Optimal Inference After Model Selection

William Fithian* Dennis Sun Jonathan Taylor

Stanford University, Department of Statistics

October 10, 2014

Abstract

To perform inference after model selection, we propose controlling the *selective type I error*; i.e., the error rate of a test given that it was performed. By doing so, we recover long-run frequency properties among selected hypotheses analogous to those that apply in the classical (non-adaptive) context. Our proposal is closely related to data splitting and has a similar intuitive justification, but is more powerful. Exploiting the classical theory of Lehmann and Scheffé (1955), we derive most powerful unbiased selective tests and confidence intervals for inference in exponential family models after arbitrary selection procedures. For linear regression, we derive new selective z-tests that generalize recent proposals for inference after model selection and improve on their power, and new selective t-tests that do not require knowledge of the error variance σ^2.

1 Introduction

A typical statistical investigation can be thought of as consisting of two stages:

1. **Selection:** The analyst chooses a probabilistic model for the data at hand, and formulates testing, estimation, or other problems in terms of unknown aspects of that model.

2. **Inference:** The analyst attempts the chosen problems using the data and the selected model.

Informally, the selection stage determines what questions to ask, and the inference stage answers those questions. In some cases, it is possible to specify the question prior to collecting the data, for example if the data are governed by some known physical law. However, in most applications, the choice of question is at least partially guided by the data. For example, we often perform exploratory analyses to decide which predictors or interactions to include in a regression model or to check that the assumptions of a test are satisfied.

If we do not account properly for model selection, the resulting inferences can have troubling frequency properties, as we now illustrate with an example.

Example 1 (File Drawer Effect)**.** Suppose a scientist observes n independent observations $Y_i \sim N(\mu_i, 1)$. He focuses only on the apparently large effects, i.e.,

$$\widehat{I} = \{i : |Y_i| > 1\}.$$

He wishes to test $H_{0,i} : \mu_i = 0$ for each $i \in \widehat{I}$ at the $\alpha = 0.05$ significance level. Most scientists intuitively recognize that the nominal test that rejects $H_{0,i}$ when $|Y_i| > 1.96$ is invalidated by the selection.

*To whom correspondence should be addressed

What exactly is "invalid" about this test? After all, the probability of falsely rejecting a given $H_{0,i}$ is still α, since most of the time, $H_{0,i}$ is simply not tested at all. Rather, the troubling feature is that the error rate among the hypotheses we *do* test is possibly much higher than α. To be precise, let n_0 be the number of true null effects and suppose $n_0 \to \infty$ as $n \to \infty$. Then, in the long run, the fraction of errors among the true nulls we test is

$$\begin{aligned}\frac{\#\text{ false rejections}}{\#\text{ true nulls selected}} &= \frac{\frac{1}{n_0}\sum_{i:\, H_{0,i}\text{ true}} 1\{i \in \widehat{I},\ \text{reject } H_{0,i}\}}{\frac{1}{n_0}\sum_{i:\, H_{0,i}\text{ true}} 1\{i \in \widehat{I}\}} \\ &\to \frac{\mathbb{P}_{H_{0,i}}(i \in \widehat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \widehat{I})} \\ &= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \widehat{I}), \end{aligned} \tag{1}$$

which for the nominal test is $\Phi(-1.96)/\Phi(-1) \approx .16$.

Thus, we see that (1), the probability of a false rejection conditional on selection, is a natural goal of inference when selection is involved. This paper will develop a theory for inference after selection, or *selective inference*, based on selective error control. Our guiding principle is:

The answer must be valid, given that the question was asked.

For all its disarming simplicity, Example 1 can be regarded as a stylized model of science. Imagine that each Y_i represents an estimated effect size from a scientific study. However, only the large estimates are ever published—a caricature which may not be too far from the truth, as recently demonstrated by Franco et al. (2014). Because of this selection bias, the error rate among published claims may be very high, leading even to speculation that "most published research findings are false" (Ioannidis, 2005). Thus, selection effects may be a partial explanation for the replicability crisis and decline effect reported on in the scientific community (Yong, 2012) and the popular media (Johnson, 2014).

1.1 Conditioning on Selection

In this article, we will argue for controlling the *selective type I error rate* (1). Before delving into selective inference, we first review some premises of classical inference.

In classical inference, the notion of "inference after selection" does not exist. The analyst must specify the model in advance of looking at the data. While this purist view avoids the selection problem altogether, it does not realistically describe most statistical practice: statisticians are trained to check their models and to tweak them if they diagnose a problem. Model checking is technically forbidden, since it leaves open the possibility that the model will change after we see the data. Under this view, a level-α test for a hypothesis H_0 under model M must control the nominal type I error:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0) \leq \alpha. \tag{2}$$

The subscript in (2) reminds us that the probability is computed under the assumption that the data Y are generated from model M, and H_0 is true; if H_0 is false, or M is misspecified, the rejection probability may be much higher.

However, one can argue that models and hypotheses are almost never truly fixed but are chosen randomly, since they are ultimately contingent upon previous experimental outcomes in the (random) scientific process. Under this view, a test should control the selective type I error

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) \leq \alpha. \tag{3}$$

In practice, we simply ignore selection and use classical tests that control (2), implicitly assuming that the randomness in the selection is independent of the data used for inference, i.e.,

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) = \mathbb{P}_{M,H_0}(\text{reject } H_0). \tag{4}$$

Although the question of whether models are fixed or random may seem like a philosophical quibble, the random viewpoint gives us a prescription for what to do when science does not dictate a model. If it is possible to split the data $Y = (Y_1, Y_2)$ with Y_1 independent of Y_2, then we can imitate the scientific process by setting aside Y_1 for selection and Y_2 for inference. If selection depends on Y_1 only, then any nominal level-α test based on the value of Y_2 will satisfy (4), so the nominal test based on Y_2 also controls the selective error (3).

This meta-algorithm for generating selective procedures from nominal ones is called *data splitting* or *sample splitting*. The idea dates back at least as far as Cox (1975), and, despite the paucity of literature on the topic, is common wisdom among practitioners. For example, it is customary in genetics to use one cohort to identify loci of interest and a separate cohort to confirm them (Sladek et al., 2007).

The popularity of data splitting owes in no small part to its transparent justification, which non-experts can easily appreciate: if we imagine that Y_1 is observed "first," then we can proceed to analyze Y_2 as though model selection took place "ahead of time." Equation (4) guarantees that this temporal metaphor will not lead us astray even if it has no basis in the physical reality of how Y_1 and Y_2 were actually collected.

