plots for the lengths of the confidence intervals. As expected, "Dist-SI" yields tighter confidence intervals and achieves higher power than the baseline procedure based on "Splitting" in all three scenarios. We

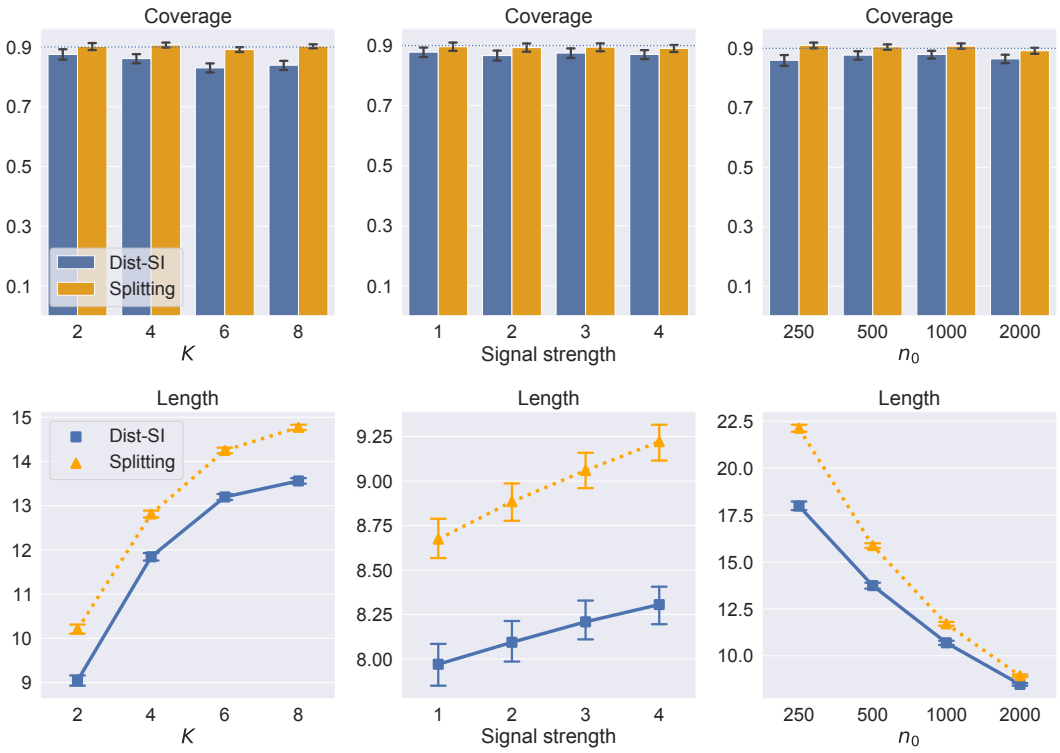


Figure 2: Coverage proportions (top panel) and average interval lengths (bottom panel) for the linear model. Left panel: varying K. Middle panel: varying signal strength. Right panel: varying n_0.

observe that interval lengths for both methods increase with K. This is because the final model, which is the union of the K models selected by local machines, is likely to be larger for larger K. In this case, the variance of $\widehat{\beta}_j$ tends to be larger. For a similar reason, interval lengths tend to increase with signal strengths as well. In Scenario 3, we see that both methods produce longer intervals when n_0 decreases, and as expected, the gap between "splitting" and "Dist-SI" is more pronounced with fewer samples at the central machine.

6.2 Experiments on p-value lotteries

In this section, we apply the suitable adaptation of our procedure to solve the p-value lottery problem, as described in Section 5. We compare our procedure with "Multi-carving" and "Multi-splitting" as proposed by Schultheiss et al. (2021) and Meinshausen et al. (2009), respectively. For the latter two algorithms, we use the implementation provided by Schultheiss et al. (2021) with code available on GitHub[1]. To avoid any confusion, we continue to refer our procedure as "Dist-SI", though we are no longer simulating distributed data sets.

1. `https://github.com/cschultheiss/Multicarving`. The original code is written in R, and we load them into Python when running our simulations, which might have contributed to slightly longer running times as reported in our findings.

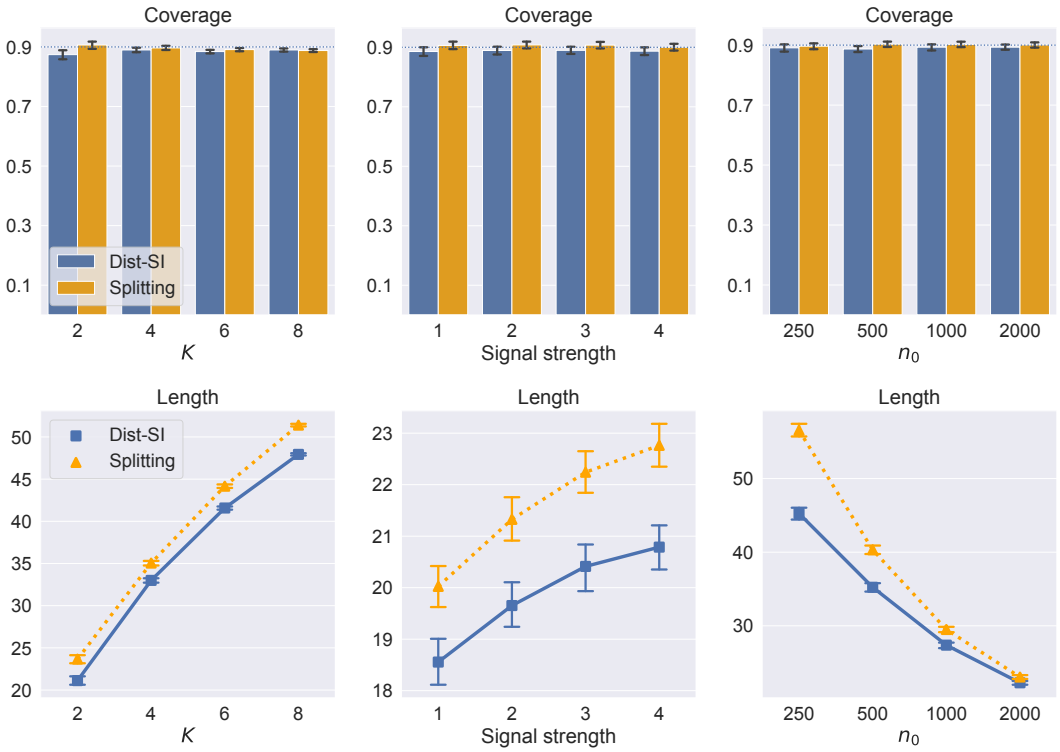


Figure 3: Coverage proportions (top panel) and average interval lengths (bottom panel) for the logistic model. Left panel: varying K. Middle panel: varying signal strength. Right panel: varying n_0.

We generate our data from the same linear model as described before, but now we use the same dimension as in Schultheiss et al. (2021). We take $n = 100$ and $p = 200$, and consider $s = 5$ nonzero coefficients with $\beta_j = \pm 2\sqrt{\log p}$. We use $B = 10$ replicates, and aggregate the p-values using formula (16) with $\gamma_{\min} = 0.1$. The proportion of samples used for variable selection is varied in the set $\{0.5, 0.6, \ldots, 0.9\}$. We fix the significance level at 0.1.

Again, the inferential target of these methods is not the actual generating β, unless the selected set E contains all the nonzero coefficients of β. Nevertheless, we can compare these methods by their accuracy in terms of detecting nonzero signals. Specifically, a coefficient β_j is predicted to be nonzero if the corresponding p-value is no larger than 0.1. We measure the accuracy of this prediction by the F-score, which is defined as

$$\text{F-score} := \frac{2 \cdot \text{True Positive}}{2 \cdot \text{True Positive} + \text{False Positive} + \text{False Negative}}$$

Note that the positives and negatives are computed with respect to the true underlying parameter β, which might be different from the inferential target post selection if the selected E fails to contain some nonzero coefficients of β. Besides computing the F-score, we compare the average run time for "Dist-SI" and "Multi-carving". The results are shown in Figure 4. In the left panel, we plot the F-scores of the three methods with varying proportions. The error bars are once again reported for 500 random repetitions. In the right panel, we plot the average run times of "Dist-SI" and "Multi-carving" on the log scale.

Observations. We find that our procedure has higher F-scores than the two previously proposed alternatives, "Multi-carving" and "Multi-splitting", for all values of sample proportion. Especially, a p-value in every replicate uses the full data after carefully discarding information that was used up for selecting predictors. The re-use of data from selection results in larger power over "Multi-splitting". Our procedure aligns with "Multi-carving", which also deploys conditional techniques to reuse data for hypothesis testing. However, a key distinction of our procedure with "Multi-carving" lies in how we use the randomization framework to characterize the selection event, and subsequently marginalize over this randomness to construct our p-values. In particular, we note that "Multi-carving" conditions on the randomization that is involved during variable selection on a random split of the data, whereas our procedure explicitly characterizes the distribution of randomization instead of simply conditioning on Ω. We believe that this difference between the two procedures shows up in our simulated findings as we note higher values of F-score with "Dist-SI". Unsurprisingly, our proposal is also faster than "Multi-carving" by about 100 times. This is because the latter procedure requires MCMC sampling, while our procedure only involves solving a convex optimization problem.

6.3 Experiments on medical data set

We illustrate an application of our procedure on a real data set that is publicly available on MIT's GOSSIS database Raffa et al. (2022). This data set contains records on intensive care unit (ICU) admissions from 192 hospitals, including patients' demographic information, and various medical measurements, and lab results. We only use the data sets from the four largest hospitals, among which three data sets are used for variable selection and the remaining one is reserved for selective inference. We focus on a regression problem with data

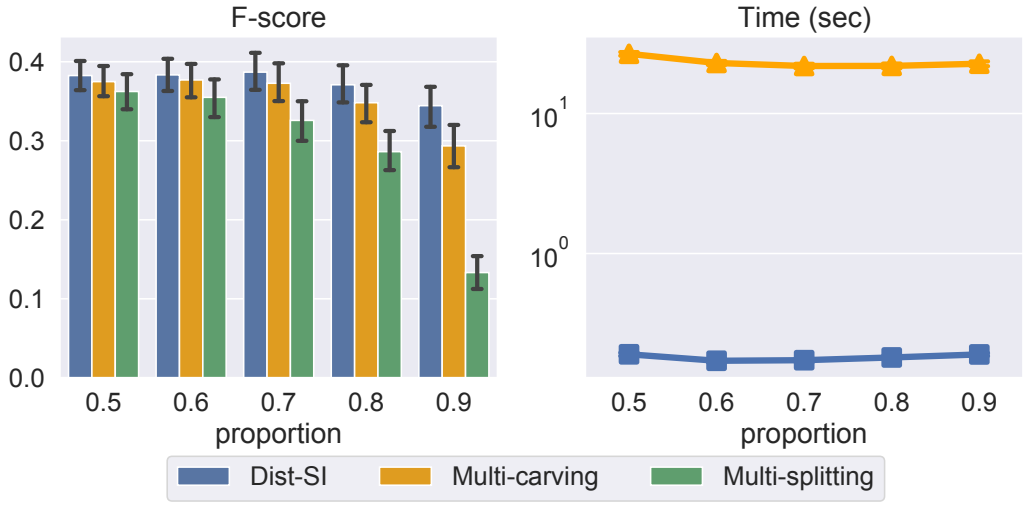


Figure 4: Compare Dist-SI with multi-carving and multi-splitting. The left panel shows the F-scores when using different proportions of samples for selection. The right panel shows the average running time.

from the first 24 hours of intensive care. The response in this problem is binary, and takes the value 1 if a patient admitted to an ICU has been diagnosed with Diabetes Mellitus, and is 0 otherwise. The same problem appeared in the 2021 Women in Data Science Datathon [2]. We remove variables with more than half missing values, and also remove rows with missing values. After preprocessing, we end up with 81 predictors. The three data sets used for variable selection have sample sizes ranging from 1633 to 1788, and the data set reserved for inference has 2000 samples.

