The Annals of Statistics
2018, Vol. 46, No. 2, 679–710
https://doi.org/10.1214/17-AOS1564

SELECTIVE INFERENCE WITH A RANDOMIZED RESPONSE

BY XIAOYING TIAN AND JONATHAN TAYLOR[1]

Stanford University

Inspired by sample splitting and the reusable holdout introduced in the field of differential privacy, we consider selective inference with a randomized response. We discuss two major advantages of using a randomized response for model selection. First, the selectively valid tests are more powerful after randomized selection. Second, it allows consistent estimation and weak convergence of selective inference procedures. Under independent sampling, we prove a selective (or privatized) central limit theorem that transfers procedures valid under asymptotic normality without selection to their corresponding selective counterparts. This allows selective inference in nonparametric settings. Finally, we propose a framework of inference after combining multiple randomized selection procedures. We focus on the classical asymptotic setting, leaving the interesting high-dimensional asymptotic questions for future work.

1. Introduction. Tukey (1980) promoted the use of *exploratory data analysis* to examine the data and possibly formulate hypotheses for further investigation. Nowadays, many statistical learning methods allow us to perform these exploratory data analyses, based on which we can posit a model on the data generating distribution. Since this model is not given a priori, classical statistical inference will not provide valid tests that control the Type-I errors.

Selective inference seeks to address this problem, see Fithian, Sun and Taylor (2014), Lee et al. (2016), Lockhart et al. (2014), Tibshirani et al. (2016). Loosely speaking, there are two stages in selective inference. The first is the *selection* stage that explores the data and formulates a plausible model for the data distribution. Then we enter the *inference* stage that seeks to provide valid inference under the selected model which is proposed after inspecting the data. Inference under different models have been studied, notably the Gaussian families Lee et al. (2016), Tian, Loftus and Taylor (2015), Tibshirani et al. (2016) as well as other exponential families Fithian, Sun and Taylor (2014). The target of inference in the selective inference problems can be adaptively chosen, which is different from other works on inference in modern regression settings Bühlmann (2013), Javanmard and Montanari (2014), van de Geer et al. (2014), Zhang and Zhang (2014).

Received March 2016; revised February 2017.
[1]Supported in part by NSF Grant DMS-12-08857 and AFOSR Grant 113039.
MSC2010 subject classifications. Primary 62M40; secondary 62J05.
Key words and phrases. Selective inference, nonparametric, differential privacy.

In this work, we consider selective inference in a general setting that include nonparametric settings. In addition, we introduced the use of *randomized response* in model selection. A most common example of randomized model selection is probably the practice of data splitting Cox (1975), Wasserman and Roeder (2009). Assuming independent sampling, we can divide the data into two subsets, using the first for model selection and the second subset for inference. Though not emphasized, this split is often *random*. Hence, data splitting can be thought of as a special case of randomized model selection. To motivate the use of randomized selection and introduce the inference problem that ensues, we consider the following example.

1.1. *A first example*. Publication bias, [also called the "file drawer effect" by Rosenthal (1979)] is a bias introduced to scientific literature by failure to report negative or nonconfirmatory results. While it is difficult to correct for the selection bias in published works without access to the original data and detailed selection procedures, it is possible to develop a framework for scientists to perform valid post-selection inference in the process of data analysis. We formulate the problem in the simple example below.

EXAMPLE 1 (File drawer problem). Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_{i,n}$$

be the sample mean of a sample of n i.i.d. draws from $\mathbb{F}_n$ in a standard triangular array. We set $\mu_n = \mathbb{E}_{\mathbb{F}_n}[X_{1,n}]$ and assume $\mathbb{E}_{\mathbb{F}_n}[(X_{1,n} - \mu_n)^2] = 1$.

Suppose that we are interested in discovering positive effects and would only report the sample mean if it survives the file drawer effect, that is,

$$n^{1/2}\bar{X}_n > 2. \tag{1}$$

Then what is the "correct" p-value to report for an observation $\bar{X}_{n,\mathrm{obs}}$ that exceeds the threshold?

If we have Gaussian family, namely $\mathbb{F}_n = N(\mu_n, 1)$, then the distribution of $\bar{X}_n$ surviving the file drawer effect (1) is a truncated Gaussian distribution. We also call this distribution the *selective distribution*. Formally, its survival function is

$$\begin{aligned} P(t) &= \mathbb{P}\big(\bar{X}_n > t | n^{1/2}\bar{X}_n > 2\big), \qquad \bar{X}_n \sim N\left(\mu_n, \frac{1}{n}\right) \\ &= \frac{1 - \Phi(n^{1/2}(t - \mu_n))}{1 - \Phi(2 - n^{1/2}\mu_n)}, \end{aligned}$$

where Φ is the CDF of an $N(0,1)$ random variable. Therefore, we get a pivotal quantity

$$P(\bar{X}_{n,\text{obs}}) = \frac{1-\Phi(n^{1/2}(\bar{X}_{n,\text{obs}}-\mu_n))}{1-\Phi(2-n^{1/2}\mu_n)} \sim \text{Unif}(0,1),$$
$$n^{1/2}\bar{X}_{n,\text{obs}} > 2, \qquad X_{n,\text{obs}} \sim N\left(\mu_n, \frac{1}{n}\right). \tag{2}$$

The pivotal quantity in (2) allows us to construct p-values or confidence intervals for Gaussian families. When the distributions $\mathbb{F}_n$'s are not normal distributions, central limit theorem states that the sample mean $\bar{X}_n$ is asymptotically normal when $\mathbb{F}_n$ has second moments. Thus, a natural question is whether the pivotal quantity in (2) is asymptotically Unif(0, 1) when $X_{i,n}$ does not come from a normal distribution?

The following lemma provides a negative answer to this question in the case when $\mathbb{F}_n$ is a translated Bernoulli distribution that has a negative mean. Essentially when the selection event $n^{1/2}\bar{X}_n > 2$ becomes a rare event with vanishing probability, the pivotal quantity in (2) no longer converges to Unif(0, 1). We defer the proof of the lemma to Section B in the Supplementary Material [Tian and Taylor (2018)].

LEMMA 1. *If $X_{i,n}$ takes values in $\{-1.5, 0.5\}$, with $\mathbb{P}(X_{i,n} = -1.5) = \mathbb{P}(X_{i,n} = 0.5) = 0.5$. Thus, $\mu_n = -0.5$. Then the pivot in (2) does not converge to* Unif(0, 1)

$$P(\bar{X}_n) \nrightarrow \text{Unif}(0,1),$$

for the $\bar{X}_n$'s surviving the file drawer effect (1).

Randomized selection circumvents this problem. In the following, we propose a randomized version of the "file drawer problem".

EXAMPLE 2 (File drawer problem, randomized). We assume the same setup of a triangular array of observations $X_{i,n}$ as in Example 1. But instead of reporting $\bar{X}_n$ when it survives the file drawer effect (1), we independently draw $\omega \sim G$, and only report $\bar{X}_n$ if

$$n^{1/2}\bar{X}_n + \omega > 2. \tag{3}$$

Note that the selection event is different from that in (1) in that we randomize the sample mean before checking whether it passes the threshold. In this case, if $\mathbb{F}_n = N(\mu_n, 1)$, the survival function of $\bar{X}_n$ is

$$\begin{aligned} P(t) &= \mathbb{P}\big(\bar{X}_n > t | n^{1/2}\bar{X}_n + \omega > 2\big), \qquad (\bar{X}_n, \omega) \sim N\left(\mu_n, \frac{1}{n}\right) \times G \\ &= \mathbb{P}\big(Z > n^{1/2}(t-\mu_n) | Z + \omega > 2 - n^{1/2}\mu_n\big), \qquad (Z, \omega) \sim N(0,1) \times G. \end{aligned} \tag{4}$$

To compute the exact form of $P(t)$, we have to compute the convolution of $N(0, 1)$ and G which has explicit forms for many distributions G. Moreover, when G is Logistic or Laplace distribution, we have

$$P(\bar{X}_{n,\text{obs}}) \to \text{Unif}(0, 1),$$

as long as $\mathbb{F}_n$ has centered exponential moments in a fixed neighborhood of 0. The convergence is in fact uniform for $-\infty < \mu_n < \infty$. For details, see Lemma 10 in Section 5.2.

The only difference between these two examples is the randomization in selection. After selection, we need to consider the conditional distribution for inference, which conditions on the selection event. If we denote by $\mathbb{F}_n^*$ the distribution used for selective inference, we have in Example 1

$$\frac{d\mathbb{F}_n^*}{d\mathbb{F}_n}(\bar{X}_n) = \frac{1_{\{n^{1/2}\bar{X}_n > 2\}}}{\mathbb{P}_{\mathbb{F}_n}(n^{1/2}\bar{X}_n > 2)}. \tag{5}$$

We also call the ratio between $\mathbb{F}_n^*$ and $\mathbb{F}_n$ the *selective likelihood ratio*. In this case, the selective likelihood ratio is simply a restriction to the $\bar{X}_n$'s that survives the file drawer effect. We observe that

$$\sqrt{n}\bar{X}_n = \sqrt{n}\mu_n + Z, \qquad Z \sim N(0, 1),$$

which leads to three scenarios for selection.

- $\mu_n > \delta > 0$, for some $\delta > 0$.
  In this case, the dominant term for selection is $\sqrt{n}\mu_n$, and since we have a big positive effect, we would always report the sample mean $\bar{X}_n$ when n is big. This corresponds to the selection event having probability tending to 1 and the selective likelihood ratio goes to 1 as well. In this case, there is very little selection bias, and the original law is a good approximation to the selective distribution for valid inference.
- $\mu_n < -\delta < 0$, for some $\delta > 0$.
  In this case, the dominant term is also $\sqrt{n}\mu_n$, but in the negative direction. As $n \to \infty$, the selection probability vanishes and the selective likelihood becomes degenerate. We almost never report the sample mean in this scenario, but in the rare event where we do, by no means can we use the original distribution for inference.
- $-\delta < n^{1/2}\mu_n < \delta$, for some $\delta > 0$.
  This corresponds to local alternatives. In this case, the selective likelihood neither converges to 1 or becomes degenerate. Rather, it becomes an indicator function of a half interval. Proper adjustment is needed for valid inference in this case.

It is in the second scenario that pivotal quantity (2) will not converge to Unif(0, 1). Different distributions will have different behaviors in the tail. Since the conditioning event $n^{1/2}\bar{X}_n > 2$ becomes a large-deviations event, we cannot expect it to behave like the normal distribution in the tail.

On the other hand, in Example 2, if we denote by $\tilde{\mathbb{F}}_n^*$ the law for selective inference, we have

$$\begin{aligned}\frac{d\tilde{\mathbb{F}}_n^*}{d\mathbb{F}_n}(\bar{X}_n) &= \frac{\bar{G}(2 - n^{1/2}\bar{X}_n)}{\mathbb{E}_{\mathbb{F}_n}(\bar{G}(2 - n^{1/2}\bar{X}_n))} \\ &= \frac{\bar{G}(2 - n^{1/2}(\bar{X}_n - \mu_n) - n^{1/2}\mu_n)}{\mathbb{E}_{\mathbb{F}_n}[\bar{G}(2 - n^{1/2}(\bar{X}_n - \mu_n) - n^{1/2}\mu_n)]},\end{aligned} \tag{6}$$

where $\bar{G}(t) = \int_t^\infty G(du)$ is the survival function of G. When $\mu_n < -\delta < 0$ for some $\delta > 0$, and G is the Laplace or Logistic distribution so that $\bar{G}$ has an exponential tail, the dominant term $\exp(n^{1/2}\mu_n)$ in both the numerator and the denominator will cancel out, making the selective likelihood ratio properly behaved in this difficult scenario.