Data splitting elegantly solves the problem of controlling selective error, but at a cost. It not only reduces the data available for inference, but also reduces the data available for selection. Furthermore, it is not always possible; for example, spatial and time series data often exhibit autocorrelation that rules out splitting the data into independent parts.

In this article, we propose directly controlling the selective error rate (3) by conditioning on the event that (M, H_0) is selected. As with data splitting, we treat the data as though it were revealed in stages: in the first stage, we "observe" just enough data to resolve the decision of whether to test (M, H_0), after which we can treat the data $(Y \mid (M, H_0) \text{ selected})$ as "not yet observed" when stage two commences.

The intuition of the above paragraph can be expressed formally in terms of the filtration

$$\mathscr{F}_0 \underbrace{\subseteq}_{\text{used for selection}} \mathscr{F}(\mathbf{1}_A(Y)) \underbrace{\subseteq}_{\text{used for inference}} \mathscr{F}(Y), \tag{5}$$

where $\mathscr{F}(Z)$ denotes the σ-algebra generated by random variable Z (informally, everything we know about the data after observing Z), $\mathscr{F}_0$ is the trivial σ-algebra (representing complete ignorance), and A is the *selection event* $\{(M, H_0) \text{ selected}\}$. We can think of "time" progressing from left to right in (5). In stage one, we learn just enough to decide whether to test (M, H_0), and no more, advancing our state of knowledge from $\mathscr{F}_0$ to $\mathscr{F}(\mathbf{1}_A(Y))$. We then begin stage two, in which we discover the actual value of Y, advancing our knowledge to $\mathscr{F}(Y)$. Because our selection decision is made at the end of stage one, everything revealed during stage two is fair game for inference.

Controlling the type I error conditional on A in effect prevents us from appealing to the fact that $Y \in A$ as evidence against H_0. Even if $Y \in A$ is extremely surprising under H_0, we still will not reject unless we are surprised anew in the second stage. In this sense, conditioning on a random variable discards the information it carries about any parameter or hypothesis of interest. By contrast with data splitting, which can be viewed as conditioning on Y_1 instead of $\mathbf{1}_A(Y_1)$, we advocate discarding as little data as possible and reserving the rest for stage two. This frugality results in a more efficient division of the information carried by Y — *data carving*, to introduce an evocative metaphor.

Although data splitting is impossible in Example 1, where there is only one observation Y_i per hypothesis $H_{0,i}$, it is a simple matter to control the selective error (3) directly:

Example 1 (continued)**.** We condition on the selection event

$$\{i \in \widehat{I}\} = \{|Y_i| > 1\}.$$

To control the *selective type I error* at 0.05, we find the critical value c solving

$$\mathbb{P}_{H_{0,i}}\left(|Y_i| > c \;\middle|\; |Y_i| > 1\right) = 0.05.$$

In this case $c = 2.41$, which is more stringent than the nominal 1.96 cutoff.

1.2 Outline

In Section 2 we formalize the problem of selective inference, discuss general properties of selective error control, and address key conceptual questions. Conditioning on the selection event effectively discards the information used for selection, but some information is left over for second-stage inference. We will also see that a major advantage of selective error control is that it allows us to consider only one model at a time when designing tests and intervals, even if *a priori* there are many models under consideration.

If $\mathcal{L}(Y)$, the law of random variable Y, follows an exponential family model, then for any event A, $\mathcal{L}(Y \,|\, A)$ follows a closely related exponential family model. As a result, selective inference dovetails naturally with the classical optimality theory of Lehmann and Scheffé (1955); Section 3 briefly reviews this theory and derives most powerful unbiased selective tests in arbitrary exponential family models after arbitrary model selection procedures. Because conditioning on more data than is necessary saps the power of second-stage tests, data splitting yields inadmissible selective tests under general conditions.

Section 4 gives some general strategies for computing rejection cutoffs for the tests prescribed in Section 3, while Sections 5–6 derive selective tests in specific examples. Section 5 focuses on the case of linear regression, generalizing the recent proposals of Taylor et al. (2014), Lee et al. (2013), and others. We derive new, more powerful selective z-tests, as well as t-tests that do not require knowledge of the error variance σ^2.

Several simulations in Section 7 compare the post-lasso selective z-test with data splitting, and illustrate a *selection–inference tradeoff*, between using more data in the initial stage and reserving more information for the second stage. Section 8 compares and contrasts selective inference with multiple inference, and Section 9 concludes.

2 The Problem of Selective Inference

2.1 Example: Regression and the Lasso

In the previous section, we motivated the idea of conditioning on selection. Arguably, the most familiar example of this "selection" is variable selection in linear regression. In regression, the observed data $Y \in \mathbb{R}^n$ is assumed to be generated from a multivariate normal distribution

$$Y \sim N_n(\mu,\ \sigma^2 I_n). \tag{6}$$

The goal is to model the mean μ as a linear function of predictors X_j, $j = 1, \ldots, p$. To obtain a more parsimonious model (or simply an identifiable model when $p > n$), researchers will often use only a subset $M \subseteq \{1, \ldots, p\}$ of the predictors. Each subset M leads to a different probabilistic model

$$\mu = X_M \beta^M, \tag{7}$$

where X_M denotes the matrix consisting of columns X_j for $j \in M$. Then, for the model that we selected, it is customary to report tests of $H_0 : \beta_j^M = 0$ for each coefficient in the model. If M

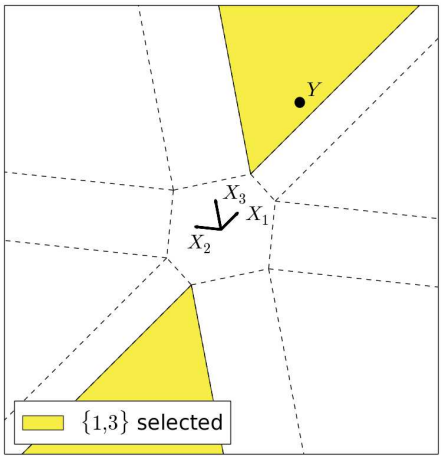


Figure 1: An example of the lasso with $n = 2$ observations and $p = 3$ variables. We base tests on the distribution of Y, conditional on its landing in the highlighted region.

was chosen in a data-dependent way, then to control selective error we must condition on having selected (M, H_0), which in this case is the same as conditioning on having selected model M.