For model selection, we run the logistic regression with lasso penalty. The regularization parameter is tuned with one extra data set with 893 samples. The selected GLM has 58 predictors. To construct confidence intervals for the 58 selected variables, we apply the proposed "Dist-SI" algorithm and "Splitting" as done in simulations. The significance level is set to be 0.1. "Dist-SI" reports 21 significant variables, while "Splitting" reports 13 significant variables, with 10 variables overlapping as significant in both methods. In Figure 5, we plot the confidence intervals for the regression coefficients that are rejected by either of the two procedures. The boxplot for the lengths of these intervals, in Figure 6, show that the median length of the "Dist-SI" intervals is smaller than the "Splitting" intervals by 67%. Additionally, the coefficient of variation is 1.8 and 3.9 for "Dist-SI" and "Splitting", respectively. This indicates that the dispersion of interval lengths for "Dist-SI" is smaller than "Splitting". On this instance, we see that "Splitting" yields a few very wide intervals. This is because the Hessian matrix based on data present at the central machine (reserved data set) is ill-conditioned. "Dist-SI" does not have this issue because it reuses data from the three hospitals for more powerful selective inference.

2. `https://www.kaggle.com/competitions/widsdatathon2021/data`. Accessed on on Dec. 17, 2022.

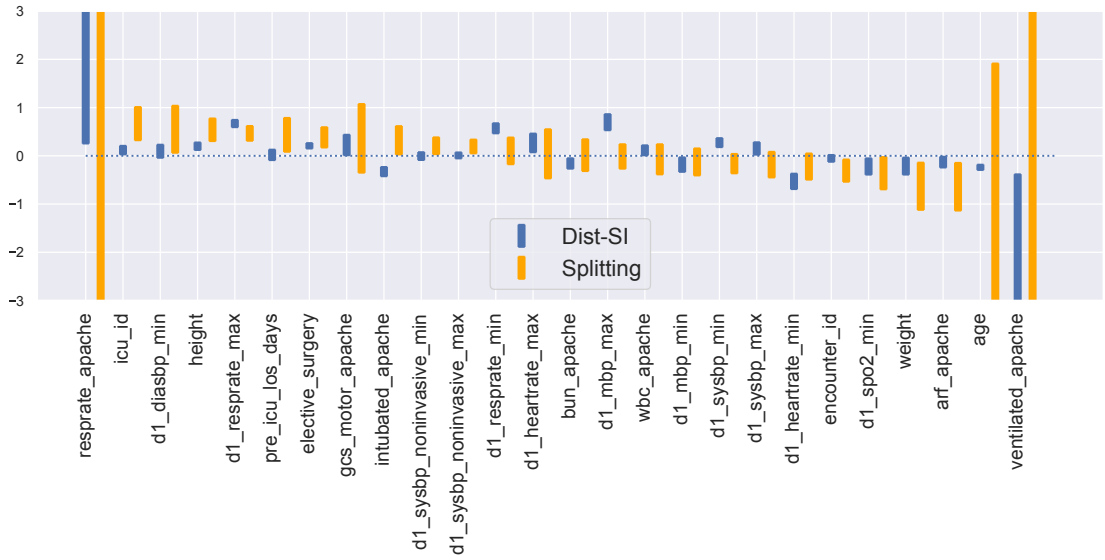


Figure 5: Confidence intervals for the coefficients that are rejected by either Dist-SI or sample splitting.

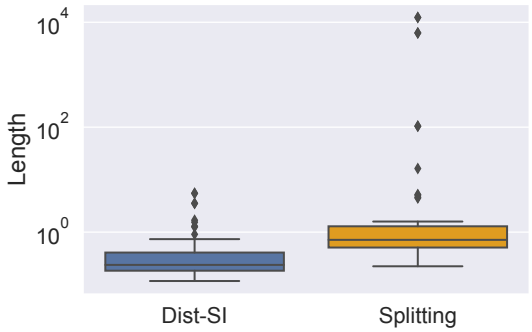


Figure 6: Boxplot of confidence interval lengths produced by Dist-SI and sample splitting. The y-axis is on the logarithmic scale.

7. Conclusion

Model selection appears to be routine practice when analyzing big data sets. Inference for data-dependent models and parameters is a very challenging goal, because sound procedures must rigorously account for randomness from the selection process. To the best of our knowledge, this paper is the first contribution that addresses selective inference with distributed data. We provide a rigorous procedure to construct confidence intervals and p-values when inference is sought in a generalized linear model with selected predictors. We identify a representation for selection in a common distributed setup, and provide an asymptotic selective likelihood by developing a novel randomized framework for our problem. Approximately-valid selective inference, based on our selective likelihood, takes a very simple form: our confidence intervals for the selected regression coefficients are centered around the MLE of the selective likelihood, and the variance of the MLE is estimated by the observed Fisher information matrix.

An appealing feature of our proposed inferential procedure is that we only require some aggregated information, with relatively low communication cost, from each machine. This feature allows an adaptation of our procedure to settings where various sites may not be willing to share their individual data sets. But, we note that there is room for improvement here, specially if various sites have not measured the same set of predictors.

Our paper also provides an efficient solution for the p-value lottery problem without relying on MCMC samplers. Our procedure bypasses the primary computational bottleneck in the earlier proposal (Schultheiss et al., 2021) by reducing selective inference to the solution of an optimization problem. The recently developed sampler in Liu (2023) is also applicable in our distributed setup to perform sampling-based inference, as an alternative to the MLE-based inference used in this paper.

Acknowledgments

S. Panigrahi's research is supported by NSF grants DMS-1951980 and DMS-2113342 and by the NSF CAREER Award DMS-2337882. S. Liu's research is partially supported by the Stanford Data Science Scholars program.

Appendix A. Proofs for Section 3

Supporting results are collected in Appendix A.3.1.

A.1 Proof of Proposition 2

Proof We write $\beta_E = \beta_{E,n}$ in the asymptotic analysis as $n \to \infty$. Define $\beta^*_{E^{(k)},n} = \mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},E}\beta_{E,n}$. We start with the decomposition

$$\frac{1}{n}X^\intercal(\nabla A(X\widehat{\beta}^{\Lambda,(k)}) - Y) = \frac{1}{n}X^\intercal(\nabla A(X_E\beta_{E,n}) - Y) + R_1^{(k)} + R_2^{(k)},$$

where

$$\begin{aligned} R_1^{(k)} &= \frac{1}{n}X^\intercal(\nabla A(X_{E^{(k)}}\beta^*_{E^{(k)},n}) - \nabla A(X_E\beta_{E,n})), \\ R_2^{(k)} &= \frac{1}{n}X^\intercal(\nabla A(X\widehat{\beta}^{\Lambda,(k)}) - \nabla A(X_{E^{(k)}}\beta^*_{E^{(k)},n})). \end{aligned}$$

In a similar fashion, we decompose the quantities based on local data $D^{(k)}$ as

$$\frac{1}{n_k}X^{(k)\intercal}(\nabla A(X^{(k)}\widehat{\beta}^{\Lambda,(k)}) - Y^{(k)}) = \frac{1}{n_k}X^{(k)\intercal}(\nabla A(X_E^{(k)}\beta_{E,n}) - Y^{(k)}) + r_1^{(k)} + r_2^{(k)}.$$

The decomposition in the above two displays allow us to write

$$\begin{aligned} \sqrt{n}\omega^{(k)} &= \sqrt{n}\Big\{\frac{1}{n}X^\intercal(\nabla A(X_E\beta_{E,n}) - Y) - \frac{1}{n_k}X^{(k)\intercal}(\nabla A(X_E^{(k)}\beta_{E,n}) - Y^{(k)}) \\ &\qquad + R_1^{(k)} + R_2^{(k)} - r_1^{(k)} - r_2^{(k)}\Big\} \\ &= \sqrt{n}\tilde{\omega}^{(k)} + \sqrt{n}R^{(k)}, \end{aligned} \tag{18}$$

where

$$R^{(k)} = R_1^{(k)} + R_2^{(k)} - r_1^{(k)} - r_2^{(k)},$$

and

$$\tilde{\omega}^{(k)} = \frac{1}{n}X^\intercal(\nabla A(X_E\beta_{E,n}) - Y) - \frac{1}{n_k}X^{(k)\intercal}(\nabla A(X_E^{(k)}\beta_{E,n}) - Y^{(k)}).$$

Let $\tilde{\Omega} \in \mathbb{R}^{pK}$ be the stack of $\tilde{\omega}^{(k)}$ for $1 \le k \le K$. It suffices to show that

$$\sqrt{n}\tilde{\Omega} \overset{d}{\Rightarrow} \mathcal{N}(\mathbf{0}, \Sigma_\Omega) \quad \text{and} \quad \sqrt{n}R^{(k)} \overset{p}{\to} 0. \tag{19}$$

To proceed with the proof, let $e_i = x_i(\nabla A(x_{i,E}^\intercal\beta_{E,n}) - y_i)$. It is easy to see that e_i are i.i.d. for all $1 \le i \le n$ with $\mathbb{E}[e_i] = 0$ and $\mathrm{Var}[e_i] \to \mathcal{I}$ under the selected model (3). It follows that

$$\sqrt{n}\tilde{\omega}^{(k)} = \sqrt{1-\rho_k}\frac{1}{\sqrt{n-n_k}}\sum_{i\notin\mathcal{C}_k} e_i - \frac{1-\rho_k}{\sqrt{\rho_k}}\frac{1}{\sqrt{n_k}}\sum_{j\in\mathcal{C}_k} e_j.$$

Hence, we have

$$\sqrt{n}\tilde{\omega}^{(k)} \overset{d}{\Rightarrow} \mathcal{N}\left(\mathbf{0}, \frac{1-\rho_k}{\rho_k}\mathcal{I}\right),$$

and

$$\mathrm{Cov}\left[\sqrt{n}\tilde{\omega}^{(j)},\, \sqrt{n}\tilde{\omega}^{(k)}\right] \to -\mathcal{I}$$

for $j \neq k$. This leads to

$$\sqrt{n}\begin{pmatrix}\tilde{\omega}^{(j)}\\ \tilde{\omega}^{(k)}\end{pmatrix} \overset{d}{\Rightarrow} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}\frac{1-\rho_j}{\rho_j}\mathcal{I} & -\mathcal{I}\\ -\mathcal{I} & \frac{1-\rho_k}{\rho_k}\mathcal{I}\end{pmatrix}\right),$$

which proves the first statement of (19). Lemma 8 and Lemma 9 below show that $\sqrt{n}R^{(k)} = o_p(1)$, thereby concluding the proof of (19).