It turns out that this selective likelihood ratio is fundamental to formalizing asymptotic properties of selective inference procedures. Its behavior determines not only the asymptotic convergence of the pivotal quantities like in (4), but also whether consistent estimation of the population parameters is possible with large samples.

Again in the negative mean scenario where $\mu_n < -\delta < 0$, the sample mean $\bar{X}_n$ surviving the nonrandomized "file drawer effect" cannot be a consistent estimator for the underlying means μ_n because it will always be positive. But if $\bar{X}_n$ is reported as in Example 2, it will be consistent for μ_n even if μ_n is negative and bounded away from 0. For detailed discussion, see Section 3.

In general, the behavior of the selective likelihood ratio can be used to study the asymptotic properties of selective inference procedures. We study consistent estimation and weak convergence for selective inference procedures in Section 3 and Section 5, respectively.

We are especially inspired by the field of differential privacy [cf. Dwork et al. (2015) and references therein] to study the use of randomization in selective inference. Privatized algorithms purposely randomize reports from queries to a database in order to allow valid interactive data analysis. To our understanding, our results are the first results related to weak convergence in privatized algorithms, as most guarantees provided in the differentially private literature are consistency guarantees. Some other asymptotic results in selective inference have also been considered in Tian and Taylor (2015), Tibshirani et al. (2015), though these have a slightly different flavor in that they marginalize over choices of models.

We conclude this section with some more examples.

1.2. *Linear regression*. Consider the linear regression framework with response $y \in \mathbb{R}^n$, and feature matrix $X \in \mathbb{R}^{n \times p}$, with X fixed. We make a homoscedasticity assumption that $\text{Cov}[y|X] = \sigma^2 I$, with σ^2 considered known. Of interest is

$$\mu = \mathbb{E}(y|X),$$

a functional of $\mathbb{F} = \mathbb{F}(X)$ the conditional law of y given X. When $\mathbb{F}$ is a Gaussian distribution, exact selective tests have been proposed for different selection procedures [Tian, Loftus and Taylor (2015), Tibshirani (1996), Tibshirani et al. (2016)]. Removing the Gaussian distribution on $\mathbb{F}$, Tian and Taylor (2015) showed that the same tests are asymptotically valid under some conditions.

Randomized selection in this setting is a natural extension of these works. Fithian, Sun and Taylor (2014) proposed to use a subset of data for model selection, which yields a significant increase in power. In this work, we study general randomized selection procedures. Consider the following example.

Due to the sparsity of the solution of LASSO Tibshirani (1996)

$$\hat{\beta}_\lambda(y) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \cdot \|\beta\|_1,$$

a small subset of variables can be chosen for which we want to report p-values or confidence intervals. This problem has been studied in Lee et al. (2016). However, instead of using the original response y to select the variables, we can independently draw $\omega \sim \mathbb{Q}$ and choose the variables using $y^* = y + \omega$. Specifically, we choose subset E by solving

$$\hat{\beta}_\lambda(y, \omega) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y^* - X\beta\|_2^2 + \lambda \cdot \|\beta\|_1, \qquad y^* = y + \omega, \tag{7}$$

and take $E = \text{supp}(\hat{\beta}_\lambda(y, \omega))$. In Section 4.2.2, we discuss how to carry out inference after this selection procedure, with much increased power. We also discuss the reason behind this increase in Section 4.2.

1.3. *Nonparametric selective inference*. All the previous works on selective inference assume a parametric model like the Gaussian family or the exponential family. In this work, we allow selective inference in a nonparametric setting. Consider the following examples.

Suppose in a classification problem we observe independent samples

$$(x_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{F}, \qquad (x_i, y_i) \in \mathbb{R}^p \times \{0, 1\},$$

with fixed p. This problem is nonparametric if we do not assume any parametric structure for $\mathbb{F}$ and are simply interested in some population parameters of the distribution $\mathbb{F}$. In Section 5, we developed asymptotic theory to construct an asymptotically valid test for the population parameters of interest. More details can be found in Section 5.4.1.

Also consider a multi-group problem where a response x is measured on p treatment groups. A special case is the two-sample problem where there are two groups. It is of interest to form a confidence interval for the effect size in the "best" treatment group. This arises often in medical experiments where multiple treatments are performed and we are interested to discover whether one of the treatment has a positive effect. The fact we have chosen to report the "best" treatment effect exposes us to selection bias and multiple testing issues [Benjamini and Hochberg (1995)] and therefore calls for adjustment after selection. Benjamini and Stark (1996) have considered the parametric setting where $x_j \overset{\text{i.i.d.}}{\sim} N(\mu_j, \sigma^2)$ for each group. Suppose for robustness, it is of interest to report the median effect size instead of the mean (assuming responses are not symmetric). Then without any assumptions on the distribution of the measurements, this also becomes a nonparametric problem. But we can apply the theory in Section 5 to cope with this problem; for details, see Section 5.3.

1.4. *Outline of the paper.* There are three main advantages of applying randomization for selective inference:

- Consistent estimation under the selective distribution.
- Increase in power for selective tests.
- Weak convergence of selective inference procedures.

In the following sections, Section 2 gives the setup of selective inference and introduced selective likelihood ratio, which is the key for studying consistent estimation and weak convergence of selective inference procedures. Section 4 focuses on linear regression models with different randomization schemes, demonstrating the increase in power. Section 5 proposes an asymptotic test for the nonparametric settings. Theorem 9 proves that the central limit theorem holds under the selective distribution with mild conditions. Applications to the two examples in Section 1.3 are discussed. This is a result for fixed dimension p. Finally, Section 6 discusses the possibility of extending our work to the setting, when multiple selection procedures are performed on different randomizations of the original data. One application is selective inference after cross validation for the square-root LASSO Belloni, Chernozhukov and Wang (2011).

2. Selective likelihood ratio. We first review some key concepts of selective inference. Our data D lies in some measurable space $(\mathcal{D}, \mathcal{F})$, with unknown sampling distribution $D \sim \mathbb{F}$. Selective inference seeks a reasonable probability model M—a subset of the probability measures on $(\mathcal{D}, \mathcal{F})$, and carry out inference in M. Central to our discussion is a *selection algorithm*, a set-valued map

$$\widehat{\mathcal{Q}} : \mathcal{D} \to \mathcal{Q}, \tag{8}$$

where $\mathcal{Q}$ is loosely defined as being made up of "potentially interesting statistical questions."

For instance, in the linear regression setting, $\mathcal{D} = \mathbb{R}^n$, our data $D = y$ and we have a fixed feature matrix $X \in \mathbb{R}^{n \times p}$. The unknown sampling distribution is $\mathbb{F} = \mathcal{L}(y|X)$, the conditional law of y given X.

In this work, the model selection procedure $\widehat{\mathcal{Q}}$ can be very general, and the models considered are not restricted to linear or even parametric models. However, as an example, a reasonable candidate for the range of $\widehat{\mathcal{Q}}$ might be all linear regression models indexed by subsets of $\{1, \ldots, p\}$ with known or unknown variance. For any selected subset of variables E, we carry out selective inference within the model $M = \{N(X_E\beta_E, \sigma^2 I), \beta_E \in \mathbb{R}^{|E|}\}$.

Since we use the data to choose the model M, it is only fair to consider the conditional distribution for inference,

$$D|M \in \widehat{\mathcal{Q}}(D), \qquad D \sim \mathbb{F}.$$

Therefore, we seek to control the selective Type-I error

$$\mathbb{P}_{M,H_0}(\text{reject } H_0|M \in \widehat{\mathcal{Q}}) \leq \alpha, \tag{9}$$

where M is the selected family of distributions in the range of $\widehat{\mathcal{Q}}$ and $H_0 \subset M$ is the null hypothesis. Selective intervals for parametric models M can then be constructed by inverting such selective hypothesis tests, though only the one-parameter case has really been considered to date.

2.1. *Randomized selection*. Randomized selection is a natural extension of the framework above. We enlarge our probability space to include some element of randomization. Specifically, let $\mathcal{H}$ denote an auxiliary probability space and $\mathbb{Q}$ is a probability measure on $\mathcal{H}$. A randomized selection algorithm is then simply

$$\widehat{\mathcal{Q}}^* : \mathcal{D} \times \mathcal{H} \to \mathcal{Q}.$$

Note the randomization is completely under the control of the data analyst and hence $\mathbb{Q}$ will be fully known. This is an extension of the nonrandomized selective inference framework in the sense that we can take $\mathbb{Q}$ to be the Dirac measure at 0. Many choices of $\widehat{\mathcal{Q}}^*$ are natural extensions of $\widehat{\mathcal{Q}}$, which we will see in many examples.

Randomized selective inference is simply based on the law $\mathbb{F}^*$, which we also call the *selective distribution*,

$$D|M \in \widehat{\mathcal{Q}}^*(D, \omega), \qquad (D, \omega) \sim \mathbb{F} \times \mathbb{Q}. \tag{10}$$

Note that although randomization is incorporated into selection, inference is still carried out using the original data D, after adjusting for the selection bias by considering the conditional distribution $\mathbb{F}^*$.

Similar to the selective inference we defined above, we seek to control the selective Type-I error,

$$\mathbb{P}_{\mathbb{F}^*}(\text{reject } H_0) = \mathbb{P}_{M,H_0}\big(\text{reject } H_0|M \in \widehat{\mathcal{Q}}^*\big) \leq \alpha. \tag{11}$$

Moreover, we also want to achieve good estimation, which makes

$$\mathbb{E}_{\mathbb{F}^*}\big((\hat{\theta}(y) - \theta(\mathbb{F}))^2\big) \tag{12}$$

small.

In Sections 3 to 5, we will discuss concrete examples of $\mathcal{D}$, D, $\mathbb{F}$ and $\widehat{\mathcal{Q}}^*$. But before that we first introduce the selective likelihood ratio, which is a crucial quantity in studying the selective distribution $\mathbb{F}^*$.

2.2. *Selective likelihood ratio.* Selective likelihood ratio provides a way of connecting the original distribution $\mathbb{F}$ and its selective counterpart $\mathbb{F}^*$. It is easy to see from (10) that the selective distribution is simply a restriction of the (D, ω)'s such that model M will be selected. Thus, $\mathbb{F}^*$ is absolutely continuous with respect to $\mathbb{F}$, and the *selective likelihood ratio* is

$$\begin{aligned} &\frac{d\mathbb{F}^*}{d\mathbb{F}}(D) = \frac{\mathbb{W}(M; D)}{\mathbb{E}_{\mathbb{F}}(\mathbb{W}(M; D))} = \ell_{\mathbb{F}}(D) \qquad \forall \mathbb{F} \in M, \\ &\mathbb{W}(M; D) = \mathbb{Q}\big(\{\omega : M \in \widehat{\mathcal{Q}}^*(D, \omega)\}\big). \end{aligned} \tag{13}$$

The numerator in $\ell_{\mathbb{F}}(D)$ is the restriction of (D, ω), integrated over the randomizations ω, and the denominator is simply a normalizing constant. One implication of the selective likelihood ratio is that for distributions $\mathbb{F}$ in parametric families, their selective counterparts may have the same parametric structure.