There are many data-driven methods for variable selection in linear regression, ranging from AIC minimization to forward stepwise selection, cf. Hastie et al. (2009). We will consider one procedure in particular, based on the lasso, mostly because selective inference in the context of the lasso (Lee et al., 2013) was a main motivation for the present work. The lasso (Tibshirani, 1996) provides an estimate of $\beta \in \mathbb{R}^p$ that solves

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; ||Y - X\beta||_2^2 + \lambda||\beta||_1, \tag{8}$$

where X is the "full" matrix consisting of all p predictors. The first term is the usual least-squares objective, while the second term encourages many of the coefficients to be exactly zero. Because of this property, it makes sense to define the model "selected" by the lasso to be the set of variables with non-zero coefficients, i.e.,

$$\widehat{M}(Y) = \{j : \hat{\beta}_j \neq 0\}.$$

Notice that $\widehat{M}(Y)$ can take on up to 2^p possible values, one for each possible subset of $\{1, ..., p\}$, so the inverse image of $\widehat{M}$ partitions $\mathbb{R}^n$ into regions that correspond to each possible model. To control the selective error, we must condition on the event that Y landed in the region corresponding to the model we chose. The partition for a lasso problem with $p = 3$ variables in $n = 2$ dimensions is shown in Figure 1. An explicit characterization of the lasso partition can be found in Lee et al. (2013). If we used a different selection procedure, we would obtain a different partitions. Characterizations of the partitions in forward stepwise selection and marginal screening can be found in Loftus and Taylor (2014) and Lee and Taylor (2014), respectively.

Let us imagine that in stage one, we loaded the data into a software package and computed $\widehat{M}(Y)$, but we remain otherwise ignorant of the value Y — that is, we have observed *which* of the regions Y falls into but not *where* Y is in that region. Now that we have chosen the model, we will construct tests of $H_0 : \beta_j^M = 0$ for each of the selected variables. In the example shown in Figure

1, we selected variables 1 and 3 and thus test the two hypotheses

$$H_{0,1}^{\{1,3\}} : \beta_1^{\{1,3\}} = 0$$
$$H_{0,3}^{\{1,3\}} : \beta_3^{\{1,3\}} = 0.$$

Notice that we have to be careful to always specify the model along with the coefficient, since the coefficient for variable j does not necessarily have a consistent interpretation across different models. Each regression coefficient summarizes the effect of that variable, adjusting for the other variables in the model. For example, "What is the effect of IQ on salary?" is a genuinely different question from "What is the effect of IQ on salary, after adjusting for years of education?" Both are meaningful and interesting in their own right, but we should not confuse the two.[1]

Having chosen the model M and conditioned on the selection, we will base our tests on the precise location of Y, which we do not know yet. Conditionally, Y is not Gaussian, but it does follow an exponential family. As a result, we can appeal to the classical theory of Lehmann and Scheffé (1955) to construct tests or confidence intervals for its natural parameters, which are β^M if σ^2 is known, and otherwise are $(\beta^M/\sigma^2, 1/\sigma^2)$.

With this concrete example in mind, we will now develop a general framework of selective inference that is much more broadly applicable. Because we are allowing models and hypotheses to be random, it is necessary to carefully define our inferential goals. We first discuss selective inference in the context of hypothesis testing. The closely related developments for confidence intervals will follow in Section 2.3.

2.2 Selective Hypothesis Tests

Now suppose that our data Y lies in some measurable space $(\mathcal{Y}, \mathscr{F})$, with unknown sampling distribution $Y \sim F$. The analyst's task is to pose a reasonable probability model M — i.e., a family of distributions which she believes contains F — and then carry out inference based on the observation Y.

Let $\mathcal{Q}$ denote the *question space* of inference problems q we might tackle. In the case of regression where we test all the variables in the model, the question space would be

$$\mathcal{Q} = \{(M, H_{0,j}^M) : M \subseteq \{1, \ldots, p\}, j \in M\}.$$

To avoid measurability issues, we will assume that $\mathcal{Q}$ is countable, although the setting can be generalized to other question spaces with additional care.

A hypothesis testing problem is a pair $q = (M, H_0)$ of a model M and null hypothesis H_0, by which we mean a submodel $H_0 \subseteq M$.[2] Without loss of generality, we assume H_0 is tested against the alternative hypothesis $H_1 = M \setminus H_0$. We model selective inference as a process with two distinct stages:

1. **Selection:** From the collection $\mathcal{Q}$ of possible questions, the analyst selects a subset $\widehat{\mathcal{Q}}(Y) \subseteq \mathcal{Q}$ to test, based on the data.

2. **Inference:** The analyst performs a hypothesis test for each $q = (M, H_0) \in \widehat{\mathcal{Q}}$.

In the case of the simple regression example shown in Figure 1, where we selected variables 1 and 3, $\widehat{\mathcal{Q}}$ would consist of the hypotheses for each of the two variables in the model. To be completely explicit,

$$\widehat{\mathcal{Q}}(Y) = \left\{ \left(\{1,3\}, H_{0,1}^{\{1,3\}} \right), \left(\{1,3\}, H_{0,3}^{\{1,3\}} \right) \right\}.$$

[1] We use the word "effect" here informally to refer to a regression coefficient, recognizing that regression cannot establish causal claims on its own.

[2] We identify a "null hypothesis" like $H_0 : \mu(F) = 0$ with the corresponding subfamily or "null model" $\{F \in M : \mu(F) = 0\}$. This should remind us that the error guarantees of a test do not necessarily extend beyond the model it was designed for.

A correctly specified model M is one that contains the true sampling distribution F. Importantly, we expressly do not assume that all, or any, of the candidate models are correctly specified. Because the analyst must choose M without knowing F, she could choose poorly, in which case there may be no formal guarantees on the behavior of the test she performs in stage two. Misspecification is possible in nearly every real statistical application whether models are specified adaptively or non-adaptively; see Section 2.5.2 for further discussion of this issue.

For our purposes, a *hypothesis test* is a function $\phi(y)$ taking values in $[0, 1]$, representing the probability of rejecting H_0 if $Y = y$. In most cases, the value of the function will be either 0 or 1, but with discrete variables, randomization may be necessary to achieve exact level α.