Lemma 8 (Rate of $R_1^{(k)} - r_1^{(k)}$) *Let*

$$R_1^{(k)} = \frac{1}{n}X^\intercal(\nabla A(X_{E^{(k)}}\beta^*_{E^{(k)},n}) - \nabla A(X_E\beta_{E,n})),$$
$$r_1^{(k)} = \frac{1}{n_k}X^{(k)\intercal}(\nabla A(X^{(k)}_{E^{(k)}}\beta^*_{E^{(k)},n}) - \nabla A(X^{(k)}_{E}\beta_{E,n})).$$

Then $\sqrt{n}(R_1^{(k)} - r_1^{(k)}) = o_p(1)$.

Proof [Proof of Lemma 8.] According to Lemma 12, we have

$$X_{E^{(k)}}\beta^*_{E^{(k)},n} - X_E\beta_{E,n} = O_p(n^{-1/2}).$$

By a Taylor approximation, we have

$$\begin{aligned}R_1^{(k)} - r_1^{(k)} = &\frac{1}{n}X^\intercal \operatorname{diag}(\nabla^2 A(X_E\beta_{E,n}))(X_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},E} - X_E)\beta_{E,n}\\ &- \frac{1}{n_k}X^{(k)\intercal}\operatorname{diag}(\nabla^2 A(X^{(k)}_E\beta_{E,n}))(X^{(k)}_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},E} - X_E)\beta_{E,n} + o_p(n^{-1/2}).\end{aligned}$$

By Assumption 1, there exists $\mathcal{E}_k \subseteq \widetilde{E}^{(k)} = E\backslash E^{(k)}$ such that for $j \in \widetilde{E}^{(k)}\backslash\mathcal{E}_k$, $X_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},j} = X_j$ and $\beta_{\mathcal{E}_k,n} = O(n^{-1/2})$. So the last display simplifies as

$$\begin{aligned}R_1^{(k)} - r_1^{(k)} = &\frac{1}{n}X^\intercal \operatorname{diag}(\nabla^2 A(X_E\beta_{E,n}))(X_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k})\beta_{\mathcal{E}_k,n}\\ &- \frac{1}{n_k}X^{(k)}\operatorname{diag}(\nabla^2 A(X^{(k)}_E\beta_{E,n}))(X^{(k)}_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k})\beta_{\mathcal{E}_k,n} + o_p(n^{-1/2}).\end{aligned}$$

If $\mathcal{E}_k$ is not empty, let

$$T_1 = \mathbb{E}\left[\frac{1}{n}X^\intercal \operatorname{diag}(\nabla^2 A(X_E\beta_{E,n}))(X_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k})\right].$$

Then

$$\begin{aligned}R_1^{(k)} - r_1^{(k)} = &\left[\frac{1}{n}X^\intercal \operatorname{diag}(\nabla^2 A(X_E\beta_{E,n}))(X_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k}) - T_1\right]\beta_{\mathcal{E}_k,n}\\ &- \left[\frac{1}{n_k}X^{(k)\intercal}\operatorname{diag}(\nabla^2 A(X^{(k)}_E\beta_{E,n}))(X^{(k)}_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k}) - T_1\right]\beta_{\mathcal{E}_k,n}\\ &+ o_p(n^{-1/2}).\end{aligned}$$

Note that $\beta_{\mathcal{E}_k} = O(n^{-1/2})$. Further, observe that

$$\frac{1}{n} X^\intercal \operatorname{diag}(\nabla^2 A(X_E \beta_{E,n})(X_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k}) - T_1 = o_p(1), \text{ and}$$

$$\frac{1}{n_k} X^{(k)\intercal} \operatorname{diag}(\nabla^2 A(X^{(k)}_E \beta_{E,n})(X^{(k)}_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k}) - T_1 = o_p(1).$$

Thus, we conclude that $R_1^{(k)} - r_1^{(k)} = o_p(n^{-1/2})$. ■

Lemma 9 (Rate of $R_2^{(k)} - r_2^{(k)}$) *Let*

$$R_2^{(k)} = \frac{1}{n} X^\intercal (\nabla A(X \widehat{\beta}^{\Lambda,(k)}) - \nabla A(X_{E^{(k)}} \beta^*_{E^{(k)},n})),$$
$$r_2^{(k)} = \frac{1}{n_k} X^{(k)\intercal} (\nabla A(X^{(k)} \widehat{\beta}^{\Lambda,(k)}) - \nabla A(X^{(k)}_{E^{(k)}} \beta^*_{E^{(k)},n})).$$

Then $\sqrt{n}(R_2^{(k)} - r_2^{(k)}) = o_p(1)$.

Proof [Proof of Lemma 9.] Based on the assertion in Lemma 12, we have

$$\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n} = O_p(n^{-1/2}).$$

Taking a Taylor expansion of $\nabla A(X \widehat{\beta}^{\Lambda,(k)})$ at $X \beta^*_{E^{(k)},n}$ for each coordinate, we obtain

$$R_2^{(k)} - r_2^{(k)} = \frac{1}{n} X^\intercal \left[\operatorname{diag}(\nabla^2 A(X_{E^{(k)}} \beta^*_{E^{(k)},n})) X_{E^{(k)}} (\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}) + o(\|\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}\|) \right] -$$
$$\frac{1}{n_k} X^{(k)\intercal} \left[\operatorname{diag}(\nabla^2 A(X^{(k)}_{E^{(k)}} \beta^*_{E^{(k)},n})) X^{(k)}_{E^{(k)}} (\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}) + o(\|\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}\|) \right].$$

Letting

$$T = \mathbb{E}\left[\frac{1}{n} X^\intercal \operatorname{diag}(\nabla^2 A(X_{E^{(k)}} \beta^*_{E^{(k)},n})) X_{E^{(k)}} \right],$$

we have

$$\frac{1}{n} X^\intercal \operatorname{diag}(\nabla^2 A(X_{E^{(k)}} \beta^*_{E^{(k)},n})) X_{E^{(k)}} = T + o_p(1),$$

and

$$\frac{1}{n_k} X^{(k)\intercal} \operatorname{diag}(\nabla^2 A(X^{(k)}_{E^{(k)}} \beta^*_{E^{(k)},n})) X^{(k)}_{E^{(k)}} = T + o_p(1).$$

Hence,

$$\begin{aligned} R_2^{(k)} - r_2^{(k)} &= \left[\frac{1}{n} X^\intercal \operatorname{diag}(\nabla^2 A(X_{E^{(k)}} \beta^*_{E^{(k)},n})) X_{E^{(k)}} - T \right] (\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}) \\ &\quad - \left[\frac{1}{n_k} X^{(k)\intercal} \operatorname{diag}(\nabla^2 A(X^{(k)}_{E^{(k)}} \beta^*_{E^{(k)},n})) X^{(k)}_{E^{(k)}} - T \right] (\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}) + o_p(n^{-1/2}) \\ &= o_p(n^{-1/2}). \end{aligned}$$

■

■

A.2 Proof of Theorem 3

Proof It follows from Assumption 2 that

$$\sqrt{n}(\widehat{\beta}_E - \beta_{E,n}) \overset{d}{\Rightarrow} \mathcal{N}(\mathbf{0},\, \mathcal{I}_{E,E}^{-1}),$$

and from Theorem 2 we have $\sqrt{n}\Omega \overset{d}{\Rightarrow} \mathcal{N}(\mathbf{0},\, \Sigma_\Omega)$.

Note that

$$\begin{aligned}
\sqrt{n}\widehat{\beta}_{-E}^{\perp} &= \frac{1}{\sqrt{n}} X_{-E}^{\intercal}(\nabla A(X_E\beta_{E,n}) - Y) + \frac{1}{\sqrt{n}} X_{-E}^{\intercal}(\nabla A(X_E\widehat{\beta}_E) - \nabla A(X_E\beta_{E,n})) \\
&= \frac{1}{\sqrt{n}} X_{-E}^{\intercal}(\nabla A(X_E\beta_{E,n}) - Y) + \frac{1}{\sqrt{n}} X_{-E}^{\intercal} W X_E^{\intercal}(\widehat{\beta}_E - \beta_{E,n}) + o_p(1) \\
&= \frac{1}{\sqrt{n}} X_{-E}^{\intercal}(\nabla A(X_E\beta_{E,n}) - Y) - \mathcal{I}_{-E,E}\mathcal{I}_{E,E}^{-1} \frac{1}{\sqrt{n}} X_E^{\intercal}(\nabla A(X_E\beta_{E,n}) - Y) + o_p(1).
\end{aligned}$$

In particular, we have

$$\mathrm{Var}\begin{pmatrix} \frac{1}{\sqrt{n}} X_E^{\intercal}(\nabla A(X_E\beta_{E,n}) - Y) \\ \frac{1}{\sqrt{n}} X_{-E}^{\intercal}(\nabla A(X_E\beta_{E,n}) - Y) \end{pmatrix} = \begin{pmatrix} \mathcal{I}_{E,E} & \mathcal{I}_{E,-E} \\ \mathcal{I}_{-E,E} & \mathcal{I}_{-E,-E} \end{pmatrix}.$$

Thus, we observe that the asymptotic variance of $\sqrt{n}\widehat{\beta}_{-E}^{\perp}$ is equal to

$$\mathcal{I}_{-E,-E} - \mathcal{I}_{-E,E}\mathcal{I}_{E,E}^{-1}\mathcal{I}_{E,-E},$$

and conclude that

$$\sqrt{n}\widehat{\beta}_{-E}^{\perp} \overset{d}{\Rightarrow} \mathcal{N}(\mathbf{0}, \mathcal{I}/\mathcal{I}_{E,E}).$$

Further, it is easy see from the asymptotic representations of $\sqrt{n}(\widehat{\beta}_E - \beta_{E,n})$ and $\sqrt{n}\widehat{\beta}_{-E}^{\perp}$ that they are mutually independent.