2.2.1. *Exponential families.* One commonly used parametric family is the exponential family. Assume that $\mathbb{F} = \mathbb{F}_\theta$ is an exponential family with natural parameter space $\mathbf{\Theta}$ and $\mathcal{D} = \mathbb{R}^n$ and the data $D = y$. Its density with respect to the reference measure $d\mathbb{F}_0$ is,

$$\frac{d\mathbb{F}_\theta}{d\mathbb{F}_0}(y) = \exp\{\theta^T T(y) - \psi(\theta)\}, \qquad \theta \in \mathbf{\Theta}. \tag{14}$$

Through the relationship in (13) we conclude, for any randomization scheme, the law $\mathbb{F}^*_{M,\theta}$ is another exponential family. Formally, is the below lemma.

LEMMA 2. *If $\mathbb{F}_\theta$ belongs to the exponential family in* (14)*, then for* any *randomized selection procedure $\widehat{\mathcal{Q}}^*$, the selective distribution is also an exponential family,*

$$\frac{d\mathbb{F}^*_{M,\theta}}{d\mathbb{F}_0}(y) \propto \mathbb{W}(M; y) \exp\{\theta^T T(y) - \psi(\theta)\}, \qquad \theta \in \mathbf{\Theta},$$

with the same sufficient statistic $T(y)$ and natural parameters θ.

Furthermore, to test $H_{0j} : \theta_j = 0$, *we consider the following law*:

$$T_j(y) \mid T_{-j}(y), \qquad y \sim \mathbb{F}^*_{M,\theta}. \tag{15}$$

The first claim of the lemma is quite straight-forward using the relationship in (13). The second claim is a Lehmann–Scheffé [cf. Chapter 4.4 in Lehmann (1986)] construction which was proposed in Fithian, Sun and Taylor (2014) to construct tests for one of the natural parameters treating the others as nuisance parameters. For detailed construction of such tests in the linear regression setting, see Section 4.

3. Consistent estimation after model selection. In this section, we leave the parametric setup and consider general models M. In particular, we study the consistency of estimators under the selective distribution for arbitrary models. We first introduce the framework of asymptotic analysis under the selective model. Then we state conditions for consistent estimation in Lemma 3 and conclude with examples.

For any model M, which is a collection of distributions, we define its corresponding *selective model*, which is the collection of corresponding selective distributions,

$$M^* = \left\{ \mathbb{F}^* : \frac{d\mathbb{F}^*}{d\mathbb{F}}(D) = \ell_{\mathbb{F}}(D), \mathbb{F} \in M \right\}, \tag{16}$$

where $\ell_{\mathbb{F}}(D)$ is the selective likelihood ratio for the selection event $\{M \in \widehat{\mathcal{Q}}^*\}$. Selective inference is carried out under the selective model M^*.

In order to make meaningful asymptotic statements, we consider a sequence of randomized selection procedures $(\widehat{\mathcal{Q}}^*_n)_{n\geq 1}$ and models $(M_n)_{n\geq 1}$ with each M_n in the range of $\widehat{\mathcal{Q}}^*_n$.

Often, we are interested in some population parameter θ_n, which can be thought be as a functional of the distribution $\mathbb{F}_n \in M_n$,

$$\theta_n : M_n \to \mathbb{R}.$$

It is worth pointing out that M_n is selected by $\widehat{\mathcal{Q}}^*_n$, which already incorporates the statistical questions we are interested in. In this sense, M_n is chosen a posteriori. The selected model M^*_n does not change our target of inference, it merely changes the distribution under which such inference should be carried out. In other words, if θ_n is the mean parameter, we are interested in the underlying mean of $\mathbb{F}_n$, not $\mathbb{F}^*_n$.

We might have a good estimator $\hat{\theta}_n : \mathcal{D} \to \mathbb{R}$ for $\theta_n(\mathbb{F}_n)$ under $\mathbb{F}_n$, namely

$$\mathbb{E}_{\mathbb{F}_n}\big[(\hat{\theta}_n - \theta_n(\mathbb{F}_n))^2\big] \to 0.$$

$\hat{\theta}_n$ is a consistent estimator if our model M_n is given a priori. But as we use data to select M_n, what really cares about is its performance under the selective distribution $\mathbb{F}^*_n$. Will this estimator still be consistent under the selective distribution $\mathbb{F}^*_n$?

Formally, we say an estimator $\hat{\theta}_n$ is uniformly consistent in L^p for $\theta_n(\mathbb{F}_n)$ under the sequence $(M_n)_{n\geq 1}$ if

$$\limsup_n \sup_{\mathbb{F}_n \in M_n} \|\hat{\theta}_n - \theta_n(\mathbb{F}_n)\|_{L^p(\mathbb{F}_n)} \to 0.$$

Similarly, we say that $\hat{\theta}_n$ is uniformly consistent in probability for the functional $\theta_n(\mathbb{F}_n)$ under the sequence $(M_n)_{n\geq 1}$ if for every $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that for all $\delta \geq \delta(\varepsilon)$

$$\limsup_n \sup_{\mathbb{F}_n \in M_n} \mathbb{F}_n\big(|\hat{\theta}_n - \theta_n(\mathbb{F}_n)| > \delta\big) \leq \varepsilon.$$

The following lemma states the conditions for consistency of $\hat{\theta}_n$ under the sequence of corresponding selective models $(M_n^*)_{n\geq 1}$.

LEMMA 3. *Consider a sequence $(\widehat{\mathcal{Q}}_n^*, M_n)_{n\geq 1}$ of randomized selection procedures and models. Suppose the selective likelihood ratios satisfies, for some $p > 1$,*

$$\limsup_n \sup_{\mathbb{F}_n \in M_n} \|\ell_{\mathbb{F}_n}\|_{L^p(\mathbb{F}_n)} < C. \tag{17}$$

Then for any sequence of estimators $\hat{\theta}_n$ uniformly consistent for $\theta_n(\mathbb{F}_n)$ in L^α, it is also uniformly consistent for $\theta_n(\mathbb{F}_n)$ in L^γ under $(M_n^*)_{n\geq 1}$, $\gamma \leq \alpha/q$, $\frac{1}{p} + \frac{1}{q} = 1$.

Further, if $\hat{\theta}_n$ is uniformly consistent for θ_n in probability, then $\hat{\theta}_n$ is uniformly consistent for θ_n in probability under the sequence $(M_n^*)_{n\geq 1}$.

The proof of the lemma is deferred to Section B in the Supplementary Material [Tian and Taylor (2018)].

The significance of Lemma 3 is that if we randomized before selection, many existing consistent estimators (like sample mean or variance) will remain consistent under moment conditions on the selective likelihood ratio $\ell_{\mathbb{F}}$. We illustrate the application of Lemma 3 through our "file drawer effect" examples in Section 1.1. We will also apply the consistency results in Section 5.5 when we plug in consistent estimators for noise variances.

3.1. *Revisit the "file drawer problem".* First, we note that in Example 1 and 2, we observe data $D_n = (X_{1,n}, \ldots, X_{n,n})$, with $X_{i,n} \sim \mathbb{F}_n$. The randomized selection in Example 2 can be realized as

$$\widehat{\mathcal{Q}}^*(D_n, \omega) = \begin{cases} \text{report } p\text{-values for } \bar{X}_n, & \text{if } \sqrt{n}\bar{X}_n + \omega > 2, \\ \text{do nothing}, & \text{if } \sqrt{n}\bar{X}_n + \omega \leq 2, \end{cases}$$

where we independently draw $\omega \sim G$.

By law of large numbers, we easily see that if we always report $\bar{X}_n$, it will be an unbiased estimator for μ_n. However, since we only observe the sample means surviving the file drawer effect. Will $\bar{X}_n$ still be consistent for μ_n?

In the most difficult scenario discussed in Section 1.1, where $\mu_n < -\delta < 0$ for some $\delta > 0$, $\bar{X}_n$ cannot be a consistent estimator for μ_n in Example 1. This is easy to see as Example 1 will only report positive sample means. A remarkable feature of randomized selection is that consistent estimation of the population parameters is possible even when the selection event has vanishing probabilities. In fact, the following lemma states that when G is a Logistic distribution, $\bar{X}_n$ is consistent for μ_n after the randomized file drawer effect in Example 2.

LEMMA 4. *Suppose as in Example 2, we observe a triangular array with* $X_{i,n} \sim \mathbb{F}_n$. $\mathbb{F}_n$ *has mean* $\mu_n = \mu < 0$. *If we draw* $\omega \sim \text{Logistic}(\kappa)$, *where* κ *is the scale of the Logistic distribution. Then the sample means* $\bar{X}_n$ *surviving the "randomized" file drawer effect are consistent for* μ,

$$\bar{X}_n \xrightarrow{p} \mu, \qquad \textit{conditional on } \sqrt{n}\bar{X}_n + \omega > 2,$$

if $\mathbb{F}_n$ *has moment generating function in a neighborhood of* 0. *Namely,* $\exists a > 0$, *such that*

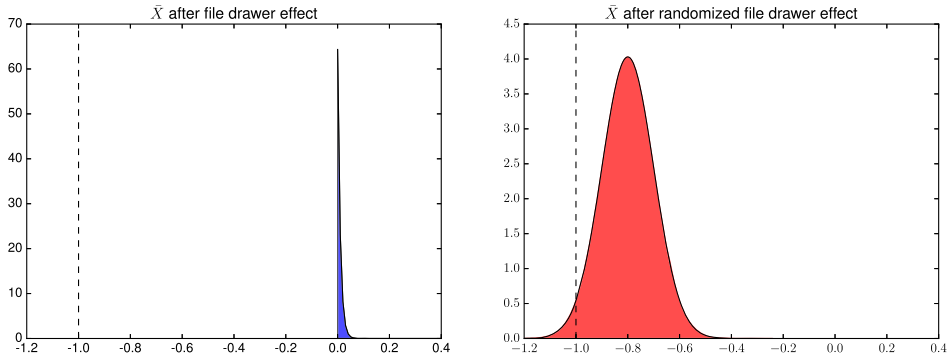
$$\mathbb{E}_{\mathbb{F}_n}\big[\exp\big(a|X_{i,n} - \mu_n|\big)\big] \le C.$$

Before we prove the lemma, we want to point out that although the selection procedure in Example 2 is different from that in Example 1 because of randomization, $\sqrt{n}\mu_n$ is still the dominant term in selection. Note that

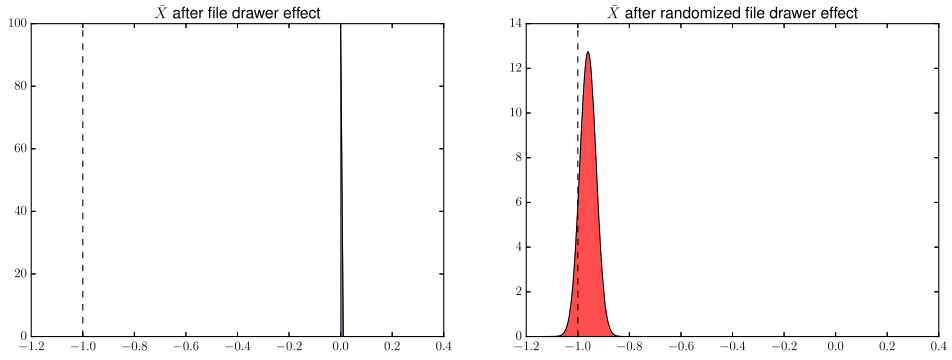
$$\sqrt{n}\bar{X}_n + \omega = \sqrt{n}\mu_n + \sqrt{n}(\bar{X}_n - \mu_n) + \omega.$$

Since both $\sqrt{n}(\bar{X}_n - \mu_n)$ and ω are $O_p(1)$ random variables, the dominant term $\sqrt{n}\mu_n \to -\infty$, would ensure that the selection event has vanishing probabilities in Example 2 as well. Thus it is particularly impressive that Example 2 gives consistent estimation where Example 1 cannot. The proof of Lemma 4 is deferred to Section C in the Supplementary Material [Tian and Taylor (2018)].