To adjust for selection, we condition on the event that the question was asked, which we describe by the selection event

$$A_q = \{q \in \widehat{\mathcal{Q}}(Y)\}, \tag{9}$$

i.e., the event that q is among the questions asked. In the regression example, we only ever test $H_{0,j}^M$ when model M is selected, so conditioning on A_q is equivalent to simply conditioning on M.

In selective inference, we are mainly interested in the properties of a test ϕ_q for a question q, conditional on A_q. We say that ϕ_q controls *selective type I error* at level α if

$$\mathbb{E}_F\left[\phi_q(Y) \,|\, A_q\right] \leq \alpha, \quad \text{for all } F \in H_0. \tag{10}$$

and define its *selective power function* as

$$\text{Pow}_{\phi_q}(F \,|\, A_q) = \mathbb{E}_F[\phi_q(Y) \,|\, A_q]. \tag{11}$$

Because $\mathcal{Q}$ is countable, the only relevant q are those for which $\mathbb{P}(A_q) > 0$.

Notice that only the model M and hypothesis H_0 are relevant for defining the selective level of a test. This means that in designing valid ϕ_q, we can concentrate on one q at a time, even if there are many mutually incompatible candidate models in $\mathcal{Q}$. As long as each ϕ_q controls the selective error at level α given its selection event A_q, then a global error is also controlled:

$$\frac{\mathbb{E}\left[\#\text{ false rejections}\right]}{\mathbb{E}\left[\#\text{ true nulls selected}\right]} \leq \alpha, \tag{12}$$

provided that the denominator is finite. (12) holds for countable $\mathcal{Q}$ regardless of the dependence structure across different q. The fact that we can design tests one q at a time makes it much easier to devise selective tests in concrete examples, which we take up in Sections 3–6.

Suppose that each scientist in a discipline controls the selective error rate for each of his or her own experiments. Then the discipline as a whole will achieve long-run control of the type I error rate among true *selected* null hypotheses, in the same sense as they would if there were no selection. No coordination is required between different research groups.

Proposition 1 (Discipline-Wide Error Control)**.** *Suppose there are n independently operating research groups in a scientific discipline with a shared, countable question space $\mathcal{Q}$. Research group i collects data $Y_i \sim F_i$, applies selection rule $\widehat{\mathcal{Q}}_i(Y_i) \subseteq \mathcal{Q}$, and carries out selective level-α tests $(\phi_{q,i}(y_i), q \in \widehat{\mathcal{Q}}_i)$. Assume each research group has probability at least $\delta > 0$ of carrying out at least one test of a true null, and for some common $B < \infty$,*

$$\mathbb{E}_{F_i}\left[\left|\widehat{\mathcal{Q}}_i(Y_i)\right|^2\right] \leq B, \quad \textit{for all } i.$$

Then as n grows, the discipline as a whole achieves long-run control over the frequentist error rate

$$\limsup_{n\to\infty} \frac{\#\text{ false rejections}}{\#\text{ true nulls selected}} \overset{a.s.}{\leq} \alpha. \tag{13}$$

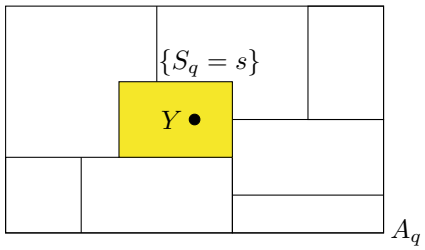


Figure 2: Instead of conditioning on the selection event A_q that question q is asked, we can condition on a finer event, the value of the random variable S_q. We call S_q the *selection variable*.

The proof is deferred to Appendix A. Though the assumption of independence across research groups can be weakened without necessarily affecting the conclusion, we do not pursue such generalizations. Note that there is no counterpart to Proposition 1 for multiple-inference error rates such as the false discovery rate (FDR) (Benjamini and Hochberg, 1995) or familywise error rate (FWER); even if every research group controls its own FWER or FDR at level α, there is no guarantee we will control FWER or FDR after aggregating across the different groups.

2.3 Selective Confidence Intervals

If the goal is instead to form confidence intervals for a parameter $\theta(F)$, it is more convenient to think of $\mathcal{Q}$ as containing pairs $q = (M, \theta(\cdot))$ of a model and a parameter. By analogy to (10), we will call a set $C(Y)$ a $(1-\alpha)$ *selective confidence set* if

$$\mathbb{P}_F(\theta(F) \in C(Y) \,|\, A_q) \geq 1 - \alpha, \quad \text{for all } F \in M. \tag{14}$$

The next result establishes that selective confidence sets can be obtained by inverting selective tests, as one would expect by analogy to the classical case.

Proposition 2 (Duality of Selective Tests and Confidence Sets)**.** *Suppose we form a confidence interval for $\theta(F)$ on the event A_q. Suppose also that on this event, we form a test ϕ_t of $H_{0,t} = \{F : \theta(F) = t\}$ for all t. Let $C(Y)$ be the set of t for which ϕ_t does not (always) reject:*

$$C(Y) = \{t : \phi_t(Y) < 1\}. \tag{15}$$

If each ϕ_t is a selective level-α test, then $C(Y)$ is a selective $(1-\alpha)$ confidence set.

Proof. The selective non-coverage probability is

$$\mathbb{P}_F(\theta(F) \notin C(Y) \,|\, A_q) \;=\; \mathbb{P}_F(\phi_{\theta(F)}(Y) = 1 \,|\, A_q) \;\leq\; \mathbb{E}_F\left[\phi_{\theta(F)}(Y) \,|\, A_q\right] \;\leq\; \alpha.$$

□

2.4 Conditioning Discards Information

Because performing inference conditional on a random variable effectively disqualifies that variable as evidence against a hypothesis, we will typically want to condition on as little data as possible in stage two. Even so, some selective inference procedures condition on more than A_q. For example, data splitting can be viewed as inference conditional on Y_1, the part of the data used for selection. More generally, we say a *selection variable* is any variable $S_q(Y)$ whose level sets partition the

sample space more finely than A_q; i.e., $A_q \in \mathscr{F}(S_q)$. Informally, we can think of conditioning on a finer partition of A_q, as shown in Figure 2.