Now, observe that $\sqrt{n}(\widehat{\beta}_E - \beta_{E,n})$ and $\sqrt{n}\widehat{\beta}_{-E}^{\perp}$ are asymptotically equivalent to sums of i.i.d. random variables z_i, and that $\sqrt{n}\omega^{(k)}$ assumes the form in (18). Thus, we can write

$$\mathrm{Cov}\left(\sqrt{n}\omega^{(k)}, \sum_{i\in[n]} z_i\right) = \sqrt{n}\mathrm{Cov}\left(\frac{1}{n}\sum_{i\in[n]} e_i - \frac{1}{n_k}\sum_{i\in\mathcal{C}_k} e_i, \sum_{i\in[n]} z_i\right) + o_p(1),$$

where $e_i = x_i(\nabla A(x_{i,E}^{\intercal}\beta_{E,n}) - y_i)$. The covariance term on the right-hand-side is $\mathbf{0}$, which completes the proof. ■

A.3 Proof of Theorem 4

Proof We start by writing

$$\sqrt{n}\omega^{(k)} = \sqrt{n}\bar{\omega}^{(k)} + o_p(1),$$

established in Lemma 13, where

$$\sqrt{n}\bar{\omega}^{(k)} = \mathbb{T}^{(k)}(\sqrt{n}\widehat{B}^{(k)}, \widehat{Z}^{(k)}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}_{-E}^{\perp}).$$

Let $\sqrt{n}\overline{\Omega} = \mathbb{T}(\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^{\perp}_{-E})$ denote the stack of $\sqrt{n}\bar{\omega}^{(k)}$ for $k \in [K]$.

We begin with p_n, the Lebesgue density of

$$\left(\sqrt{n}(\widehat{\beta}_E - \beta_{E,n}),\ \sqrt{n}\widehat{\beta}^{\perp}_{-E},\ \sqrt{n}\overline{\Omega}\right).$$

We then apply the change of variables

$$\sqrt{n}\overline{\Omega} \to (\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}})$$

through the mapping $\sqrt{n}\overline{\Omega} = \mathbb{T}(\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^{\perp}_{-E})$. Because the mapping is linear, the density of

$$\left(\sqrt{n}(\widehat{\beta}_E - \beta_{E,n}),\ \sqrt{n}\widehat{\beta}^{\perp}_{-E},\ \sqrt{n}\widehat{\mathbf{B}},\ \widehat{\mathbf{Z}}\right)$$

is proportional to

$$p_n(\sqrt{n}(\widehat{\beta}_E - \beta_{E,n}),\ \sqrt{n}\widehat{\beta}^{\perp}_{-E},\ \mathbb{T}(\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^{\perp}_{-E})).$$

Now, the condition in Assumption 3 allows us to replace p_n by the limiting Gaussian density in Theorem 3 which gives us the corresponding asymptotic density function

$$\begin{aligned}\varphi(\sqrt{n}\widehat{\beta}_E; \sqrt{n}\beta_{E,n}, \mathcal{I}^{-1}_{E,E}) \times \varphi(\sqrt{n}\widehat{\beta}^{\perp}_{-E}; \mathbf{0}, (\mathcal{I}/\mathcal{I}_{E,E})^{-1}) \\ \times \varphi(\mathbb{T}(\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^{\perp}_{-E}); \mathbf{0}, \Sigma_\Omega).\end{aligned} \tag{20}$$

Furthermore, if we condition on $\widehat{\mathbf{Z}} = \mathbf{z}$ and ignore constants, the asymptotic likelihood can be simplified to

$$\varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_{E,n} + \kappa; \Theta) \cdot \varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_E + \tau; \Xi),$$

where $\Pi, \kappa, \Theta, \Psi, \tau, \Xi$ are defined in Theorem 4. See details of the simplification in Lemma 10 below. The selection event, when conditioned on $\widehat{\mathbf{Z}} = \mathbf{z}$, is equivalent to $\widehat{\mathbf{B}} \in \mathcal{O}$. So the conditional density of $\widehat{\beta}_E, \widehat{\mathbf{B}}$ is given by

$$\frac{\varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_{E,n} + \kappa; \Theta) \cdot \varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_{E,n} + \tau; \Xi) \cdot \mathbf{1}\left\{\widehat{\mathbf{B}} \in \mathcal{O}\right\}}{\int \varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_{E,n} + \kappa; \Theta) \cdot \varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_E + \tau; \Xi) \cdot \mathbf{1}\left\{\widehat{\mathbf{B}} \in \mathcal{O}\right\} \mathrm{d}\widehat{\beta}_E \mathrm{d}\widehat{\mathbf{B}}}.$$

Ignoring the factors that do not depend on $\beta_{E,n}$ gives the selective likelihood function.

Lemma 10 (Matrix simplification) *The joint density of* $(\sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\mathbf{B}}, \widehat{\beta}^{\perp}_{-E})$ *when conditioned on* $\widehat{\mathbf{Z}} = \mathbf{z}$ *is equal to*

$$\varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_{E,n} + \kappa; \Theta) \cdot \varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_E + \tau; \Xi) \cdot \varphi(\sqrt{n}\widehat{\beta}^{\perp}_{-E}; \mathbf{0}, (\mathcal{I}/\mathcal{I}_{E,E})^{-1}).$$

Proof [Proof of Lemma 10.] Denote

$$\mathbb{Q}_1 = \begin{pmatrix} \mathcal{I}_{\cdot,E} \\ \vdots \\ \mathcal{I}_{\cdot,E} \end{pmatrix}, \quad \mathbb{Q}_2 = \begin{pmatrix} \mathcal{I}_{\cdot,E^{(1)}} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{\cdot,E^{(2)}} & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \dots & \mathbf{0} & \mathcal{I}_{\cdot,E^{(K)}} \end{pmatrix}. \tag{21}$$

Let $r^{(k)} = \Lambda \begin{pmatrix} s^{(k)} \\ z^{(k)} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \sqrt{n}\widehat{\beta}^{\perp}_{-E} \end{pmatrix}$, and let $\mathbf{r}$ be the stack of $r^{(1)}, \ldots, r^{(K)}$. Observe, the mapping $\mathbb{T}$, given in Lemma 13, can be written as

$$\sqrt{n}\overline{\Omega} = -\mathbb{Q}_1\sqrt{n}\widehat{\beta}_E + \mathbb{Q}_2\sqrt{n}\widehat{\mathbf{B}} + \mathbf{r}. \tag{22}$$

It follows from Equation (20) that the joint density of $(\sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\mathbf{B}})$ after conditioning on $\widehat{\mathbf{Z}} = \mathbf{z}$ is proportional to

$$\begin{aligned}
&\exp\Bigg[-\frac{1}{2}(\sqrt{n}\widehat{\beta}_E - \sqrt{n}\beta_{E,n})^{\intercal}\mathcal{I}_{E,E}(\sqrt{n}\widehat{\beta}_E - \sqrt{n}\beta_{E,n}) \\
&\qquad -\frac{1}{2}(-\mathbb{Q}_1\sqrt{n}\widehat{\beta}_E + \mathbb{Q}_2\sqrt{n}\widehat{\mathbf{B}} + \mathbf{r})^{\intercal}\Sigma_{\Omega}^{-1}(-\mathbb{Q}_1\sqrt{n}\widehat{\beta}_E + \mathbb{Q}_2\sqrt{n}\widehat{\mathbf{B}} + \mathbf{r})\Bigg] \\
&\propto \exp\Bigg[-\frac{1}{2}(\sqrt{n}\widehat{\mathbf{B}})^{\intercal}\mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_2(\sqrt{n}\widehat{\mathbf{B}}) + (\sqrt{n}\widehat{\mathbf{B}})^{\intercal}\mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}(\mathbb{Q}_1\sqrt{n}\widehat{\beta}_E - \mathbf{r}) \\
&\qquad -\frac{1}{2}(\sqrt{n}\widehat{\beta}_E)^{\intercal}(\mathcal{I}_{E,E} + \mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_1)(\sqrt{n}\widehat{\beta}_E) + (\sqrt{n}\widehat{\beta}_E)^{\intercal}(\mathcal{I}_{E,E}\sqrt{n}\beta_{E,n} + \mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbf{r})\Bigg].
\end{aligned}$$

For $\Xi^{-1} = \mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_2$, $\Psi = \Xi\mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_1$, $\tau = -\Xi\mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}\mathbf{r}$, we observe that this density is proportional to

$$\begin{aligned}
\varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_E + \tau, \Xi) \cdot \exp\Bigg[&-\frac{1}{2}(\sqrt{n}\widehat{\beta}_E)^{\intercal}(\mathcal{I}_{E,E} + \mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_1 - \Psi^{\intercal}\Xi^{-1}\Psi)(\sqrt{n}\widehat{\beta}_E) \\
&+(\sqrt{n}\widehat{\beta}_E)^{\intercal}(\Psi^{\intercal}\Xi^{-1}\tau + \mathcal{I}_{E,E}\sqrt{n}\beta_{E,n} + \mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbf{r})\Bigg].
\end{aligned}$$

The density in the last display is proportional to

$$\varphi(\sqrt{n}\widehat{\beta}_E; \Pi\sqrt{n}\beta_{E,n} + \kappa; \Theta) \cdot \varphi(\sqrt{n}\widehat{\mathbf{B}}; \Psi\sqrt{n}\widehat{\beta}_E + \tau; \Xi)$$

for $\Theta^{-1} = \mathcal{I}_{E,E} + \mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_1 - \Psi^{\intercal}\Xi^{-1}\Psi$, $\Pi = \Theta\mathcal{I}_{E,E}$, $\kappa = \Theta(\Psi^{\intercal}\Xi^{-1}\tau + \mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbf{r})$.