We also verified this theory of consistent estimation through simulations. Figure 1 shows the empirical distributions of the sample mean $\bar{X}_n$ after the file drawer effect in Example 1 or the "randomized" file drawer effect in Example 2. They are marked with "blue" colors or "red" colors, respectively. We set the true underlying mean to be $\mu_n = \mu = -1$ and mark it with the dotted vertical line in Figure 1. The upper panel Figure 1(a) is simulated with $n = 100$ and the lower panel Figure 1(b) is simulated with $n = 1000$. We notice that in both simulations, the sample mean in Example 1 concentrates around the thresholding boundary, which is positive. Thus, these sample means can not be possibly for the underlying mean $\mu = -1$. However, the existence of randomization allows us to report negative sample means. As a result, the sample mean in Example 2 will be consistent for $\mu = -1$. We see that as we increase sample size n, the sample means concentrates closer to $\mu = -1$.



(a) $n = 100$



(b) $n = 1000$

FIG. 1. *Empirical distributions of sample means $\bar{X}_n$ in Example 1 and Example 2, with original or randomized file drawer effect. For the randomization, we draw $\omega \sim \text{Logistic}(\kappa)$, with $\kappa = 0.5$. The true mean -1 is marked with the dashed vertical line.*

4. Inference in linear regression models. In the linear regression setting, we assume a fixed feature matrix $X \in \mathbb{R}^{n \times p}$, and observe the response vector $D = y \in \mathbb{R}^n$. We assume the noises are normally distributed. There are two ways to parametrize a linear model, and both belong to some exponential family. Now we introduce the *selected model*,

$$M_{\text{sel}}(E) = \{N(X_E \beta_E, \sigma^2 I) : \beta_E \in \mathbb{R}^{|E|}\}, \qquad E \subset \{1, \ldots, p\} \tag{18}$$

with σ^2 known or unknown or the *saturated model*,

$$M_{\text{sat}} = \{N(\mu, \sigma^2 I) : \mu \in \mathbb{R}^n\} \tag{19}$$

with known variance. Now we consider some randomized selection procedures and inference after selection.

4.1. *Data splitting and data carving.* In the Introduction, we introduced *data splitting* [Cox (1975)] as a special case of randomized selective inference. In Fithian, Sun and Taylor (2014), the term *data carving* was introduced to demonstrate that data splitting is inadmissible. In data splitting (and data carving) inference makes most sense in the selected model $M_{\text{sel}}(E)$, hence we should think of $\widehat{\mathcal{Q}}$ as returning a subset E of variables selected.

Let us formalize this notion in our notation. Let $\mathbb{Q}$ be some measure on assignments of n data points into groups and $\widehat{\mathcal{Q}}$ a selection algorithm defined on datasets of any size. The distribution $\mathbb{Q}$ determines a randomized selective inference procedure with selection algorithm $\widehat{\mathcal{Q}}^*$, an algorithm applied to subsets of the original data set. In this case, it is easy to see that

$$\mathbb{W}(E; y) \overset{D}{=} \mathbb{W}\big(M_{\text{sel}}(E); y\big) \propto \sum_{\omega} q_\omega \cdot 1_{\{M_{\text{sel}}(E) \in \widehat{\mathcal{Q}}(y_1(y,\omega))\}},$$

where q_ω is the mass assigned to assignment ω by $\mathbb{Q}$. Multiple assignments or splits considered in Meinshausen and Bühlmann (2010), Meinshausen, Meier and Bühlmann (2009) can be formalized in a similar fashion. We can construct UMPU tests for β_E in the selected model $M_{\text{sel}}(E)$ by using Lemma 2 [also see Fithian, Sun and Taylor (2014)]. We note that in Fithian, Sun and Taylor (2014) the authors conditioned unnecessarily on the split ω, and we would expect that aggregating over splits would yield a more powerful procedure.

However, there are two disadvantages with this randomization scheme. First, it is computationally difficult to aggregate over all random splits. Second, it seems difficult to consider the saturated model M_{sat} for inference, which is more robust to model misspecifications. To overcome those difficulties, we introduce other randomization schemes below.

4.2. *Additive noise and more powerful tests.* Our second randomization scheme in linear regression involves additive noise. Specifically, we draw $\omega \sim \mathbb{Q}$ and use the randomized response $y^*(y, \omega) = y + \omega$ for selection In this case, we can consider both the selected model $M_{\text{sel},E}$ and the saturated model $\mathbb{M}_{\text{sat}}$. Per Lemma 2, we can perform valid inference for β_E in $M_{\text{sel},E}$ or linear functionals of μ in M_{sat}.

One major advantage of using a randomized response y^* for selective inference is that these procedures yield much more powerful tests, at a small cost of on the quality of the selected models. In other words, small amount of randomization causes a small loss in the model selection stage, but we gain much more power in the inference stage.

The reason for increased power can be explained by a notion called *leftover Fisher information* first introduced in Fithian, Sun and Taylor (2014). Since selective inference is essentially inference under the selective distribution $\mathbb{F}_n^*$, the Fisher information under $\mathbb{F}_n^*$ would determine how efficient the selective tests are. In the saturated model with Gaussian noise M_{sat}, $\frac{y-\mu}{\sigma^2}$ is the score statistic and its

variance under $\mathbb{F}_n^*$ is exactly the leftover Fisher information (a similar relationship holds in the selected model $M_{\text{sel},E}$). Lemma 5 gives a lower bound on this leftover Fisher information when the randomization noise $\mathbb{Q} = N(0, \gamma^2 I)$.

LEMMA 5. *For either M_{sat} or $M_{\text{sel}}(E)$, if we use Gaussian randomization noise $\mathbb{Q} = N(0, \gamma^2)$, and the selection is based on $\widehat{\mathcal{Q}}(y^*) = \widehat{\mathcal{Q}}(y + \omega)$, then the leftover Fisher information is bounded below by*

$$(1 - \tau)\mathcal{I}(\theta), \qquad \tau = \sigma^2/(\sigma^2 + \gamma^2),$$

and $\mathcal{I}(\theta)$ is the nonselective Fisher information for θ in M_{sat} or $M_{\text{sel}}(E)$. The parameters θ depend on which of the two models we are considering.

The proof of the lemma is deferred to Section B in the Supplementary Material [Tian and Taylor (2018)].

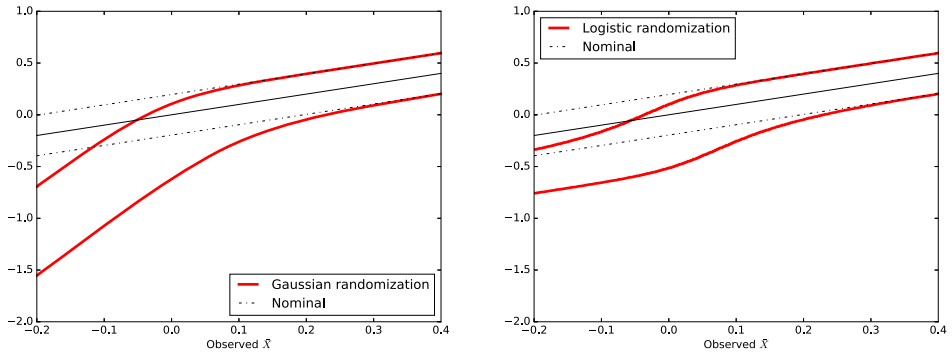
It is worth noting that the scale of the added randomization noise is an important topic. It is analogous to choosing the sample size to hold out when using data splitting (or data carving) for valid selective inference. Unfortunately, there is no obvious objective function in the data splitting scenario or the additive randomization scheme. Typically we have chosen the scale of the additive noise so that its variance is some small to medium multiple of the variance of the score statistic for the specific model we are considering. In this article, we choose this multiple to be around 0.25 ($\gamma = 0.5$), which corresponds to holding out about 20% of the data. In some related work with a different objective Harris (2016), one of the authors advocates a shrinking multiple of order $n^{-1/4}$.

When there is no randomization $\gamma = 0$, we potentially have no leftover Fisher information. This corresponds to a very rare selection event. However, after randomization, even with very extreme selection, there is always leftover Fisher information, which makes the selective tests more powerful. Consider the following examples.

4.2.1. *Revisit the "file drawer problem".* In Example 1 and Example 2, if we assume $\mathbb{F}_n = N(\mu, 1)$, they are a special case of the linear regression model, with the feature matrix $X = \mathbf{1}$, the all ones vector.

In this case, $n\bar{X}_n$ is the score statistic, and its variance under the selective distribution is the Fisher information. Lemma 5 states that the leftover Fisher information is lower bounded by $n(1 - \tau)$ if we draw randomize using Gaussian variables, $\mathbb{Q} = N(0, \gamma^2)$, $\tau = 1/(1 + \gamma^2)$.

Moreover, the increase in leftover Fisher information with randomization is not specific to Gaussian randomizations. For example, in Figure 1 when we use Logistic randomization, we also observe that under the selective distribution with randomization, $\bar{X}_n$ has a much bigger variance than without randomization. As discussed above, this variance multiplied by n^2 is exactly the leftover Fisher information, which explains why selective procedures after randomization will have better performances than without.



(a) Gaussian added noise (b) logistic added noise

FIG. 2. *Selective confidence intervals for different added noise. The solid black line is the observed value, and the dashed lines are the nominal confidence intervals which will not have the proper coverage for the true mean. The red lines are the selective confidence intervals that have exact* 90% *coverage for the true mean. Due to the selection bias, the observed value* $\bar{X}$ *will be biased up for the true mean.*

We investigate the relationship between the leftover Fisher information and the length of confidence intervals constructed by inverting the pivot in (4). Specifically, in Example 2, after observing a reported sample mean, we want to report confidence intervals for the underlying mean μ.

Figure 2 demonstrates the selective intervals (solid lines) after (3) with ω being either Gaussian or Logistic noises. The sample size $n = 100$. Unlike the nominal confidence intervals (dashed lines), the selective intervals are valid with 90% coverage for the underlying mean. Since Lemma 3 gives a lower bound of $(1-\tau)\mathcal{I}(\mu)$, we would intuitively expect the selective confidence intervals to be $1/(1-\tau)$ the length of the nominal intervals. This is verified in Figure 2(a), when we observe really negative sample means. (The sample means can be negative because we added randomization.) On the other hand, for Logistic randomization in Figure 2(b), the intervals are slightly wider than the nominal intervals around the $2/\sqrt{n}$, but narrow to roughly the nominal size on both sides of the truncation point. This indicates that added logistic noise might preserve more information than Gaussian additive noise. Both additive noises improve significantly over a nonrandomization scheme [cf. Figure 3 in Fithian, Sun and Taylor (2014)].

Of course, the increase in power and shortening of selective confidence intervals does not come without a price. Because we select with a randomized response, we are likely to select a worse model. But the trade-off between model quality and power is highly in favor of randomization. See the following example.