We say ϕ controls the *selective type I error with respect to* S_q at level α if the error rate is less than α given $S_q = s$ for $\{S_q = s\} \subseteq A_q$. More formally,

$$\mathbb{E}_F\left[\phi(Y)\mathbf{1}_{A_q}(Y) \,|\, S_q\right] \overset{\text{a.s.}}{\leq} \alpha, \quad \text{for all } F \in H_0 \tag{16}$$

Taking $S_q(y) = \mathbf{1}_{A_q}(y)$, the coarsest possible selection variable, recovers the baseline selective type I error in (10). The definition of a selective confidence set may be generalized in the same way.

Generalizing (5) to finer selection variables gives

$$\mathscr{F}_0 \underbrace{\subseteq}_{\text{used for selection}} \mathscr{F}(S(Y)) \underbrace{\subseteq}_{\text{used for inference}} \mathscr{F}(Y), \tag{17}$$

suggesting that the more we refine $S(Y)$, the less data we have left for second-stage inference. Indeed, the finer S is, the more stringent is the requirement (16):

Proposition 3 (Monotonicity of Selective Error)**.** *Suppose $\mathscr{F}(S_1) \subseteq \mathscr{F}(S_2)$. If ϕ controls the type I error rate at level α for $q = (M, H_0)$ w.r.t. the finer selection variable S_2, then it also controls the type I error rate at level α w.r.t. the coarser S_1.*

Proof. If $F \in H_0$, then

$$\mathbb{E}_F\left[\phi(Y)\mathbf{1}_A(Y) \,|\, S_1\right] \;=\; \mathbb{E}_F\left[\,\mathbb{E}_F\big[\phi(Y)\mathbf{1}_A(Y) \,|\, S_2\big] \,|\, S_1\right] \;\overset{\text{a.s.}}{\leq}\; \alpha.$$

□

Because $S(y) = \mathbf{1}_A(y)$ is the coarsest possible choice, a test controlling the type I error w.r.t. any other selection variable also controls the selective error in (10). At the other extreme, if $S(y) = y$, then we cannot improve on the trivial "coin-flip" test $\phi(y) \equiv \alpha$. Proposition 3 suggests that we will typically sacrifice power as we move from coarser to finer selection variables. Even so, refining the selection variable can be useful for computational reasons. For example, in the case of the lasso, by conditioning additionally on the signs of the nonzero $\hat{\beta}_j$, the selection event becomes a convex region instead of up to $2^{|\widehat{M}|}$ disjoint convex regions (Lee et al., 2013). Another valid reason to refine S_q beyond $\mathbf{1}_{A_q}$ is to strengthen our inferential guarantees in a meaningful way; for example, we can achieve achieve false coverage-statement rate (FCR) control by using $S_q = (\mathbf{1}_{A_q}(Y), |\widehat{\mathcal{Q}}(Y)|)$ (see Section 8, Proposition 11).

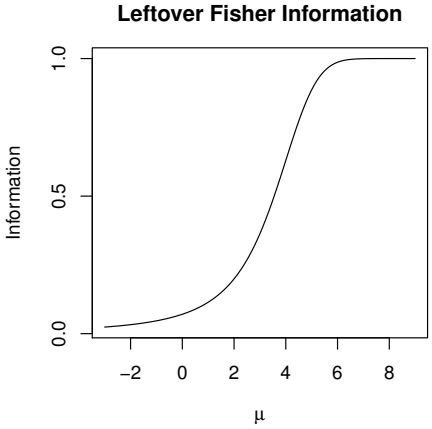
Data splitting corresponds to setting every selection variable equal to $S = Y_1$. As a result, data splitting does not use all the information that remains after conditioning on A, as we see informally in the filtration

$$\mathscr{F}_0 \underbrace{\subseteq}_{\text{used for selection}} \mathscr{F}(\mathbf{1}_A(Y_1)) \underbrace{\subseteq}_{\text{wasted}} \mathscr{F}(Y_1) \underbrace{\subseteq}_{\text{used for inference}} \mathscr{F}(Y_1, Y_2). \tag{18}$$

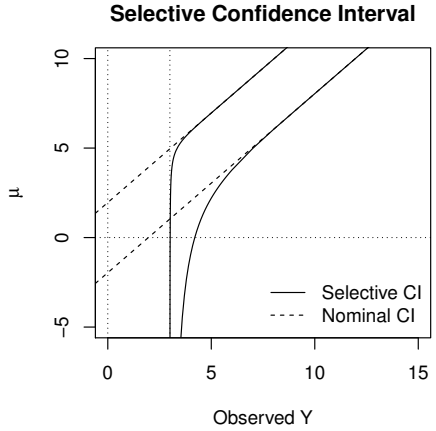
As we will see in Section 3.2, this waste of information means that data splitting is inadmissible under fairly general conditions.

We can quantify the amount of leftover information in terms of the Fisher information that remains in the conditional law of Y given S. In a smooth parametric model, we can decompose the Hessian of the log-likelihood as

$$\nabla^2 \ell(\theta;\, Y) = \nabla^2 \ell(\theta;\, S) + \nabla^2 \ell(\theta;\, Y \,|\, S) \tag{19}$$



(a) Leftover Fisher information as a function of μ. For $\mu \ll 3$, then there is very little information in the conditional distribution, since Y is conditionally highly concentrated on 3. For $\mu \gg 3$, then $\mathbb{P}_\mu(A) \approx 1$ and virtually no information is lost.



(b) Confidence intervals from inverting the UMPU tests of Section 3. For $Y \gg 3$, the interval essentially coincides with the nominal interval $Y \pm 1.96$. For Y close to 3, the wide interval reflects potentially severe selection bias.

Figure 3: Univariate Gaussian. $Y \sim N(\mu, 1)$ with selection event $A = \{Y > 3\}$.

The conditional expectation

$$\mathcal{I}_{Y \mid S}(\theta; S) = -\mathbb{E}\left[\nabla^2 \ell(\theta;\ Y \mid S) \mid S\right] \tag{20}$$

is the *leftover Fisher information* after selection at $S(Y)$. Taking expectations in (19), we obtain

$$\mathbb{E}\left[\mathcal{I}_{Y \mid S}(\theta; S)\right] \ = \ \mathcal{I}_Y(\theta) - \mathcal{I}_S(\theta) \ \preceq \ \mathcal{I}_Y(\theta). \tag{21}$$

Thus, on average, the price of conditioning on S — the price of selection — is the information S carries about θ.[3] In some cases this loss may be quite small, which a simple example elucidates.