To further simplifying the matrices in the likelihood, we note that $\Sigma_{\Omega}^{-1} = U^{-1} \otimes \mathcal{I}^{-1}$. Therefore, we can write $\mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_2$ in the form of $K \times K$ blocks, where the (j,k)-block is

$$\left(\frac{\rho_j\rho_k}{\rho_0} + \rho_j\delta_{j,k}\right)\mathcal{I}_{E^{(j)},E^{(k)}}.$$

Similarly, $\mathbb{Q}_2^{\intercal}\Sigma_{\Omega}^{-1}\mathbf{r}$ has K blocks where the k-th block is

$$\rho_k J_{E^{(k)}}\mathbf{r}^{(k)} + \frac{\rho_k}{\rho_0}\sum_{j=1}^{K}\rho_j J_{E^{(k)}}\mathbf{r}^{(j)},$$

where J_E is the matrix that selects the coordinates in E, i.e. $J_E x = x_E$. Since $J_{E^{(k)}}\begin{pmatrix} 0_E \\ \widehat{\beta}^{\perp}_{-E} \end{pmatrix} = 0$ and $J_{E^{(k)}}\gamma^{(j)} = g_k^{(j)}$, the above display is equal to $\rho_k g_k^{(k)} + (\rho_k/\rho_0)\sum_{j\in[K]}\rho_j g_k^{(j)}$. Similarly, $\mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbf{r} = \sum_{k\in[K]}\frac{\rho_k}{\rho_0}J_E\gamma^{(k)} = \sum_{k\in[K]}\frac{\rho_k}{\rho_0}g^{(k)}$ and

$$\mathbb{Q}_1^{\intercal}\Sigma_{\Omega}^{-1}\mathbb{Q}_1 = \left(\sum_{k\in[K]}\rho_k + \sum_{j,k\in[K]}\frac{\rho_j\rho_k}{\rho_0}\right)\mathcal{I}_{E,E} = \frac{1-\rho_0}{\rho_0}\mathcal{I}_{E,E}.$$

Thus $\Theta^{-1} = \frac{1}{\rho_0}\mathcal{I}_{E,E} - \Psi^\intercal \Xi^{-1}\Psi$.

■

Remark 11 *We note that when using the union aggregation rule, the quantities $\Pi, \kappa, \Theta, \Psi, \tau, \Xi$ do not depend on $\widehat{\beta}^{\perp}_{-E}$. If $E^{(k)}$ is not necessarily a subset of E, then*

$$J_{E^{(k)}} \begin{pmatrix} 0_E \\ \widehat{\beta}^{\perp}_{-E} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{E^{(k)} \cap E} \\ \widehat{\beta}^{\perp}_{E^{(k)} \setminus E}, \end{pmatrix}$$

and thus

$$J_{E^{(k)}} \mathbf{r}^{(j)} = J_{E^{(k)}} \gamma^{(j)} + \begin{pmatrix} \mathbf{0}_{E^{(k)} \cap E} \\ \widehat{\beta}^{\perp}_{E^{(k)} \setminus E} \end{pmatrix}.$$

But, we only need to re-define $g_k^{(j)}$ as

$$g_k^{(j)} = J_{E^{(k)}} \gamma^{(j)} + \begin{pmatrix} \mathbf{0}_{E^{(k)} \cap E} \\ \widehat{\beta}^{\perp}_{E^{(k)} \setminus E} \end{pmatrix},$$

and this change only affects τ. For the complete procedure when using general aggregation rules, see Appendix D.

■

A.3.1 Supporting Results

Lemma 12 *Let $\beta^*_{E^{(k)},n} = \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},E} \beta_{E,n}$. Under Assumption 1,*

$$X_{E^{(k)}} \beta^*_{E^{(k)},n} - X_E \beta_{E,n} = O_p(n^{-1/2}), \text{ and}$$
$$\widehat{\beta}^{(k),\Lambda}_{E^{(k)}} - \beta^*_{E^{(k)},n} = O_p(n^{-1/2}).$$

Proof Note that

$$\begin{aligned} X_{E^{(k)}} \beta^*_{E^{(k)}} - X_E \beta_E &= (X_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},E} - X_E)\beta_E \\ &= \left(X_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},E\setminus E^{(k)}} - X_{E \setminus E^{(k)}} \right) \beta_{E \setminus E^{(k)}} + \\ &\quad \left(X_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},E\cap E^{(k)}} - X_{E \cap E^{(k)}} \right) \beta_{E \cap E^{(k)}} \\ &= \left(X_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},E\setminus E^{(k)}} - X_{E \setminus E^{(k)}} \right) \beta_{E \setminus E^{(k)}}. \end{aligned}$$

By Assumption 1, let $\mathcal{E}_k$ be the set such that if $j \in (E \setminus E^{(k)}) \setminus \mathcal{E}_k$,

$$X_{E^{(k)}} \mathcal{I}^{-1}_{E^{(k)},E^{(k)}} \mathcal{I}_{E^{(k)},j} - X_j = 0,$$

and $\beta_{\mathcal{E}_k} = O(n^{-1/2})$. Then we have

$$X_{E^{(k)}}\beta^*_{E^{(k)},n} - X_E\beta_E = (X_{E^{(k)}}\mathcal{I}^{-1}_{E^{(k)},E^{(k)}}\mathcal{I}_{E^{(k)},\mathcal{E}_k} - X_{\mathcal{E}_k})\beta_{\mathcal{E}_k} = O_p(n^{-1/2}).$$

This proves the first claim.

By the KKT condition of lasso problem (1), we have

$$\frac{1}{n_k}{X^{(k)}_{E^{(k)}}}^\intercal(\nabla A(X^{(k)}_{E^{(k)}}\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}}) - Y^{(k)}) = -\frac{1}{\sqrt{n}}\Lambda_{E^{(k)}}s_k.$$

Denote the left-hand-side as $\nabla\ell^{(k)}(\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}})$. Taking a Taylor expansion at $\beta^*_{E^{(k)},n}$, we get

$$\begin{aligned}-\frac{1}{\sqrt{n}}\Lambda_{E^{(k)}}s_k = \nabla\ell^{(k)}(\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}}) &= \nabla\ell^{(k)}(\beta^*_{E^{(k)},n}) + \nabla^2\ell^{(k)}(\beta^*_{E^{(k)},n})(\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}) \\ &\quad + o(\|\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)}}\|_2).\end{aligned} \tag{23}$$

Note that

$$\begin{aligned}\nabla\ell^{(k)}(\beta^*_{E^{(k)}}) &= \frac{1}{n_k}{X^{(k)}_{E^{(k)}}}^\intercal(\nabla A(X^{(k)}_{E^{(k)}}\beta^*_{E^{(k)},n}) - Y^{(k)}) \\ &= \frac{1}{n_k}{X^{(k)}_{E^{(k)}}}^\intercal(\nabla A(X^{(k)}_{E^{(k)}}\beta^*_{E^{(k)},n}) - \nabla A(X^{(k)}_E\beta_{E,n}))+ \\ &\qquad \frac{1}{n_k}{X^{(k)}_{E^{(k)}}}^\intercal(\nabla A(X^{(k)}_E\beta_{E,n}) - Y^{(k)}).\end{aligned}$$

The first term is of order $O_p(n^{-1/2})$ due to the first claim. The second term is of order $O_p(n^{-1/2})$ because it is an average of n_k i.i.d. random variables of mean zero and finite variance. This proves

$$\nabla\ell^{(k)}(\beta^*_{E^{(k)}}) = O_p(n^{-1/2}).$$

Moreover,

$$\begin{aligned}\nabla^2\ell^{(k)}(\beta^*_{E^{(k)}}) &= \frac{1}{n_k}{X^{(k)}_{E^{(k)}}}^\intercal \operatorname{diag}(\nabla^2 A(X_{E^{(k)}}\beta^*_{E^{(k)}}))X^{(k)}_{E^{(k)}} \\ &= \frac{1}{n_k}{X^{(k)}_{E^{(k)}}}^\intercal \operatorname{diag}(\nabla^2 A(X_E\beta_E))X^{(k)}_{E^{(k)}} + O_p(n^{-1/2}) \\ &= \mathcal{I}_{E^{(k)},E^{(k)}} + O_p(n^{-1/2}).\end{aligned}$$

Substituting into Equation (23) gives

$$\sqrt{n}(\widehat{\beta}^{\Lambda,(k)}_{E^{(k)}} - \beta^*_{E^{(k)},n}) = O_p(1).$$

■

Lemma 13 *The randomization variables have the expression*

$$\sqrt{n}\omega^{(k)} = \sqrt{n}\bar{\omega}^{(k)} + o_p(1),$$

where

$$\begin{aligned}\sqrt{n}\bar{\omega}^{(k)} &= \mathbb{T}^{(k)}(\sqrt{n}B^{(k)}, Z^{(k)}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^{\perp}_{-E}) \\ &= \mathcal{I}_{\cdot,E^{(k)}}\sqrt{n}B^{(k)} - \mathcal{I}_{\cdot,E}\sqrt{n}\widehat{\beta}_E + \Lambda\begin{pmatrix} S^{(k)} \\ Z^{(k)}\end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \sqrt{n}\widehat{\beta}^{\perp}_{-E}\end{pmatrix}.\end{aligned}$$

Proof The K.K.T. conditions of stationarity for the lasso on machine k are summarized by

$$\begin{aligned}\sqrt{n}\omega^{(k)} &= \frac{1}{\sqrt{n}}X^{\intercal}(\nabla A(X\widehat{\beta}^{(k),\Lambda}) - Y) + \gamma^{(k)} \\ &= \frac{1}{\sqrt{n}}X^{\intercal}(\nabla A(X_E\widehat{\beta}_E) - Y + \nabla A(X_{E^{(k)}}B^{(k)}) - \nabla A(X_E\widehat{\beta}_E)) + \gamma^{(k)} \\ &= \begin{pmatrix} \mathbf{0} \\ \sqrt{n}\widehat{\beta}^{\perp}_{-E}\end{pmatrix} + \frac{1}{\sqrt{n}}X^{\intercal}\operatorname{diag}(\nabla^2 A(X_E\beta_{E,n}))X_{E^{(k)}}B^{(k)} \\ &\qquad - \frac{1}{\sqrt{n}}X^{\intercal}\operatorname{diag}(\nabla^2 A(X_E\beta_{E,n}))X_E\widehat{\beta}_E + \gamma^{(k)} + o_p(1) \\ &= \mathcal{I}_{\cdot,E^{(k)}}\sqrt{n}B^{(k)} - \mathcal{I}_{\cdot,E}\sqrt{n}\widehat{\beta}_E + \Lambda\begin{pmatrix} S^{(k)} \\ Z^{(k)}\end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \sqrt{n}\widehat{\beta}^{\perp}_{-E}\end{pmatrix} + o_p(1) \\ &= \sqrt{n}\bar{w}^{(k)} + o_p(1).\end{aligned} \tag{24}$$

■

Appendix B. Proofs for Section 4

B.1 Proof of Theorem 5

Before proving Theorem 5, we provide a supporting result in Lemma 14.

First, let $\underline{s}^{(k)} = \begin{pmatrix} \gamma^{(k)}_{E^{(k)}} \\ \mathbf{0}\end{pmatrix} \in \mathbb{R}^p$ be the active components of the subgradient vector $\gamma^{(k)}$ padded with a vector of all zeros. Now, let $\underline{\mathbf{s}} \in \mathbb{R}^{pK}$ be formed by stacking the vectors $\underline{s}^{(k)}$ for $k \in [K]$. Recall that $\overline{\Omega}$ is formed by stacking the vectors $\bar{\omega}^{(k)}$, $k \in [K]$ that we previously defined in Lemma 13. Suppose that $\sqrt{n}\tilde{\Omega} = \sqrt{n}\overline{\Omega} - \underline{\mathbf{s}}$.