4.2.2. *Linear regression with added noise.* Back to the general setup of linear regression models, we select a model by solving LASSO with the randomized

response $y^* = y + \omega$ and return the active set E of the solution [as in (7)]. Then per Lemma 2, we can construct valid selective tests in both M_{sat} and $M_{\text{sel}}(E)$. For instance, in $M_{\text{sel}}(E)$, we can construct tests for the hypothesis $H_{0j} : \beta_j = 0$, $j \in E$ based on the law,

$$\eta^T y | A_E(y + \omega) \leq b_E, P_{E\setminus j} y, \qquad (y, \omega) \sim N\big(X_E \beta_E, \sigma^2 I\big) \times \mathbb{Q}, \beta_j = 0, \tag{20}$$

where $\eta = (X_E^{\dagger})^T e_j$, e_j is the jth column of the identity matrix, $P_{E\setminus j}$ is the projection matrix onto the column space of E but orthogonal to η, and A_E, b_E are the appropriate matrix and vector corresponding to LASSO selection. This is a UMPU test due to the Lehmann–Scheffé construction [Fithian, Sun and Taylor (2014)] and controls the selective Type-I error (11). Although, we cannot compute the explicit forms of (20); the selection events in (20) are polyhedrons and thus a hit-and-run or Hamiltonian Monte Carlo algorithm [Pakman and Paninski (2014)] can be used for sampling.

Figure 3 compares inference in the additive Gaussian noise scheme to the data carving procedure proposed in Fithian, Sun and Taylor (2014) as well as data splitting. In $M_{\text{sel}}(E)$, the probability of screening (i.e., selecting E including all the nonzero β's) is a surrogate for the quality of the model. As additive noise uses a different randomization scheme than data splitting and data carving, we vary the amount of randomization used in each scheme and match on the probability of screening. Thus Figure 3 is like an ROC curve for the trade-off between model quality and power of tests. The x-axis goes in the direction of increased randomization, with the left most point corresponding to no randomization at all. We see even with a small randomization that barely affects model selection, we can substantially lower the Type-II error from 0.2 to less than 0.05. The trade-off is highly in favor of (small) randomization. We see in Figure 3 that additive noise lowers

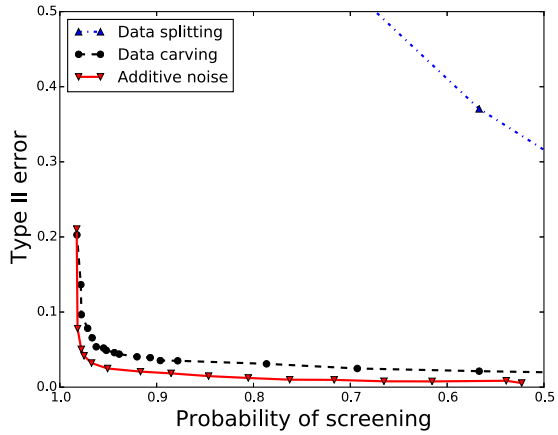


FIG. 3. *Comparison of inference in additive noise randomization vs. data carving.*

the Type-II error by almost half than data carving for the same screening probability and they both clearly dominate data splitting. For the concrete setup of the simulation, see Chapter 7 of Fithian, Sun and Taylor (2014).

5. Weak convergence and selective inference for statistical functionals. In the nonparametric setting, we assume a triangular array of data, $D_n = (d_{1,n}, \ldots, d_{n,n})$, and $d_{i,n} \overset{\text{i.i.d.}}{\sim} \mathbb{F}_n$. When $\mathbb{F}_n = \mathbb{F}$, it is the special case of independent sampling. We are interested in some functional of the distribution $\mu_n = \mu(\mathbb{F}_n)$. Associated with μ_n is our statistic T which is a *linearizable statistic* [Chung and Romano (2013)].

DEFINITION 6 (Linearizable statistic). Suppose $d_{i,n} \overset{\text{i.i.d.}}{\sim} \mathbb{F}_n$, we call T a linearizable statistic for $\mu_n = \mu(\mathbb{F}_n)$ if, for any sample size n,

$$
\begin{aligned}
T(D_n) &= \frac{1}{n}\sum_{i=1}^{n} \xi_{i,n} + R, \qquad \xi_{i,n} = \xi(d_{i,n}),\\
\mathbb{E}[\xi_{i,n}] &= \mu_n \in \mathbb{R}^p, \qquad \operatorname{Cov}[\xi_{i,n}] = \Sigma_n \in \mathbb{R}^{p\times p},
\end{aligned} \tag{21}
$$

where ξ a function of the data and R is bounded with probability 1, $R = o_p(n^{-\frac{1}{2}})$ under $\mathbb{F}$. We use the slight abuse of notation to denote $\xi_{i,n}$ as i.i.d. random variables as well.

Throughout this section, we assume the dimension p is fixed. We are interested in establishing a pivotal quantity for $T_n = T(D_n)$ like (4) in Example 2 where T_n is the sample mean after the randomized "file drawer effect." It turns out we have an exact pivotal quantity if T_n is normally distributed. To lighten notation, we suppress the script n in the following lemma, which is a finite sample result valid for any n. We prove the lemma in Section B.

LEMMA 7. *If the statistic T is normally distributed from $N(\mu, \frac{\Sigma}{n})$ and the model M is selected by randomized selection $\widehat{\mathcal{Q}}^*(T, \omega)$, where $\omega \sim \mathbb{Q}$. Then for any contrast η, which could depend on the outcome of selection $\widehat{\mathcal{Q}}^*$, we have*

$$
\begin{aligned}
P\big(T; \eta^T\mu, \Sigma\big) &= \frac{\int_{\eta^T T}^{\infty} \mathbb{Q}(t; V_\eta)\cdot \exp(-n(t-\eta^T\mu)^2/2\sigma_\eta^2)\,dt}{\int_{-\infty}^{\infty} \mathbb{Q}(t; V_\eta)\cdot \exp(-n(t-\eta^T\mu)^2/2\sigma_\eta^2)\,dt}\\
&\overset{\mathbb{F}^*}{\sim} \operatorname{Unif}(0,1),
\end{aligned} \tag{22}
$$

where

$$
\sigma_\eta^2 = \eta^T\Sigma\eta, \qquad V_\eta = \left(I - \frac{1}{\sigma_\eta^2}\Sigma\eta\eta^T\right)T,
$$

$$
\mathbb{Q}(t, V_\eta) = \mathbb{Q}\big(\{\omega : M \in \widehat{\mathcal{Q}}^*\big(t\cdot \Sigma\eta/\sigma_\eta^2 + V_\eta, \omega\big)\}\big).
$$

REMARK 8. In selected models $M_{\text{sel},E}$, the selection is often made not only based on (T, ω), but also other statistic of the data, which we call the null statistic N. Thus the selection event should be expressed as $\{M \in \widehat{\mathcal{Q}}^*((T, N), \omega)\}$. To make notation simpler, we exclude such possibilities. But a slightly modified pivot where we replace $\mathbb{Q}(t; V_\eta)$ with $\mathbb{Q}(t; V_\eta, N)$ in (22) and integrate over N, is still Unif(0, 1) distributed.

Note that Lemma 7 provides a valid pivotal quantity for any randomized selection procedure $\widehat{\mathcal{Q}}^*$ and any randomization noise $\mathbb{Q}$ provided that T is normally distributed. In fact, Lemma 7 does not require T to be a linearizable statistic. In some sense, the lemma is a reformulation (after rescaling) of the selective tests constructed in the linear regression model with additive noises (see Section 4.2.2). For example, in the selected model $M_{\text{sel},E}$, to test the hypothesis $H_{0j} : \beta_j = 0$, $j \in E$, we consider the law (20). After introducing the null statistic $N = P_E^\perp y$, the pivot in (22) is in fact the CDF transform of this law, taking $T = P_E y$, $\Sigma = n\sigma^2 P_E$, and the selection event $\{M \in \widehat{\mathcal{Q}}^*((T, N), \omega)\}$ to be the affine selection event defined in (20). With simple calculation, it is easy to see $V_\eta = (P_E - \|\eta\|^{-2}\eta^T\eta)y = P_{E\setminus\eta}y$, which we condition on in both (22) and (20).

Of course the pivot in (22) is very difficult to compute explicitly, and we need to use sampling schemes like in (20). But in a nutshell, $P(T; \eta^T\mu, \Sigma)$ is simply a CDF transform of the law

$$\eta^T T \mid V_\eta, M \in \widehat{\mathcal{Q}}^*(T, \omega), \qquad (T, \omega) \sim N\left(\mu, \frac{\Sigma}{n}\right) \times \mathbb{Q}. \tag{23}$$

After introducing the null statistic, Lemma 7 is agnostic to the selected model $M_{\text{sel},E}$, where $\mu = X_E\beta_E$ or the saturated model M_{sat}, where the parameter is simply μ. The nuances between the two models in terms of sampling is that the saturated model conditions on N (treating it as part of V_η), but selected model integrates over N.

Lemma 7 is written with T implicitly being the approximate average of n i.i.d. variables, hence the distribution $N(\mu, \frac{\Sigma}{n})$. Linearizable statistics are of particular interest as they converge to $N(\mu, \frac{\Sigma}{n})$ due to central limit theorem. In the following, we seek to establish conditions under which the pivot $P(T; \mu, \Sigma)$ will be asymptotically Unif(0, 1).

5.1. *Selective central limit theorem.* In other work on asymptotics of selective inference [Tian and Taylor (2015), Tibshirani et al. (2015)], the setup considered is usually the saturated model M_{sat}. These works considered asymptotics of selective inference marginalized over the range of $\widehat{\mathcal{Q}}^*$. In contrast, we consider the convergence for any particular selected model M_n, under the conditional law of the selection event $\{M_n \in \widehat{\mathcal{Q}}_n^*\}$. Specifically, we allow weak convergence of the pivot in (22) in the sequence of selected models $(M_n)_{n\geq 1}$. As explained above, selected models integrate over the null statistics while saturated models condition on

those, thus the selective tests should have more power provided that the selected model is believable. In the saturated model, our result provides a finer measure of convergence than in Tian and Taylor (2015). On the other hand, Tian and Taylor (2015) allows high-dimensional setting in some cases while we consider fixed dimension p.

Similar to the asymptotic setting in Section 3, we consider the convergence of $P(T_n; \eta^T \mu_n, \Sigma_n)$ under a sequence of models $(M_n)_{n\geq 1}$ selected by a sequence of selection procedures $(\widehat{\mathcal{Q}}_n^*)_{n\geq 1}$. $(T_n)_{n\geq 1}$ is a sequence of linearizable statistics defined in Definition 6, with asymptotic mean μ_n and asymptotic covariance matrix $\frac{\Sigma_n}{n}$.

It turns out that in this setting, the selective likelihood ratio $\ell_{\mathbb{F}_n}$ again plays an important role in the convergence of the pivot. Recall that with randomized selection $\widehat{\mathcal{Q}}^*(T_n, \omega)$, the selective likelihood is

$$\begin{aligned} \ell_{\mathbb{F}_n}(T_n; M_n) &= \frac{\mathbb{W}(T_n; M_n)}{\mathbb{E}_{\mathbb{F}_n}[\mathbb{W}(T_n; M_n)]}, \\ \mathbb{W}(T_n; M_n) &= \mathbb{Q}\big(\{\omega : M_n \in \widehat{\mathcal{Q}}_n^*(T_n, \omega)\}\big). \end{aligned} \tag{24}$$

It will be convenient to rewrite the likelihood ratio in terms of the normalized vector $Z_n = \sqrt{n}(T_n - \mu_n)$

$$\bar{\ell}_{\mathbb{F}_n}(Z_n) = \ell_{\mathbb{F}}\big(n^{-1/2} Z_n + \mu_n\big), \tag{25}$$

as well as the pivot (22)

$$\bar{P}_{\mathbb{F}_n}(Z_n) = P\big(n^{-1/2} Z_n + \mu_n; \eta_n^T \mu_n, \Sigma_n\big). \tag{26}$$

Our approach is basically a comparison of how the pivot will behave under $\mathbb{F}_n$ and its Gaussian counterpart $\Phi_n = N(\mu(\mathbb{F}_n), \Sigma(\mathbb{F}_n))$. Specifically, it is a modification of the proof of Theorem 1.1 of Chatterjee (2005), modified to allow for the fact the derivatives of the pivot and the likelihood are not required to be uniformly bounded. Given a norm Ω on $\mathbb{R}^p$, define

$$\lambda_r^{\Omega}(f) = \sup_{\substack{s\in\mathbb{R}^p \\ 1\leq k\leq r}} \big\{\big\|\partial^k f(s)\big\|^{r/k} \exp\big(-r\Omega(s)\big) : 1 \leq k \leq r\big\}, \tag{27}$$

where ∂^k denotes the k-fold differentiation with respect to the p-dimensional vector s, $\|\cdot\|$ denotes element wise maximum.