Example 2. Consider selective inference under the univariate Gaussian model

$$Y \sim N(\mu, 1), \tag{22}$$

after conditioning on the selection event $A = \{Y > 3\}$.

Figure 3a plots the leftover information as a function of μ. If $\mu \ll 3$, there is very little information in the conditional distribution, since Y is conditionally highly concentrated on 3. By contrast, if $\mu \gg 3$, $\mathbb{P}_\mu(A) \approx 1$ and virtually no information is lost in the conditioning.

Figure 3b shows the confidence intervals that result from inverting the tests described in Section 3. When $Y \gg 3$, the interval essentially coincides with the nominal interval $Y \pm 1.96$ because there is hardly any selection bias and no real adjustment is necessary. By contrast, when Y is close to 3 it is potentially subject to severe selection bias. This fact is reflected by the confidence interval, which is both longer than the nominal interval and centered at a value significantly less than Y. Several methods for constructing conditional confidence intervals for the thresholded univariate Gaussian were analyzed by Weinstein et al. (2013), with a view toward FCR control.

[3]Note that we do not necessarily have $\mathcal{I}_{Y \mid S}(\theta; S) \preceq \mathcal{I}_Y(\theta)$ for every S. In fact there are interesting counterexamples where $\mathcal{I}_{Y \mid A}(\theta) \gg \mathcal{I}_Y(\theta)$ for certain θ, but we will not take them up here.

2.5 Conceptual Questions

We now pause to address conceptual objections that may have occurred to a skeptical reader.

2.5.1 How Can the Model Be Random?

In our framework, inference is based on a probabilistic model that is allowed to be chosen randomly, based on the data Y; the reader may wonder whether that randomness muddies the interpretation of whatever inference is carried out in the second stage.

First, note that in our framework the *true* sampling distribution F is not "selected" in any sense; it is entirely outside the analyst's control. The only thing selected is the *working model*, a tool the analyst uses to carry out inference, which may or may not include the true F.

Thoughtful skeptics may find reasons for concern about any approach — data carving, data splitting, or selecting a model based on a previous experiment — in which a random model is selected, believing that statistical testing is only appropriate when a probabilistic model can be based purely on convincing theoretical considerations. We answer only that this point of view would rule out most scientific inquiries for which statistics is currently used. However, if one is comfortable with choosing a random model based on data splitting or a previous experiment, we see no special reason to be any more concerned about choosing a random model based on data carving.

In any case, it is by no means required that the model M be random. For example, in our clinical trial example of Section 6.1, M is always the same but the null hypothesis H_0 is chosen based on inspecting the data. The same is true of the saturated-model selective z-test described in 5.1.

2.5.2 What if the Selected Model is Wrong?

If we were not writing about model selection, we might have begun by stating a formal mathematical assumption that the sampling distribution F belongs to some model M, and then devised a test ϕ that behaves well when $F \in M$. The same ϕ might not work well at all if $F \notin M$: for example, if we choose to apply the one-sample t-test of $\mu = 0$ to a sample $Y_1, \ldots, Y_n$ whose observations are highly correlated, then the probability of rejection may be a great deal larger than the nominal α, even if $\mathbb{E}[Y_i] = 0$. This is not a mistake in the formal theory, nor does it make the t-test an invalid test; rather, the validity or invalidity of a test is defined with respect to its behavior when $F \in H_0 \subseteq M$.

In any given application, the analyst must choose from among many statistical methods knowing that each one is designed to work under a particular set of assumptions about F — i.e., under a particular model M. Because our theory encompasses both the choice and the subsequent analysis, it would not be sensible to assume that the analyst is infallible. Typically some candidate models M are correctly specified (i.e., $F \in M$), others are not ($F \notin M$), and the analyst can never know for sure which are which.

Of course, the possibility of misspecification is not restricted to adaptive procedures like data carving and data splitting: selecting an inappropriate model *after* seeing the data leaves us no better or worse off than if we had chosen the same inappropriate model *before* seeing the data. Each ϕ_q or C_q is designed with respect to a particular model M, and properties like selective type I error control or selective coverage only constrain its behavior for $F \in M$. [4]

One benefit of our selective inference framework is that it lets us perform diagnostic model checks to guard against potentially troublesome forms of misspecification like interactions, correlated errors, or overdispersion. Because such checks are formally disallowed under the purist view,

[4] There is of course a separate question of robustness: if $F \notin M$ but is "close" in some sense, we may still want our procedure to behave predictably.

concerns about misspecification should typically lead us to prefer adaptive procedures like data splitting and data carving over non-adaptive ones.

Despite the formal proscription against model checking in classical inference, most practitioners still insist on doing it. This reflects a widespread believed that usually the "data snooping" aspects of model checking are relatively harmless compared to the alternative of blind faith in the first model we think of. Our selective error framework supports that intuition, as we see next.

Suppose that after initially selecting $q = (M, H_0)$, we only carry out ϕ_q if model M passes a specification test ϕ^M — i.e., a suite of model checks with overall probability α^M of rejecting model M if $F \in M$. We can adjust for the model checks by inflating our (selective) p-values by a factor $(1 - \alpha^M)^{-1}$.

Proposition 4 (Universal Correction for Model Checking)**.** *Let $\widetilde{A}_q$ denote the event that q is initially selected, and assume $\phi^M(y)$ is a selective level-α^M specification test given $\widetilde{A}_q$.*

If ϕ_q has selective level $(1-\alpha^M)\,\alpha$ given $\widetilde{A}_q$, then it has selective level α given $A_q = \widetilde{A}_q \cap \{\phi^M(Y) < 1\}$.

Proof. For $F \in H_0$,

$$\mathbb{E}_F[\phi_q(Y) \,|\, A_q] \;=\; \frac{\mathbb{E}_F[\phi_q(Y)\mathbf{1}_{A_q}(Y) \,|\, \widetilde{A}_q]}{\mathbb{P}_F(A_q \,|\, \widetilde{A}_q)} \;\leq\; \frac{(1-\alpha^M)\,\alpha}{1-\alpha^M} \;=\; \alpha.$$

□

George Box's dictum that "all models are wrong, but some are useful" is a commonplace among statisticians, who recognize that even an imperfect model may acceptably approximate the truth. However, we urge caution in one regard: even if some model gives a reasonable approximation to $\mathcal{L}(Y)$, there is no guarantee that the induced model for $\mathcal{L}(Y \,|\, A)$ is reasonable, since conditioning can introduce new robustness problems. For example, suppose that a test statistic $Z_n(Y)$ tends in distribution to $N(0,1)$ under H_0 as $n \to \infty$. In a non-selective setting, we might be comfortable modeling it as Gaussian as a basis for hypothesis testing. In this case it is also true that $\mathcal{L}(Z_n \,|\, Z_n > c)$ converges to a truncated Gaussian law for any fixed $c \in \mathbb{R}$, but the approximation may be much poorer for intermediate values of n. Worse, if $c_n \to \infty$ with n, the truncated Gaussian approximation may never be reasonable.