Lemma 14 *Define*

$$V_n = \begin{pmatrix} \sqrt{n}\widehat{\beta}_E \\ \sqrt{n}\widehat{\beta}^{\perp}_{-E} \\ \sqrt{n}\tilde{\Omega}\end{pmatrix}.$$

Let $\mathcal{S}$ be a convex subset of $\mathbb{R}^{p(K+1)}$. Under the conditions in Assumption 4 and Assumption 5, we have

$$\lim_{n\to\infty} -\frac{1}{a_n^2}\log\mathbb{P}\left[\frac{1}{a_n}V_n \in \mathcal{S}\right] = \inf_{b,b^{\perp},\omega\in\mathcal{S}} R(b, b^{\perp}, \omega),$$

where

$$R(b, b^{\perp}, \omega) = \left\{ \frac{1}{2}(b - \beta_E^*)^{\intercal} \mathcal{I}_{E,E} (b - \beta_E^*) + \frac{1}{2}(b^{\perp})^{\intercal} (\mathcal{I}/\mathcal{I}_{E,E})^{-1} b^{\perp} + \frac{1}{2} \omega^{\intercal} \Sigma_{\Omega}^{-1} \ \omega \right\}. \tag{25}$$

Proof Because the difference between $\sqrt{n}\bar{\Omega}$ and $\sqrt{n}\tilde{\Omega}$ is $O(1)$, based on the condition in Assumption 5, we have

$$\lim_{n \to \infty} \frac{1}{a_n^2} \left\{ \log \mathbb{P} \left[\frac{1}{a_n} V_n \in \mathcal{S} \right] - \log \mathbb{P} \left[\frac{1}{a_n} \begin{pmatrix} \sqrt{n}\widehat{\beta}_E \\ \sqrt{n}\widehat{\beta}_{-E}^{\perp} \\ \sqrt{n}\Omega \end{pmatrix} \in \mathcal{S} \right] \right\} = 0. \tag{26}$$

From the proof of Proposition 2, we have the representation

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_E \\ \widehat{\beta}_{-E}^{\perp} \\ \Omega \end{pmatrix} = \sqrt{n}\bar{E}_n + R_n,$$

where the limiting density of $\sqrt{n}\bar{E}_n$, at $(b, b^{\perp}, \omega)$, is proportional to $\varphi(b; \beta_E^*, \mathcal{I}_{E,E}^{-1}) \cdot \varphi(b^{\perp}; 0, \mathcal{I}/\mathcal{I}_{E,E}) \cdot \varphi(\omega; 0, \Sigma_{\Omega})$. So we have the following large-deviation limit

$$\lim_{n \to \infty} -\frac{1}{a_n^2} \log \mathbb{P} \left[\frac{1}{a_n} \begin{pmatrix} \sqrt{n}\widehat{\beta}_E \\ \sqrt{n}\widehat{\beta}_{-E}^{\perp} \\ \sqrt{n}\Omega \end{pmatrix} \in \mathcal{S} \right] = \inf_{b, b^{\perp}, \omega \in \mathcal{S}} R(b, b^{\perp}, \omega)$$

under Assumption 4, where R is the rate function (25).

The assertion follows directly by using (26). ■

Now we are ready to prove Theorem 5.

Proof First let us define some notations which we need for the proof. Recall that $r^{(k)} = \Lambda \begin{pmatrix} s^{(k)} \\ z^{(k)} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \sqrt{n}\widehat{\beta}_{-E}^{\perp} \end{pmatrix}$, and that $\mathbf{r}$ is the stack of $r^{(k)}$ for $k \in [K]$, and that $\underline{\mathbf{s}} \in \mathbb{R}^{pK}$ is the stack of $\underline{s}^{(k)} = \begin{pmatrix} \gamma_{E^{(k)}}^{(k)} \\ \mathbf{0} \end{pmatrix}$. In matrix form, we can write

$$\mathbf{r} = \mathbb{Q}_3 \widehat{\mathbf{Z}} + \mathbb{Q}_4 \sqrt{n}\widehat{\beta}_{-E}^{\perp} + \underline{s}$$

for fixed matrices $\mathbb{Q}_3$ and $\mathbb{Q}_4$. Let $\frac{1}{a_n}\mathbf{z} = \zeta$.

Let $\mathbb{Q}_1$, $\mathbb{Q}_2$ be defined according to (21). Then, note that we have

$$\begin{aligned} \sqrt{n}\tilde{\Omega} &= \sqrt{n}\bar{\Omega} - \underline{s} \\ &= -\mathbb{Q}_1 \sqrt{n}\widehat{\beta}_E + \mathbb{Q}_2 \sqrt{n}\widehat{\mathbf{B}} + \mathbf{r} - \underline{s} \\ &= -\mathbb{Q}_1 \sqrt{n}\widehat{\beta}_E + \mathbb{Q}_2 \sqrt{n}\widehat{\mathbf{B}} + \mathbb{Q}_3 \widehat{\mathbf{Z}} + \mathbb{Q}_4 \sqrt{n}\widehat{\beta}_{-E}^{\perp}. \end{aligned}$$

If we define:

$$h(B, Z; b, b^\perp) = \mathbb{Q}_2 B + \mathbb{Q}_3 Z - \mathbb{Q}_1 b + \mathbb{Q}_4 b^\perp.$$

then

$$\sqrt{n}\tilde{\Omega} = h(\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^\perp_{-E}).$$

Denote by U_n the vector

$$\begin{pmatrix} \sqrt{n}\widehat{\beta}_E \\ \sqrt{n}\widehat{\beta}^\perp_{-E} \\ \sqrt{n}\widehat{\mathbf{B}} \\ \widehat{\mathbf{Z}} \end{pmatrix},$$

and consider V_n as defined in Lemma 14. Observe that

$$V_n = \begin{pmatrix} \sqrt{n}\widehat{\beta}_E \\ \sqrt{n}\widehat{\beta}^\perp_{-E} \\ h\left(\sqrt{n}\widehat{\mathbf{B}}, \widehat{\mathbf{Z}}; \sqrt{n}\widehat{\beta}_E, \sqrt{n}\widehat{\beta}^\perp_{-E}\right) \end{pmatrix} \tag{27}$$

is a bijective linear function of U_n. Let us define $\mathbb{H} : \mathbb{R}^{p(K+1)} \to \mathbb{R}^{p(K+1)}$ as the function such that

$$\frac{1}{a_n} V_n = \mathbb{H}(\frac{1}{a_n} U_n).$$

Applying the contraction principle for large-deviation limit together with Lemma 14, we conclude that the vector $\frac{1}{a_n} U_n = \mathbb{H}^{-1}(\frac{1}{a_n} V_n)$ satisfies a large deviation principle with the rate function $R \circ \mathbb{H}$, where R is the rate function of $\frac{1}{a_n} V_n$ given in Equation (25). Thus, it follows that

$$\begin{aligned}
\lim_{n\to\infty} -\frac{1}{a_n^2} \log \mathbb{P}\left[\frac{\sqrt{n}}{a_n}\widehat{\mathbf{B}} \in \mathcal{O} \;\middle|\; \frac{1}{a_n}\widehat{\mathbf{Z}} = \zeta\right] &= \inf_{b, b^\perp, B \in \mathcal{O}} R \circ \mathbb{H}(b, b^\perp, B, \zeta) \\
&= \inf_{b, b^\perp, B \in \mathcal{O}} R(b, b^\perp, h(B, \zeta; b, b^\perp)) \\
&= \inf_{b, b^\perp, B \in \mathcal{O}} \frac{1}{2}(b - \beta^*_E)^\intercal \mathcal{I}_{E,E}(b - \beta^*_E) + \frac{1}{2} b^\perp (\mathcal{I}/\mathcal{I}_{E,E})^{-1} b^\perp \\
&\qquad + \frac{1}{2} h(B, \zeta; b, b^\perp)^\intercal \Sigma_\Omega^{-1} h(B, \zeta; b, b^\perp).
\end{aligned}$$

By a similar matrix simplification as lemma 10, this is equal to (up to constant independent of β^*_E)

$$\inf_{b, B \in \mathcal{O}} \frac{1}{2}(b - \Pi\beta^*_E - \bar{\kappa})^\intercal \Theta^{-1}(b - \Pi\beta^*_E - \bar{\kappa}) + \frac{1}{2}(B - \Psi b - \bar{\tau})^\intercal \Xi^{-1}(B - \Psi b - \bar{\tau}),$$

where $\bar{\tau} = -\Xi \mathbb{Q}_2^\intercal \Sigma_\Omega^{-1}(\mathbb{Q}_3\zeta + \mathbb{Q}_4 b^\perp)$, $\bar{\kappa} = \Theta(\Psi^\intercal \Xi^{-1} \bar{\tau} + \mathbb{Q}_1^\intercal \Sigma_\Omega^{-1}(\mathbb{Q}_3\zeta + \mathbb{Q}_4 b^\perp))$. The matrices Π, Θ, Ψ, Ξ are the same as in Lemma 10.

It remains to show that

$$\begin{aligned}
&\inf_{b, B \in \mathcal{O}} \frac{1}{2}(b - \Pi\beta^*_E - \bar{\kappa})^\intercal \Theta^{-1}(b - \Pi\beta^*_E - \bar{\kappa}) + \frac{1}{2}(B - \Psi b - \bar{\tau})^\intercal \Xi^{-1}(B - \Psi b - \bar{\tau}) \\
&= \lim_{n\to\infty} \inf_{b, B \in \mathcal{O}} \frac{1}{2}(b - \Pi\beta^*_E - \frac{1}{a_n}\kappa)^\intercal \Theta^{-1}(b - \Pi\beta^*_E - \bar{\kappa}) + \frac{1}{2}(B - \Psi b - \bar{\tau})^\intercal \Xi^{-1}(B - \Psi b - \frac{1}{a_n}\tau)
\end{aligned}$$

Recall from that Lemma 10 that $\mathbb{Q}_2^\intercal \Sigma_\Omega^{-1}$ has (j,k) block equal to $(\frac{\rho_j \rho_k}{\rho_0} + \rho_k \delta_{j,k}) J_{E^j}$, and $\mathbb{Q}_1^\intercal \Sigma_\Omega^{-1}$ has $(1,k)$ block equal to $\frac{\rho_k}{\rho_0} J_E$. The matrix $\mathbb{Q}_4$ has block k equal to $J_{-E}^\intercal$. Thus, $\mathbb{Q}_2^\intercal \Sigma_\Omega^{-1} \mathbb{Q}_4 = 0$, $\mathbb{Q}_1^\intercal \Sigma_\Omega^{-1} \mathbb{Q}_4 = 0$. So $\bar{\tau} = -\Xi \mathbb{Q}_2^\intercal \Sigma_\Omega^{-1} \mathbb{Q}_3 \zeta$. Recall that $\mathbf{z} = a_n \zeta$, so

$$\tau = -\Xi \mathbb{Q}_2^\intercal \Sigma_\Omega^{-1} (\mathbb{Q}_3 a_n \mathbf{z} + \underline{s}) = a_n \bar{\tau} - \Xi \mathbb{Q}_2^\intercal \Sigma_\Omega^{-1} \underline{s}.$$

Because $a_n \to \infty$, $\bar{\tau} = \frac{1}{a_n}\tau + o(1)$. Similarly, $\bar{\kappa} = \frac{1}{a_n}\kappa + o(1)$. This proves the last claim.