Now we state our selective central limit theorem, which we prove in Section B.

THEOREM 9 (Selective central limit theorem). *Suppose the statistics $T_n = T(D_n)$ are linearizable statistics according to Definition 6. We also assume the norms $\Omega : \mathbb{R}^p \to \mathbb{R}$ are such that for each $f \in \{\bar{P}_n, \bar{\ell}_{\mathbb{F}_n}, \bar{\ell}_{\Phi_n}\}$, it satisfies*

$$\sup_{\mathbb{F}_n \in M_n} \lambda_3^{\Omega}(f) \leq C_1. \tag{28}$$

Moreover, assume $\xi_{i,n}$ has uniformly bounded moment generating function in some neighborhood of 0. *Namely,* $\exists a > 0$, *such that*

$$\sup_{n\geq 1}\sup_{\mathbb{F}_n\in M_n}\mathbb{E}_{\mathbb{F}_n}\big(\exp\big(a\big\|\xi_{i,n}-\mu(\mathbb{F}_n)\big\|_1\big)\big)\leq C_2. \tag{29}$$

Furthermore, we assume

$$\limsup_n n^{1/2}\cdot\frac{\mathbb{P}_{(\mathbb{F}_n\times\mathbb{Q})}[M_n\in\widehat{\mathcal{Q}}_n^*]-\mathbb{P}_{(\Phi_n\times\mathbb{Q})}[M_n\in\widehat{\mathcal{Q}}_n^*]}{\mathbb{P}_{(\Phi_n\times\mathbb{Q})}[M_n\in\widehat{\mathcal{Q}}_n^*]}\leq C_3. \tag{30}$$

Then, for any g with uniformly bounded derivatives up to third order

$$\left|\mathbb{E}_{\mathbb{F}_n^*}\big[g\big(P(T_n)\big)\big]-\int_0^1 g(x)\,dx\right|\leq n^{-1/2}K(g,C_1,C_2,C_3,p),\qquad n\geq n_0, \tag{31}$$

where K depends only on the bounds on the derivatives of g, the constants C_1, C_2, C_3 and the dimension p. Thus the convergence is uniform in $(M_n)_{n\geq 1}$ for models satisfying (28), (29) *and* (30).

Theorem 9 provides a finite sample bound on the convergence of the pivot $P(T_n)$. Since we allow g to be functions with uniformly bounded derivatives up to the third order, (31) implies convergence of $P(T_n)$ to Unif(0, 1) under $\mathbb{F}_n^*$. In the following examples, we show how to verify conditions (28), (29) and (30).

5.2. *Revisit the "file drawer problem".* In Examples 1 and 2, we considered only reporting an interval or a p-value about μ_n when $n^{1/2}\bar{X}_n > 2$ or $n^{1/2}\bar{X}_n+\omega > 2$. This is an example where we do not really select a model, but rather select only a proportion of the data to report. The selective distribution simply refers to the law of the reported sample means, which pass the threshold.

The data we observe is $D_n = (X_{1,n},\dots,X_{n,n})$ with the linearizable statistic T_n simply being the sample mean $\bar{X}_n$. Example 1 corresponds to the degenerate randomization of adding 0 to $\bar{X}_n$. Tian and Taylor (2015) shows that in order for the corresponding pivot to converge weakly we can take, for $\Delta < 0$ fixed,

$$M_n = \big\{F : \mathbb{E}_F[\bar{X}_n] > n^{-1/2}\Delta, \mathbb{E}_F\big[X_{i,n}^3\big] < \infty\big\}. \tag{32}$$

That is, $\bar{X}_n$ will satisfy a selective CLT when the population mean is not too negative.

On the other hand, in Example 2, the pivot in (22) is of the form

$$P(\bar{X}_n) = \frac{\int_{\bar{X}_n}^{\infty}\bar{G}(2-\sqrt{n}t)e^{-n(t-\mu_n)^2/2}\,dz}{\int_{-\infty}^{\infty}\bar{G}(2-\sqrt{n}t)e^{-n(t-\mu_n)^2/2}\,dz}, \tag{33}$$

and likelihood $\ell_{\mathbb{F}_n}(\bar{X}_n)$ is defined in (6).

When G is the Logistic noise, then condition (28) and (30) can be verified. Formally, we have the following lemma whose proof we defer to Section C in the Supplementary Material [Tian and Taylor (2018)].

LEMMA 10. *If $G = \text{Logistic}(\kappa)$, with κ being the scale parameter, then if centered $X_{i,n}$'s have moment generating functions in the neighborhood of zero, then the pivot $P(\bar{X}_n)$ is asymptotically* Unif(0, 1).

In other words, with Logistic randomization noise, we can take the sequence of models to be

$$M_n = \{\mathbb{F}_n : \mathbb{E}_{\mathbb{F}_n}[\exp(a|X_{1,n} - \mu_n|)] < \infty\} \qquad \text{for some } a > 0. \tag{34}$$

Requiring exponential moments is stricter than the third moment condition in (32), but we would have a stronger conclusion, namely weak convergence uniformly over all μ_n's.

5.3. *Two-sample median problem.* In the two-sample median problem, we have two treatment groups from which we take measurements, $x_{1i} \overset{\text{i.i.d.}}{\sim} \mathbb{F}_1$ and $x_{2i} \overset{\text{i.i.d.}}{\sim} \mathbb{F}_2$; for simplicity of notation, we assume we observe n samples from each group, and drop n in the subscript. We will report the bigger median from this group in the nonrandomized setting. Exact formulation of randomized selection will be discussed below.

Suppose our underlying distribution is $\mathbb{F} = \mathbb{F}_1 \times \mathbb{F}_2$. Let $\mu = (\mu_1, \mu_2)$ be the population median of the two groups, and $T = (T_1, T_2)$ be the sample median. The well-known result by Bahadur (1966) states that the sample median is a linearizable statistic for the median when the CDF of the distribution F has positive density f, and f' is bounded in a neighborhood of the population median m. Formally, if $x_i \overset{\text{i.i.d.}}{\sim} F$, then the sample median

$$T(x_1, \ldots, x_n) = m + \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}\{x_i > m\} - 1/2}{F'(m)} + R_n, \tag{35}$$

with $R = O(n^{-3/4}\log n)$ with probability 1.

Our (randomized) selection algorithm $\widehat{\mathcal{Q}}^*$ reports

$$\begin{cases} P(T; \mu_1, \Sigma), & \text{if } T_1 > T_2 + n^{-1/2}\omega, \\ P(T; \mu_2, \Sigma), & \text{if } T_1 \le T_2 + n^{-1/2}\omega, \end{cases}$$

where $\omega \sim \mathbb{Q}$ and $\Sigma = \text{diag}(\frac{1}{4}f_1(\mu_1)^{-2}, \frac{1}{4}f_2(\mu_2)^{-2})$ is a diagonal matrix. f_1, f_2 are the densities of $\mathbb{F}_1$ and $\mathbb{F}_2$. Without loss of generality, we suppose M_1 is selected, i.e. the first group is the "best" group.

We choose the randomization noise $\mathbb{Q}$ to be a Logistic(κ) with mean 0 and κ is the scale, and let G_κ be the CDF. The resulting pivot for μ_1 is

$$P(T; \mu_1, \Sigma) = \frac{\int_{T_1}^{\infty} G_\kappa(\sqrt{n}t - \sqrt{n}T_2) \cdot \exp(-n(t-\mu_1)^2/2\sigma_1^2)\,dt}{\int_{-\infty}^{\infty} G_\kappa(\sqrt{n}t - \sqrt{n}T_2) \cdot \exp(-n(t-\mu_1)^2/2\sigma_1^2)\,dt},$$

$$\sigma_1^2 = \frac{1}{4f_1(\mu_1)^2}.$$

This pivot strikes a similarity with the pivot in (33) for Example 2 with the truncation threshold 2 being replaced by $\sqrt{n}T_2$ and plugging in the appropriate means and variances of the medians. A result similar to Lemma 10 can be established, which ensures convergence of the pivot uniformly for any underlying medians (μ_1, μ_2).

In order to construct the above pivot, we need knowledge of the variance σ_1^2. Without selection, there are natural estimates of this variance. One may ask, how will inference be affected if we plug this estimate into our pivot? We revisit this question in Section 5.5.

5.4. *Affine selection events*. In this section, we discuss the special case of affine selection events (regions). This combined with the asymptotic result in Theorem 9 applies to more general settings. In particular, it allows us to approximate nonaffine regions. For a concrete example, see Section 5.4.1.

We drop the subscript n where possible to simplify notation. Suppose for our model M, the selection is based on (T, ω), and the selection event $\{M \in \widehat{\mathcal{Q}}^*\}$ can be described as

$$\{\sqrt{n}A_M T + \omega \in K_M\},$$

where the affine matrix $A_M \in \mathbb{R}^{d\times p}$ and K_M is a region in $\mathbb{R}^d$. Many examples of nonrandomized selective inference can be expressed in this way [cf. Fithian et al. (2015), Lee et al. (2016), Tibshirani et al. (2016)]. In this section, we provide conditions under which Theorem 9 can be applied.

We again normalize T to be $Z = \sqrt{n}(T - \mu)$, then the selection event can be rewritten as

$$\{A_M(Z + \Delta) + \omega \in K_M\}, \tag{36}$$

where $\sqrt{n}\mu = \Delta$, Z converges to $N(0, \Sigma)$.

Suppose $\omega \sim \mathbb{Q}$, which has distribution function G. Then we introduce some conditions on the selection region K_M and the added noise distribution G,

Lower bound: We assume there is some norm h, such that

$$\int_{K_M-\theta} G(dw) \geq C^- \exp\Big[-\inf_{w\in K_M-\theta} h(w)\Big], \qquad \forall\theta \in \mathbb{R}^d.$$

Smoothness: Suppose G has density g, we assume the first 3 derivatives of g are integrable,

$$\int_{\mathbb{R}^d} \|\partial^j g(w)\|\, dw \leq C_j, \qquad j = 0, 1, 2, 3,$$

where the norm on the left-hand side is the maximum element-wise of the partial derivatives.

The above two conditions essentially require G to be differentiable and have heavier tails than (or equal to) exponential tails. In fact we prove that the lower bound and smoothness conditions ensure that (28) are satisfied under the *local alternatives* introduced below.

DEFINITION 11 (Local alternatives). For the sequence of selected model $(M_n)_{n\geq 1}$, we define the local alternatives of radius of B to be the set all sequences $(\mu_n)_{n\geq 1}$, such that

$$d_h(0, K_{M_n} - A_{M_n}\Delta) \leq B, \qquad \Delta = \sqrt{n}\mu_n,$$

where $d_h(\cdot, \cdot)$ is the distance induced by the norm h.

The notion of local alternatives is natural in the asymptotic setting as we expect even a small effect size will be more prominent when we collect more and more data.