2.6 Prior Work on Selective Inference

This article takes its inspiration from a recent ferment of work on the problem of inference in linear regression models after model selection. Lockhart et al. (2014) derive an asymptotic test for whether the nonzero coefficients at a given knot in the lasso path contain all of the true nonzero coefficients. Taylor et al. (2014) provided an exact (finite-sample) version of this result and extended it to the LARS path, while Lee et al. (2013), Loftus and Taylor (2014), and Lee and Taylor (2014) used similar approaches to derive exact tests for the lasso with a fixed value of regularization parameter λ, forward stepwise regression, and regression after marginal screening, respectively. All of the above approaches are derived assuming that the error variance σ^2 is known or an independent estimate is available.

The present work attempts to unify the above approaches under a common theoretical framework generalizing the classical optimality theory of Lehmann and Scheffé (1955) and elucidate previously unexplored questions of power. It also lets us generalize the results to the case of unknown σ^2, and to arbitrary exponential families after arbitrary selection events.

Olshen (1973) anticipated our requirement of selective error control in the context of a two-stage multiple comparison procedure in which an F-test is first performed, then Scheffé's S-method applied if the F-test rejects. He shows that the conditional coverage in the second stage may be less than $1 - \alpha$ conditional on rejection in stage one.

Other works have viewed selective inference as a multiple inference problem. Recent work in this vein can be found in Berk et al. (2013) and Foygel Barber and Candes (2014). Section 8 argues that inference after model selection and multiple inference are distinct problems with different scientific goals. An empirical Bayes approach for selection-adjusted estimation can be found in Efron (2011).

There has also recently been work on inference in high-dimensional linear regression models, notably Javanmard and Montanari (2013) and Dezeure et al. (2014). The main difference between these works and this paper is the notion of *selective inference*. These works focus on approximate asymptotic inference for a fixed model with many variables, while we consider finite-sample inference after selecting a smaller submodel.

By far the most similar to ours is the proposal of Benjamini and Yekutieli (2005) to control the FCR, who focus as we do on problems of selective inference. They propose constructing multiple confidence intervals for an adaptively selected set of parameters that achieve "coverage on the average, over the selected ones." We see in Section 8 that selective error control is closely related to FCR control, and the former can always be adapted to achieve the latter.

3 Selective Inference in Exponential Families

As discussed in Section 2.2, we can construct selective tests "one at a time" for each model M and hypothesis H_0, conditional on the corresponding selection event and ignoring any other models that were previously under consideration. This is because the other candidate models and hypotheses are irrelevant to satisfying (10). For that reason, we suppress the explicit dependence on $q = (M, H_0)$ except where it is necessary to resolve ambiguity.

Our framework for selective inference is especially convenient when M corresponds to a multiparameter exponential family

$$Y \sim f_\theta(y) = \exp\{\theta' T(y) - \psi(\theta)\}\, f_0(y) \tag{23}$$

with respect to some dominating measure. Then, the conditional distribution given $Y \in A$ for any measurable A is another exponential family with the same natural parameters and sufficient statistics but different carrier distribution and normalizing constant:

$$(Y \,|\, Y \in A) \sim \exp\{\theta' T(y) - \psi_A(\theta)\}\, f_0(y)\, \mathbf{1}_A(y) \tag{24}$$

This fact lets us draw upon the rich theory of inference in multiparameter exponential families.

3.1 Conditional Inference and Nuisance Parameters

Classically, conditional inference in exponential families arises as a means for inference in the presence of nuisance parameters, as in Model 5 below.

Model 5 (Exponential Family with Nuisance Parameters)**.** *Y follows a p-parameter exponential family with sufficient statistics $T(y)$ and $U(y)$, of dimension k and $p-k$ respectively:*

$$Y \sim f_{\theta,\zeta}(y) = \exp\{\theta' T(y) + \zeta' U(y) - \psi(\theta, \zeta)\}\, f_0(y), \tag{25}$$

with $(\theta, \zeta) \in \Theta \subseteq \mathbb{R}^p$ open.

Assume θ corresponds to a parameter of interest and ζ to an unknown nuisance parameter. The conditional law $\mathcal{L}(T(Y) \,|\, U(Y))$ depends only on θ:

$$(T \,|\, U = u) \sim g_\theta(t \,|\, u) = \exp\{\theta' t - \psi_g(\theta \,|\, u)\}\, g_0(t \,|\, u), \tag{26}$$

letting us eliminate ζ from the problem by conditioning on U. For $\theta \in \mathbb{R}$, we obtain a single-parameter family for T.

We say a level-α selective test $\phi(y)$ is *selectively unbiased* if

$$\mathrm{Pow}_\phi(\theta \,|\, A) = \mathbb{E}_\theta[\phi(Y) \,|\, A] \geq \alpha, \quad \text{for all } \theta \in \Theta, \tag{27}$$

(27) specializes to the usual definition of an unbiased test if there is no conditioning (i.e., if $A = \mathcal{Y}$). Unbiasedness rules out tests that privilege some alternatives to the detriment of others, such as one-sided tests of two-sided alternatives.

A *uniformly most powerful unbiased* (UMPU) selective level-α test is one whose selective power is uniformly highest among all level-α tests satisfying (27). A selectively unbiased confidence region is one that inverts a selectively unbiased test, and confidence regions inverting UMPU selective tests are called uniformly most accurate unbiased (UMAU). All of the above specialize to the usual definitions when $A = \mathcal{Y}$.

See Lehmann and Romano (2005) or Brown (1986) for thorough reviews of the rich literature on testing in exponential family models. In particular, the following classic result of Lehmann and Scheffé (1955) gives a simple construction of UMPU tests in exponential family models.