To conclude the proof, we observe that

$$\begin{aligned}
&\lim_{n\to\infty} \inf_{b,B} \left\{ \frac{1}{2} \left(b - \Pi \beta_E^* - \frac{1}{a_n}\kappa \right)^\intercal \Theta^{-1} \left(b - \Pi \beta_E^* - \frac{1}{a_n}\kappa \right) \right. \\
&\qquad \left. + \frac{1}{2} \left(B - \Psi b - \frac{1}{a_n}\tau \right)^\intercal \Xi^{-1} \left(B - \Psi b - \frac{1}{a_n}\tau \right) + \frac{1}{a_n^2} \mathrm{Barr}_{\mathcal{O}}(a_n B) \right\} \\
&= \lim_{n\to\infty} \inf_{b,B \in \mathcal{O}} \frac{1}{2} (b - \Pi \beta_E^* - \frac{1}{a_n}\kappa)^\intercal \Theta^{-1} (b - \Pi \beta_E^* - \frac{1}{a_n}\kappa) + \frac{1}{2} (B - \Psi b - \frac{1}{a_n}\tau)^\intercal \Xi^{-1} (B - \Psi b - \frac{1}{a_n}\tau).
\end{aligned}$$

This is because the sequence of convex objectives in the left-hand side display converge (in a pointwise sense) to the convex objective on the right-hand side display which has a unique minimum. ■

B.2 Proof of Theorem 6

Proof Observe, the approximate selective likelihood is equal to

$$(\sqrt{n}\widehat{\beta}_E)^\intercal \Theta^{-1} (\sqrt{n}\Pi \beta_{E,n} + \kappa) - Q_n^* \left(\Theta^{-1} (\sqrt{n}\Pi \beta_{E,n} + \kappa) \right),$$

where

$$Q_n^*(\alpha) = \sup_v (\sqrt{n}v)^\intercal \alpha - Q_n(\sqrt{n}v) \tag{28}$$

and

$$\begin{aligned}
Q_n(\sqrt{n}v) = \frac{1}{2} (\sqrt{n}v)^\intercal \Theta^{-1} \sqrt{n}v + \inf_V \left\{ \frac{1}{2} \left(\sqrt{n}V - \sqrt{n}\Psi v - \tau \right)^\intercal \Xi^{-1} \left(\sqrt{n}V - \sqrt{n}\Psi v - \tau \right) \right. \\
\left. + \mathrm{Barr}_{\mathcal{O}}(\sqrt{n}V) \right\}.
\end{aligned}$$

The score, based on the approximate selective likelihood, is equal to

$$\sqrt{n}\Pi^\intercal \Theta^{-1} \left(\sqrt{n}\widehat{\beta}_E - \nabla Q_n^* \left(\Theta^{-1} (\sqrt{n}\Pi \beta_{E,n} + \kappa) \right) \right).$$

Thus, the selective MLE is given by

$$\begin{aligned}
\Theta^{-1} (\sqrt{n}\Pi \widehat{\beta}_{E,n}^{(S)} + \kappa) &= (\nabla Q_n^*)^{-1} (\sqrt{n}\widehat{\beta}_E) \\
&= \nabla Q_n (\sqrt{n}\widehat{\beta}_E) \\
&= \Theta^{-1} \sqrt{n}\widehat{\beta}_E - \Psi^\intercal \Xi^{-1} \left(\sqrt{n}\widehat{V}^\star_{\widehat{\beta}_E} - \sqrt{n}\Psi \widehat{\beta}_E - \tau \right).
\end{aligned}$$

That is,

$$\begin{aligned}\sqrt{n}\widehat{\beta}_E^{(S)} &= \sqrt{n}\Pi^{-1}\widehat{\beta}_E - \Pi^{-1}\kappa + \Pi^{-1}\Theta\Psi^\intercal\Xi^{-1}\left(\Psi\sqrt{n}\widehat{\beta}_E + \tau - \sqrt{n}\widehat{V}^\star_{\widehat{\beta}_E}\right)\\ &= \sqrt{n}\Pi^{-1}\widehat{\beta}_E - \Pi^{-1}\kappa + \mathcal{I}_{E,E}^{-1}\Psi^\intercal\Theta^{-1}\left(\Psi\sqrt{n}\widehat{\beta}_E + \tau - \sqrt{n}\widehat{V}^\star_{\widehat{\beta}_E}\right).\end{aligned}$$

Let v^* be the solution of (28) when $\alpha = \sqrt{n}\Pi\widehat{\beta}_E^{(S)} + \kappa$. The selective obs-FI matrix, derived from the curvature of the approximate selective likelihood, is given by

$$\begin{aligned}\widehat{\mathcal{I}}_{E,E}^{(S)} &= \Pi^\intercal\Theta^{-1}\nabla^2 Q_n^*\left(\sqrt{n}\Pi\widehat{\beta}_E^{(S)} + \kappa\right)\Theta^{-1}\Pi\\ &= \Pi^\intercal\Theta^{-1}\left(\nabla^2 Q_n\left(\sqrt{n}v^*\right)\right)^{-1}\Theta^{-1}\Pi\\ &= \Pi^\intercal\Theta^{-1}\left(\Theta^{-1} + \Psi^\intercal\Xi^{-1}\Psi - \Psi^\intercal\Xi^{-1}\left(\Xi^{-1} + \nabla^2\mathrm{Barr}\left(\sqrt{n}\widehat{V}^\star_{\widehat{\beta}_E}\right)\right)^{-1}\Xi^{-1}\Psi\right)^{-1}\Theta^{-1}\Pi\\ &= \mathcal{I}_{E,E}\left(\Theta^{-1} + \Psi^\intercal\Xi^{-1}\Psi - \Psi^\intercal\Xi^{-1}\left(\Xi^{-1} + \nabla^2\mathrm{Barr}\left(\sqrt{n}\widehat{V}^\star_{\widehat{\beta}_E}\right)\right)^{-1}\Xi^{-1}\Psi\right)^{-1}\mathcal{I}_{E,E}.\end{aligned}$$

■

Appendix C. Sampling subsets with replacement

Proof [Proof of Lemma 7] It suffices to prove that for $j \neq k$ $\mathrm{Cov}\left[\sqrt{n}\omega^{(j)}, \sqrt{n}\omega^{(k)}\right] \to 0$. Following the proof A.1, we only need to show

$$\mathrm{Cov}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(j)}} e_i,\ \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(k)}} e_i\right] \to 0.$$

Let $\mathcal{C}^{(j)}$ denote the index set of $D^{(j)}$. Since $\mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(j)}} e_i \mid \mathcal{C}^{(j)}\right] = 0$, it remains to show that

$$\mathbb{E}\left[\mathrm{Cov}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(j)}} e_i,\ \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(k)}} e_i \mid \mathcal{C}^{(j)}, \mathcal{C}^{(k)}\right]\right] = 0.$$

Note that

$$\begin{aligned}&\mathrm{Cov}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(j)}} e_i,\ \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i - \frac{1}{\sqrt{n}\rho}\sum_{i\in\mathcal{C}^{(k)}} e_i \mid \mathcal{C}^{(j)}, \mathcal{C}^{(k)}\right]\\ &= \mathrm{Cov}\left[e_i\right] - \mathrm{Cov}\left[e_i\right] - \mathrm{Cov}\left[e_i\right] + |\mathcal{C}^{(j)} \cap \mathcal{C}^{(k)}|\frac{1}{n\rho^2}\mathrm{Cov}\left[e_i\right]\\ &= (|\mathcal{C}^{(j)} \cap \mathcal{C}^{(k)}|\frac{1}{n\rho^2} - 1)\,\mathrm{Cov}\left[e_i\right].\end{aligned}$$

The proof is completed by the fact that $\mathbb{E}\left[|\mathcal{C}^{(j)} \cap \mathcal{C}^{(k)}|\right] = n\rho^2$. ■

In this setting, the matrices $\Xi, \Theta, \Psi, \tau, \nu, \Pi, \kappa$ are similarly found by Theorem 4 and its proof with $\Sigma_\Omega = \frac{1-\rho}{\rho} I_K \otimes \mathcal{I}$. So now $\Sigma_\Omega^{-1} = \frac{\rho}{1-\rho} I_K \otimes \mathcal{I}^{-1}$. Then

$$\{\Xi^{-1}\}_{j,k} = \{\mathbb{Q}_2 \Sigma_\Omega^{-1} \mathbb{Q}_2\}_{j,k} = \frac{\rho}{1-\rho} \mathcal{I}_{E^{(j)}, E^{(j)}}$$

if $j = k$ and $\mathbf{0}$ otherwise. Other matrices are similarly computed.

Appendix D. Selective inference with general aggregation rules

D.1 Algorithm

In the main manuscript, we focused on the union aggregation rule, i.e., the final model E is the union of selected variables in the base models $E^{(k)}, 1 \leq k \leq K$. We show that with a slight modification, our procedure can be adapted to accommodate other aggregation rules.

The new procedure is summarized in Algorithm 4. We note that the procedure remains almost the same, except that

$$g_k^{(j)} = J_{E^{(k)}} \gamma^{(j)} + \begin{pmatrix} \mathbf{0}_{E^{(k)} \cap E} \\ \widehat{\beta}^{\perp}_{E^{(k)} \setminus E} \end{pmatrix}, \tag{29}$$

where

$$\widehat{\beta}^{\perp} = \frac{1}{n} X^{\intercal} (Y - X \widehat{\beta}_E).$$

If there exists a variable that is selected by machine k but is not selected in the final model E, then we must compensate the subgradients by the correlation between that variable and the residual vector for a more general aggregation rule.