Formally, we have the following lemma, the proof of which is deferred to Section D in the Supplementary Material [Tian and Taylor (2018)].

LEMMA 12. *Suppose G, K_M satisfy the lower bound and smoothness conditions, then condition* (28) *are satisfied under the local alternatives.*

Now, we are left to verify conditions (29) and (30). Condition (29) is essentially a moment condition on the centered statistics $\xi_{i,n} - \mu_n$, which we have to assume. Condition (30) can be verified using the well-known results in multivariate CLT [see Götze (1991)]. To be rigorous, we state the following lemma, which we also prove in Section D of the Supplementary Material [Tian and Taylor (2018)].

LEMMA 13. *If $\mathbb{F}_n$ is such that the centered statistics $\xi_{i,n} - \mu_n$ have finite third moments, then under the local alternatives, condition* (30) *is satisfied.*

To summarize, Lemma 12 and Lemma 13 state that if G has integrable derivatives and exponential tails, then the pivot in (22) converges to Unif(0, 1) uniformly for $\mathbb{F}_n^*$ so long as $\mathbb{F}_n$'s are such that $\xi_{i,n} - \mu_n$ have exponential moments in a neighborhood of 0.

Unlike the sample mean and sample median examples, the pivot is difficult to compute explicitly in this case. However, as we discuss in the beginning of Section 5, the pivot is essentially the CDF transform of the conditional law (23), which we can sample from. As discussed above, we can just take ω to be from a Logistic distribution.

Now we apply the above theory to logistic regression.

5.4.1. *Example*: *Randomized logistic lasso*. Suppose we observe independent samples, $d_i = (y_i, x_i) \overset{\text{i.i.d.}}{\sim} \mathbb{F}$, where y_i's are binary observations and $x_i \in \mathbb{R}^p$. The ordinary logistic regression solves the following problem:

$$\begin{aligned}\bar{\beta} &= \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \ell(\beta) \\ &= \operatorname*{argmin}_{\beta\in\mathbb{R}^p} -\left[\sum_{i=1}^{n} y_i \log \pi(x_i\beta) + (1-y_i)\log\bigl(1-\pi(x_i\beta)\bigr)\right],\end{aligned} \tag{37}$$

where $\pi(x) = \exp(x)/(1+\exp(x))$. This is a nonparametric setting as we do not assume any parametric structure for $\mathbb{F}$.

The randomized logistic lasso adds an ℓ_1 penalty, a randomization term and a small quadratic term,

$$\hat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{\sqrt{n}}\ell(\beta) + \omega^T\beta + \|\Lambda\beta\|_1 + \frac{1}{2\sqrt{n}}\|\beta\|_2^2, \tag{38}$$

where $\omega_j \overset{\text{i.i.d.}}{\sim} \text{Logistic}(\kappa)$ is the perturbation to the gradient and Λ is a diagonal matrix which introduces (possibly) unequal feature weights, κ controls the amount of randomization added. The addition of the quadratic term ensures that (38) is strictly convex, thus has a unique solution. A similar formulation for linear regression has been proposed in Meinshausen and Bühlmann (2010).

Selective inference in this setting has not been considered before. Without the Gaussian assumptions Lee et al. (2016) does not apply. The parametric setting of this problem has been discussed in Fithian, Sun and Taylor (2014), but computation of the selective tests is mostly infeasible for general X. Finally, the asymptotic result by Tian and Taylor (2015) does not apply here as the framework requires exactly affine selection regions, which is not the case in this setting.

Suppose the solution to (38) has nonzero entry set E, then our target of inference β_E^*, the unique population minimizer restricted to E which satisfies

$$\mathbb{E}_{\mathbb{F}}\bigl[X_E^T\bigl(y - \pi\bigl(X_E\beta_E^*\bigr)\bigr)\bigr] = 0. \tag{39}$$

Note that a parametric model $y_i|x_i \sim \text{Bernoulli}(\pi(x_{i,E}\beta_E^*))$ with independently sampled x_i's will have β_E^* satisfying (39). But we by no means assume such an underlying distribution. Rather, for any well-behaved distribution $\mathbb{F}$, β_E^* can be thought of of a statistical functional of the underlying distribution $\mathbb{F}$, depending on the outcome of selection E.

Selective inference in this setting is carried out conditioned on (E, s_E), the active set and its signs. We first introduce the following notation:

$$\begin{aligned}&\pi_E(\beta_E) = \frac{\exp(X_E\beta_E)}{1+\exp(X_E\beta_E)}, \qquad W_E(\beta_E) = \operatorname{diag}\bigl(\pi_E(\beta_E)\bigl(1-\pi_E(\beta_E)\bigr)\bigr), \\ &Q_E(\beta_E) = \frac{1}{n}X_E^T W_E(\beta_E)X_E, \qquad C_E(\beta_E) = \frac{1}{n}X_{-E}^T W_E(\beta_E)X_E, \\ &D_E(\beta_E) = C_E(\beta_E)Q_E^{-1}(\beta_E),\end{aligned}$$

where X is the feature matrix, and X_E, X_{-E} is the columns corresponding to the active set and inactive set respectively. By law of large numbers, we have

$$
\begin{aligned}
&Q_E(\beta_E^*) \xrightarrow{p} \mathbb{E}_{\mathbb{F}} Q_E(\beta_E^*) \stackrel{\text{def}}{=} Q, \qquad C_E(\beta_E^*) \xrightarrow{p} \mathbb{E}_{\mathbb{F}} C_E(\beta_E^*) \stackrel{\text{def}}{=} C, \\
&D_E(\beta_E^*) \xrightarrow{p} C Q^{-1} \stackrel{\text{def}}{=} D.
\end{aligned} \tag{40}
$$

Now we introduce our linearizable statistics and show that the conditioning event (E, s_E) can be expressed as affine regions of these statistics.

LEMMA 14. *Suppose E is the active set of the solution of (38), and we denote*

$$
\bar{\beta}_E = \underset{\beta_E \in \mathbb{R}^E}{\arg\min} - \left[\sum_{i=1}^{n} y_i \log \pi(x_{i,E}\beta_E) + (1 - y_i) \log\bigl(1 - \pi(x_{i,E}\beta_E)\bigr) \right]
$$

as the unpenalized MLE restricted to the selected variables E.

The following statistic T is linearizable with asymptotic mean (β_E^*, ρ) and variance Σ/n,

$$
T = \begin{pmatrix} \bar{\beta}_E \\ \dfrac{1}{n} X_{-E}^T \bigl[y - \pi_E(\bar{\beta}_E) \bigr] \end{pmatrix} + R,
$$

where $R = o_p(n^{-1/2})$ is a small residual, and $\rho = \mathbb{E}[x_{i,-E}^T (y_i - \pi(x_{i,E}\beta_E^*))]$. Moreover, the selection event $\{\hat{E}, z_{\hat{E}} = (E, s_E)\}$ can be characterized as the affine region $\{\sqrt{n} A_M T + B_M \omega \le b_M\}$, where

$$
A_M = \begin{pmatrix} -S_E & 0 \\ 0 & I_{-E} \\ 0 & -I_{-E} \end{pmatrix}, \qquad B_M = \begin{pmatrix} S_E Q^{-1} & 0 \\ D & -I_{-E} \\ -D & I_{-E} \end{pmatrix},
$$

$$
b_M = \begin{pmatrix} -S_E Q^{-1} \Lambda_E s_E \\ \lambda_{-E} - D \Lambda_E s_E \\ \lambda_{-E} + D \Lambda_E s_E \end{pmatrix},
$$

where I_{-E} denotes the identity matrix of $n - |E|$ dimensions and Λ_E, Λ_{-E} denote the active block and the inactive block of Λ respectively, and λ is the diagonal elements of Λ, $S_E = \operatorname{diag}(s_E)$.

The proof of this lemma is also deferred to Section B the Supplementary Material [Tian and Taylor (2018)].

Thus using Lemma 12 and Lemma 13, we can conclude under local alternatives, the pivot (22) converges to Unif(0, 1). To test $H_{0j} : \beta_j^* = 0$, we take $\eta = e_j$, and sample

$$
\eta^T T \mid V_\eta, \sqrt{n} A_M T + B_M \omega \le b_M, \qquad (T, \omega) \sim N\left(\begin{pmatrix} \beta_E^* \\ \rho \end{pmatrix}, \frac{\Sigma}{n} \right) \times G,
$$

where $\rho = \mathbb{E}[x_{i,-E}^T(y_i - \pi(x_{i,E}\beta_E^*))]$. Since ρ is the nuisance parameter for testing H_{0j}, $j \in E$, the conditional law above will not depend on its value. A hit-and-run algorithm for sampling this law can be implemented. Moreover, recent development by Harris et al. (2016), Tian, Bi and Taylor (2016) proposes more general and efficient sampling schemes for this law. For details see, for example, Chapter 3.2 Tian, Bi and Taylor (2016) where the sampling scheme for this very example is considered and simulation results are provided.

In Lemma 14, we assume the covariance matrix Σ is known. In applications, we can bootstrap it. But is it valid to plug in the bootstrap estimate of Σ?

5.5. *Plugging in variance estimates.* In Section 5.3, we derived quantities that were asymptotically pivotal for the best median, up to an unknown variance. In the sample median case, by (35), the variance of the sample median is approximately $[4nf(m)^2]^{-1}$, where $f(m)$ is the PDF evaluated at the median m. A simple consistent estimator for $f(m)$ (in probability) is to take $1/2 \pm \frac{1}{\sqrt{n}}$ quantiles a_n and b_n, then

$$f(m) \approx \frac{2}{\sqrt{n}(b_n - a_n)} \tag{41}$$

is consistent (in probability) for $f(m)$ based on which we get a consistent estimator for σ_1^2.

More generally, computing the pivot (22) requires knowledge of Σ. In practice, we usually do not have prior knowledge of the variance Σ and need a consistent estimate for Σ. We might use a bootstrap or jackknife estimator. When p is fixed, the bootstrap estimator is consistent and thus we get a consistent estimator $\hat{\Sigma}_n$, that is

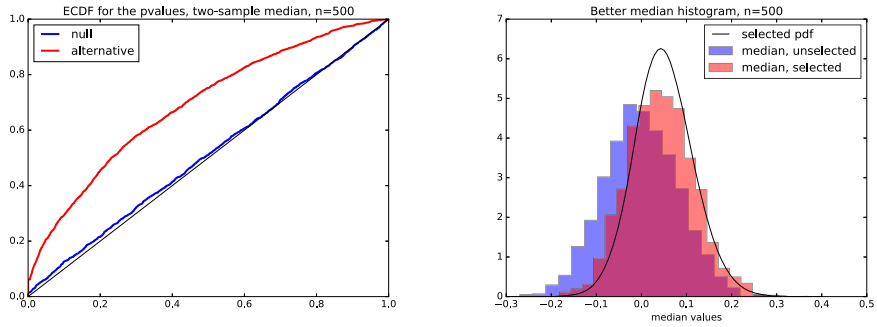
$$\hat{\Sigma}_n \xrightarrow{\mathbb{F}} \Sigma \text{ in probability.}$$

Lemma 3 states that under moment conditions on the likelihood, $\hat{\Sigma}_n$ will be consistent for Σ under the selective distribution $\mathbb{F}^*$ as well. Namely, if we randomize before selection,

$$\hat{\Sigma}_n \xrightarrow{\mathbb{F}^*} \Sigma \text{ in probability.}$$

This justifies the plug-in estimator for Σ when computing the pivot (22). More technical details can be found in Section E in the Supplementary Material [Tian and Taylor (2018)].

Figure 4 is some simulation results for the two-sample medians problem. In each case, we take the sample size for each treatment group to be 500, and generate the noise from a skewed distribution $N(0, 1) + 0.5\,\mathrm{Exp}(1)$. We standardize it such that the noise has median 0 under the null hypothesis. We use additive logistic noise with scale $\kappa = 0.5$ for randomization. The better group is decided using the randomized sample median, and selective inference is carried out. In Figure 4(a),



(a) Null and alternative pivot for the "better" median (b) Selective v.s. unselective distribution, and theoretical PDF

FIG. 4. *Asymptotic distribution of the median for the selected group.*

the pivot with plug-in variance estimate $\widehat{\sigma}$ in (41) is plotted under both the null hypothesis $H_0 : \mu_{\text{better}} = 0$ and the $H_A : \mu_{\text{better}} > \frac{1}{\sqrt{n}}$. The pivot has reasonable power even for identifying local alternatives. The pivot is almost exactly Unif(0, 1) under the null hypothesis with the sample size $n = 500$. In fact, it is very close at a relatively small sample size $n = 50$ justifying the application of asymptotics in the nonparametric setting. Figure 4(b) further illustrates the difference in the unselective v.s. selective distribution and its convergence to its theoretical limit. We see that there is a clear shift in selective distribution that calls for adjustment for the selection. For sample size $n = 500$, the empirical selective distribution converges to our theoretical distribution.

6. Multiple randomizations of the data. Most of the examples above focus on a single randomization ω on the data, which we use for model selection. We naturally want to extend it to multiple randomizations, and multiple randomized selections, which will collectively suggest a model for inference. In this section, we allow multiple randomizations in a possibly sequential fashion and discuss how inference can be carried out.

6.1. *Selective inference after cross-validation.* Consider the case where we first choose a regularization parameter by cross-validation, and then fit the square-root LASSO problem [Belloni, Chernozhukov and Wang (2011), Sun and Zhang (2012)] at this parameter,

$$\hat{\beta}_\lambda(y; X) = \arg\min_\beta \|y - X\beta\|_2 + \lambda\|\beta\|_1, \tag{42}$$

where λ is picked from a fixed grid $\Lambda = [\lambda_1, \ldots, \lambda_k]$. The discussion below is not specific to selection by square-root LASSO.

The model selected by cross-validated square-root LASSO involves two steps of selection. We denote by y_{CV} the response for selecting the randomization parameter, and y_{select} the response vector for fitting the square-root LASSO at the selected regularization parameter λ. Both vectors are randomized version of the original vector y. Inference after cross validation requires combining two steps of randomized selection. Consider the following procedure.

First, we randomize y to get the vector y_{CV} and y_{select}

$$\begin{aligned} y_{\mathrm{inter}}|y, X &\sim N\bigl(y, \sigma_1^2 I\bigr), \\ y_{\mathrm{CV}}|y_{\mathrm{inter}}, y, X &\sim N\bigl(y_{\mathrm{inter}}, \sigma_{2,\mathrm{CV}}^2 I\bigr), \\ y_{\mathrm{select}}|y_{\mathrm{inter}}, y, X &\sim N\bigl(y_{\mathrm{inter}}, \sigma_{2,\mathrm{select}}^2 I\bigr). \end{aligned} \tag{43}$$

Note the intermediate vector y_{inter} is introduced for convenience of sampling. The above is just one of the plausible randomization schemes.

After having randomized, we select λ with K-fold cross-validation using y_{CV}:

$$\hat{\lambda} = \hat{\lambda}(y_{\mathrm{CV}}, X) = \operatorname*{argmin}_{\lambda \in \Lambda} \mathrm{CV}_K(y_{\mathrm{CV}}, X, \lambda), \tag{44}$$

where $\mathrm{CV}_K(y, X, \lambda)$ is the usual K-fold cross-validation score with coefficients estimated by the square-root LASSO. Alternatively, one could compute the cross-validation score using the OLS estimators of the selected variables. Note that we have left implicit the randomization that splits observations into groups. That is $\hat{\lambda}$ in (44) above is a function of $(y_{\mathrm{CV}}, X, \omega)$ where ω is a random partition of $\{1, \ldots, n\}$ into K groups. When we sample y_{CV} below, we redraw ω each time.

The subset of variables and signs is selected using the square-root LASSO with response y_{select}:

$$\begin{aligned} \hat{E}(y_{\mathrm{CV}}, y_{\mathrm{select}}, X) &= \{j : \hat{\beta}_{\hat{\lambda}(y_{\mathrm{CV}}, X), j} \neq 0\}, \\ z_{\hat{E}}(y_{\mathrm{CV}}, y_{\mathrm{select}}, X) &= \operatorname{sign}(\hat{\beta}_{\hat{\lambda}(y_{\mathrm{CV}}, X)}). \end{aligned} \tag{45}$$

After seeing the selected variables $\hat{E}$, we perform inference in the selected model $M_{\mathrm{sel}}(\hat{E})$. Since $M_{\mathrm{sel}}(\hat{E})$ is an exponential family, we will still have an exponential family after selection. Per Lemma 2, we sample from the following law:

$$\mathcal{L}\bigl(X_j^T y|\hat{\lambda}(y_{\mathrm{CV}}, X) = \lambda, (\hat{E}, z_{\hat{E}}) = (E, z_E), P_{E\setminus j} y\bigr). \tag{46}$$

The additional conditioning on the signs is for computational reasons. In fact, recent development in Harris et al. (2016) proposes sampling schemes that overcome these difficulties, so that we do not need to condition on this additional information.

A detailed sampling scheme for (46) is included in Section A in the Supplementary Material [Tian and Taylor (2018)].

6.2. *Collaborative selective inference.* One of the motivations of the reusable holdout described in Dwork et al. (2015) is that it allows a data analyst to repeatedly query a database yet still be able to approximately estimate expectations even after asking many questions about the data. Another version of this model may be that several groups wish to model the same data and then, as a consortium, decide on a final model and be able to approximately estimate expectations in this final model. We might call this *collaborative selective inference*.

Formally, suppose each of L groups has its own preferred method of model selection, encoded as selection procedures $(\widehat{\mathcal{Q}}_l)_{1\le l\le L}$. We assume there is a central "data" bank that decides what "data" each group is allowed to see. We express this is as a sequence of randomization schemes $(y_l^*)_{1\le l\le L}$. Formally, this is equivalent to enlarging the probability space to $\mathcal{D}\times\mathcal{B}$ with measure $\mathbb{F}\times\mathbb{B}$ and fixing a function $y^*(y,\omega)=(y_1^*(y,\omega),\dots,y_l^*(y,\omega))$. It may be desirable to choose the law of $y^*|y$ so that the coordinates are conditionally independent given y, though it is not necessary.

Now suppose that the L groups choose models $\hat{M}_l^*=\widehat{\mathcal{Q}}_l(y_l^*)\in\sigma(y_l^*)$ and convene to discuss what the best model is M. For every choice of L models $(M_1,\dots,M_L)$ and final model M, the following selective distribution can be used for valid selective inference

$$\frac{d\mathbb{F}^*}{d\mathbb{F}}(y)=\frac{\mathbb{B}(\omega:\bigcap_{l=1}^L\widehat{\mathcal{Q}}_l(y_l^*(y,\omega))=M_l)}{(\mathbb{F}\times\mathbb{B})(\bigcap_{l=1}^L\widehat{\mathcal{Q}}_l^*=M_l)}. \tag{47}$$

When the y_l^*'s are conditionally independent given y then it is clear that

$$\mathbb{B}\left(\bigcap_{l=1}^L\widehat{\mathcal{Q}}_l\big(y^*(y,\omega)\big)=M_l\right)=\prod_{l=1}^L\mathbb{B}\big(\widehat{\mathcal{Q}}_l\big(y_l^*(y,\omega)\big)=M_l\big).$$

It is possible that the consortium has beforehand decided on an algorithm that will choose a best model automatically, determined by some function $\mathcal{S}(M_1,\dots,M_L)$. In this case, one should use the selective distribution

$$\frac{d\mathbb{F}^*}{d\mathbb{F}}(y)=\frac{\mathbb{B}(\omega:\mathcal{S}(M_1^*(y,\omega),\dots,M_L^*(y,\omega))=M)}{(\mathbb{F}\times\mathbb{B})(\mathcal{S}(M_1^*,\dots,M_L^*)=M)}. \tag{48}$$

When the models in question are parametric, perhaps Gaussian distributions, and the randomization is additive Gaussian noise the central data bank can explicitly lower bound the leftover information by

$$\operatorname{Var}\big(y|y_1^*,\dots,y_L^*\big).$$

This quantity is expressible in terms of the marginal variance of y and the central data bank's noise generating distribution for $y^*(y,\omega)=(y+\omega_1,\dots,y+\omega_L)$. By maintaining a lower bound on the above quantity, the central data bank can maintain a minimum prescribed information in the data for final estimation and/or inference. In a sequential setting, where valid inference is desired at each step,

maintaining a lower bound may involve releasing noisier and noisier versions of y. Sampling under this scheme seems quite difficult, and we leave it as an area of interesting future research.

SUPPLEMENTARY MATERIAL

Supplement to "Selective inference with a randomized response" (DOI: 10.1214/17-AOS1564SUPP; .pdf). We provide additional sampling schemes, technical details for plugin variance estimators and proofs for all the theorems and lemmas in the supplementary materials.

REFERENCES

BAHADUR, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580. MR0189095

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and STARK, P. B. (1996). Nonequivariant simultaneous confidence intervals less likely to contain zero. *J. Amer. Statist. Assoc.* **91** 329–337. MR1394088

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

CHATTERJEE, S. (2005). A simple invariance theorem. Preprint. Available at arXiv:math/0508213.

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189

DWORK, C., FELDMAN, V., HARDT, M., PITASSI, T., REINGOLD, O. and ROTH, A. (2015). Preserving statistical validity in adaptive data analysis [extended abstract]. In *STOC'*15—*Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 117–126. ACM, New York. MR3388189

FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Available at arXiv:1410.2597.

FITHIAN, W., TAYLOR, J., TIBSHIRANI, R. and TIBSHIRANI, R., (2015). Selective sequential model selection. Available at arXiv:1512.02565.

GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724–739. MR1106283

HARRIS, X. T. (2016). Prediction error after model search. Preprint. Available at arXiv:1610.06107.

HARRIS, X. T., PANIGRAHI, S., MARKOVIC, J., BI, N. and TAYLOR, J. (2016). Selective sampling after solving a convex problem. Preprint. Available at arXiv:1609.05609.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York. MR0852406

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. MR2758523

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584

PAKMAN, A. and PANINSKI, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *J. Comput. Graph. Statist.* **23** 518–542. MR3215823

ROSENTHAL, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* **86** 638.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

TIAN, X., BI, N. and TAYLOR, J. (2016). Magic: A general, powerful and tractable method for selective inference. Preprint. Available at arXiv:1607.02630.

TIAN, X., LOFTUS, J. R. and TAYLOR, J. E. (2015). Selective inference with unknown variance via the square-root lasso. Preprint. Available at arXiv:1504.08031.

TIAN, X. and TAYLOR, J. (2015). Asymptotics of selective inference. Available at arXiv:1501.03588.

TIAN, X. and TAYLOR, J. (2018). Supplement to "Selective inference with a randomized response." DOI:10.1214/17-AOS1564SUPP.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. Preprint. Available at arXiv:1506.06266.

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689

TUKEY, J. W. (1980). We need both exploratory and confirmatory. *Amer. Statist.* **34** 23–25.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: xtian@stanford.edu
jonathan.taylor@stanford.edu