To compute the vector $g_k^{(j)}$, we note that the central machine requires $\widehat{\beta}^{\perp}_{E^u \setminus E}$, where $E^u = \cup_{k \in [K]} E^{(k)}$ is the union of the base models. Thus, each local machine must send

$$\widehat{\beta}^{\perp,(k)}_{E^u \setminus E} = X^{(k),\intercal}_{E^u \setminus E} (Y^{(k)} - X^{(k)} \widehat{\beta}_E)$$

to the central machine. Because this quantity depends on the MLE $\widehat{\beta}_E$, which is computed on the central machine, the central machine must first send $\widehat{\beta}_E$ to the local machines. Our modified procedure in Algorithm 4, therefore, involves two more exchanges between the central machine and local machines: (1) the central machine sends $\widehat{\beta}_E$ to local machines; (2) local machines send $\widehat{\beta}^{\perp,(k)}_{E^u \setminus E}$ to the central machine. In comparison with Algorithm 1, the communication cost is $|E^u \cup E|$ per local machine. Note that this cost is comparable to the overall cost of order $O(|E|^2)$, as long as $|E^u|$ is about the same order as $|E|$.

Of course, the modified $g_k^{(j)}$ in (29) change some matrices in the optimization that we solve for approximately-valid selective inference. Theoretically, our selective likelihood is now obtained by conditioning further on $\widehat{\beta}^{\perp}_{E^u \setminus E}$ besides the information from the subgradient vectors.

Algorithm 4: General aggregation rules.

STEP 1: Variable Selection at Local Machines
Machine k solves (1) and sends $E^{(k)} = \text{Support}(\widehat{\beta}^{\Lambda,(k)})$ to the central machine.

STEP 2: Modeling with selected predictors
Central Machine aggregates $E^{(k)}$ to get the final model E and forms the aggregated model E

STEP 3: Communication with Central Machine
Exchange 1: Central machine sends the set E as well as $E^u = \cup_{k\in[K]} E^{(k)}$ to the local machines.
Exchange 2: Local machines send back the following information

$$\text{local estimators: } \widehat{\beta}_E^{(k)},\ \widehat{\mathcal{I}}_{E,E}^{(k)}; \qquad \text{subgradient at } \widehat{\beta}^{\Lambda,(k)}\text{: } \gamma_{E^u}^{(k)}.$$

Exchange 3: Central Machine computes the MLE $\widehat{\beta}_E$ and sends to local machines.
Exchange 4: Local machines compute $\widehat{\beta}^{\perp,(k)}_{E^u\setminus E} = X^{(k),\intercal}_{E^u\setminus E}(Y^{(k)} - X^{(k)}\widehat{\beta}_E)$ and send to Central Machine.

STEP 4: Selective Inference at Central Machine

(A) Compute $\widehat{\beta}^{\perp}_{E^u\setminus E} = \frac{1}{n}X^{(0),\intercal}_{E^u\setminus E}(Y^{(0)} - X^{(0)}\widehat{\beta}_E) + \frac{1}{n}\sum_{k=1}^{K}\widehat{\beta}^{\perp,(k)}_{E^u\setminus E}$

(B) Compute $\widehat{\Xi}, \widehat{\Psi}, \widehat{\tau}, \widehat{\Theta}, \widehat{\Pi}, \widehat{\kappa}$ as defined in Theorem 4, with $g_j^{(k)}$ defined according to Equation (29).

(C) Apply Algorithm 2 to perform inference based on selective MLE.

D.2 Experiments

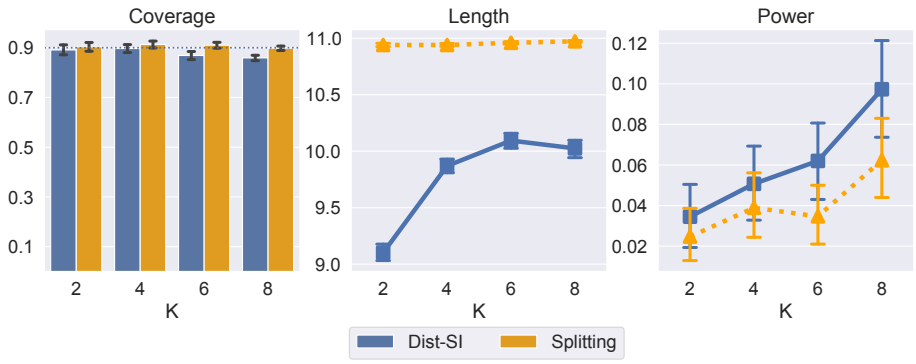
We illustrate the performance of Algorithm 4 on simulated data. In the following experiment, we consider the same setting as that in Section 6 with prespecified groups of correlated predictors. More specifically, we consider 20 groups of predictors with size 5; distinct groups of predictors are uncorrelated, while all pairs of predictors within the same group have correlation equal to 0.9. As before, we assume there are 5 non-zero coefficients β_j and these nonzero coefficients are present in 5 different groups.

Suppose that

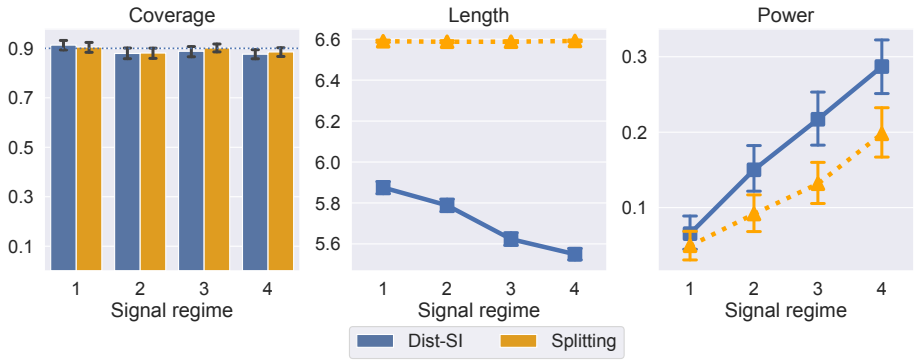
$$G = \bigcup_{k \in [K]} \{G_j : j \in E^{(k)}\},$$

i.e., G contains groups which have at least one predictor selected by at least one of the K local machines. Our final aggregated model is formed by randomly picking one predictor from each of the selected groups (in G) with highly correlated predictors.

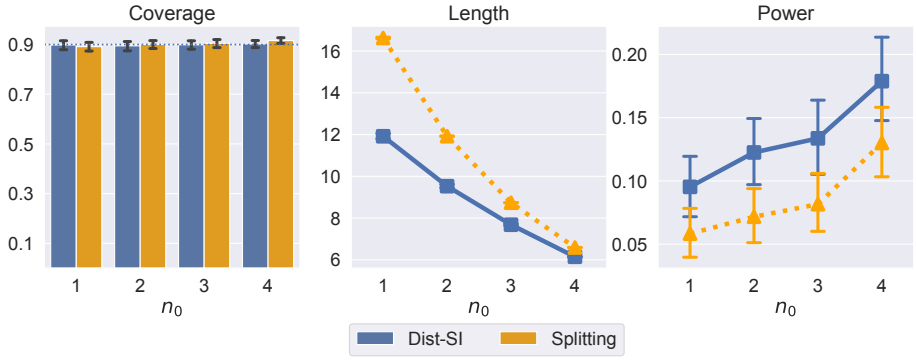
The results of our experiment are shown in Figure 7. We see that our proposed method achieves the desired coverage probability. Similar patterns hold up for the lengths and power of our confidence intervals as was already noted for the previous aggregation rule.



(a) Varying K. Each local machine has $[8000/K]$ data points for each K.



(b) Varying signal strength. The nonzero β_j equals $\pm\sqrt{2c\log p}$ with random signs for $c = 0.3, 0.5, 0.7, 0.9$ in the four signal regimes.



(c) Varying n_0, the sample size in the central machine.

Figure 7: Results for the grouped aggregation rule

References

François Bachoc, Hannes Leeb, and Benedikt Pötscher. Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics*, 47(3):1475–1504, 2019.

Maria Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *Conference on Learning Theory*, pages 26–1. JMLR Workshop and Conference Proceedings, 2012.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102 (1):77–94, 2015.

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100 (469):71–81, 2005.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Ali Charkhi and Gerda Claeskens. Asymptotic post-selection inference for the Akaike information criterion. *Biometrika*, 105(3):645–664, 2018.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684, 2014.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Zaijing Huang and Andrew Gelman. Sampling for Bayesian computation with large datasets. *Available at SSRN 1010107*, 2005.

Sangwon Hyun, Max G'Sell, and Ryan Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2019.

Vo Nguyen Le Duy and Ichiro Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *Journal of Machine Learning Research*, 23 (300):1–37, 2022.

Jason Lee, Yuekai Sun, Qiang Liu, and Jonathan Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.

Jason Lee, Dennis L Sun, Yuekai Sun, and Jonathan Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Nan Lin and Ruibin Xi. Aggregated estimating equation estimation. *Statistics and Its Interface*, 4(1):73–83, 2011.

Sifan Liu. An exact sampler for inference after polyhedral model selection. *arXiv preprint arXiv:2308.10346*, 2023.

Sifan Liu, Jelena Markovic, and Jonathan Taylor. Black-box selective inference via bootstrapping. *arXiv preprint arXiv:2203.14504*, 2022.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon Mann. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in neural information processing systems*, 22, 2009.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18(1):4488–4527, 2017.

Willie Neiswanger, Chong Wang, and Eric P Xing. Asymptotically exact, embarrassingly parallel mcmc. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 623–632, 2014.

Snigdha Panigrahi. Carving model-free inference. *The Annals of Statistics*, 51(6):2318–2341, 2023.

Snigdha Panigrahi and Jonathan Taylor. Scalable methods for Bayesian selective inference. *Electronic Journal of Statistics*, 12(2):2355–2400, 2018.

Snigdha Panigrahi and Jonathan Taylor. Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association*, 118(544):2810–2820, 2023.

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

Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Integrative methods for post-selection inference under convex constraints. *The Annals of Statistics*, 49(5):2803–2824, 2021.

Snigdha Panigrahi, Jingshen Wang, and Xuming He. Treatment effect estimation with efficient data aggregation. *arXiv preprint arXiv:2203.12726*, 2022.

Jesse Raffa, Alistair Johnson, Zach O'Brien, Tom Pollard, Roger Mark, Leo Celi, David Pilcher, and Omar Badawi. The global open source severity of illness score (GOSSIS). *Critical Care Medicine*, 2022.

D García Rasines and GA Young. Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614, 2023.

Jonathan Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

Christoph Schultheiss, Claude Renaux, and Peter Bühlmann. Multicarving for high-dimensional post-selection inference. *Electronic Journal of Statistics*, 15(1):1695–1742, 2021.

Steven Scott, Alexander Blocker, Fernando Bonassi, Hugh Chipman, Edward George, and Robert McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

Sanvesh Srivastava, Cheng Li, and David Dunson. Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

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

Xiangyu Wang and David Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Yuchen Zhang, Martin Wainwright, and John Duchi. Communication-efficient algorithms for statistical optimization. *Advances in neural information processing systems*, 25, 2012.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Tijana Zrnic and Michael I Jordan. Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691, 2023.