Theorem 6 (Lehmann and Scheffé (1955))**.** *Under Model 5 with $k = 1$, consider testing the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{against} \quad H_1 : \theta \neq \theta_0 \tag{28}$$

at level α. There is a UMPU test of the form $\phi(T(y), U(y))$ with

$$\phi(t,u) = \begin{cases} 1 & t < c_1(u) \text{ or } t > c_2(u) \\ \gamma_i & t = c_i(u) \\ 0 & c_1(u) < t < c_2(u) \end{cases} \tag{29}$$

for which c_i and γ_i solve

$$\mathbb{E}_{\theta_0}\left[\phi(Y) \,|\, U(Y) = u\right] = \alpha \tag{30}$$

$$\mathbb{E}_{\theta_0}\left[T(Y)\phi(Y) \,|\, U(Y) = u\right] = \alpha\, \mathbb{E}_{\theta_0}\left[T(Y) \,|\, U(Y) = u\right]. \tag{31}$$

The condition (30) constrains the power to be α at $\theta = \theta_0$, and (31) is obtained by differentiating the power function and setting its derivative to 0 at $\theta = \theta_0$.

Because $\mathcal{L}(Y \,|\, A)$ is an exponential family, we can simply apply Theorem 6 to the conditional law $\mathcal{L}(Y \,|\, A)$ to obtain an analogous construction in the selective setting.

Corollary 7 (UMPU Selective Tests)**.** *Under Model 5 with $k = 1$, consider testing the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{against} \quad H_1 : \theta \neq \theta_0 \tag{32}$$

at selective level α on selection event A. There is a UMPU selective test of the form $\phi(T(y), U(y))$ with

$$\phi(t,u) = \begin{cases} 1 & t < c_1(u) \text{ or } t > c_2(u) \\ \gamma_i & t = c_i(u) \\ 0 & c_1(u) < t < c_2(u) \end{cases} \tag{33}$$

for which c_i and γ_i solve

$$\mathbb{E}_{\theta_0}\left[\phi(Y) \,|\, U(Y) = u,\ Y \in A\right] = \alpha \tag{34}$$

$$\mathbb{E}_{\theta_0}\left[T(Y)\phi(Y) \,|\, U(Y) = u,\ Y \in A\right] = \alpha\, \mathbb{E}_{\theta_0}\left[T(Y) \,|\, U(Y) = u,\ Y \in A\right]. \tag{35}$$

In the same way, we can generalize Corollary 7 to obtain an optimal selective test with respect to a finer selection variable S. Because selective error control is a more stringent requirement the finer S is, the power of the optimal test will tend to attenuate as we refine S. Unless we have a good computational or inferential reason to do so, we should never condition on more than the selection event A.

There is no reason to insist absolutely on unbiasedness; it is merely a natural way to choose a test when there is no completely UMP one. For example, another simple choice is to use the equal-tailed test from the same conditional law (26). The equal-tailed level-α rejection region is simply the union of the one-sided level-$\alpha/2$ rejection regions. While the equal-tailed and UMPU tests have different ways of choosing c_i and γ_i, both tests take the form (29). There is a good reason why, as we are about to see.

3.2 Conditioning, Admissibility, and Data Splitting

The selective level-α test ϕ is *inadmissible* on selection event A if there exists another selective level-α test ϕ^* for which

$$\mathbb{E}_{\theta,\zeta}[\phi^*(Y)|A] \geq \mathbb{E}_{\theta,\zeta}[\phi(Y)|A], \quad \text{for all } (\theta,\zeta) \in \Theta, \tag{36}$$

with the inequality strict for at least one (θ, ζ).

For $k \geq 1$, we say $\phi(y)$ has *convex U-sections* if for each u there is some closed convex set $C(u)$ for which

$$\phi(y) = \begin{cases} 0 & T(y) \in \text{int } C(U(y)) \\ \gamma(y) & T(y) \in \partial C(U(y)) \\ 1 & T(y) \notin C(U(y)) \end{cases}, \tag{37}$$

which generalizes (29) to $k \geq 1$. Let $\mathscr{C}(T;U) \subseteq m\mathscr{F}(T,U)$ denote all tests of the form (37), as well as tests that are a.s. equivalent to a test in $\mathscr{C}(T;U)$. Here, $m\mathscr{F}(T,U)$ is the space of functions measurable w.r.t. $\mathscr{F}(T,U)$; informally, all functions of the form $f(y) = g(T(y), U(y))$.

Generalizing results of Birnbaum (1955) Matthes and Truax (1967, Theorem 3.1) show that, for testing

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0 \tag{38}$$

for $k \geq 1$ in Model 5, all admissible tests in $m\mathscr{F}(T,U)$ are in $\mathscr{C}(T;U)$.

If $\mathcal{L}(T \mid U)$ is absolutely continuous, then all admissible tests must be functions of the sufficient statistics T and U on A.

Corollary 8. *Under Model 5, consider a selective level-α test ϕ of the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{against} \quad H_1 : \theta \neq \theta_0 \tag{39}$$

If $\mathcal{L}(T \mid U)$ is absolutely continuous, then ϕ is inadmissible unless $\phi \in m\mathscr{F}(T,U,A)$.

Proof. Define the Rao-Blackwellized critical function

$$\overline{\phi}(y) = \mathbb{E}\left[\phi(Y) \mid T(Y), U(Y), Y \in A\right] \tag{40}$$

The selective power functions of ϕ and $\overline{\phi}$ are identical, so ϕ is admissible if and only if $\overline{\phi}$ is. Assuming $\overline{\phi}$ is admissible, by Matthes and Truax (1967, Theorem 3.1) applied to $\mathcal{L}(Y \mid A)$, it must have convex U-sections on A. Because $\mathcal{L}(T \mid U)$, and thus $\mathcal{L}(T \mid U, A)$, is absolutely continuous, the boundaries of the $C(u)$ have probability zero, so $\overline{\phi}(Y) \in \{0,1\}$ a.s. on A. But because $\phi(y) \in [0,1]$, we must also have $\phi(Y) \overset{\text{a.s.}}{=} \overline{\phi}(Y)$ there. □

Corollary 8 makes it hard for a data-splitting test to be admissible.

Model 9 (Exponential Family with Data Splitting)**.** *Model random variables* $Y_1 \perp\!\!\!\perp Y_2$ *as*

$$Y_i \sim \exp\left\{\theta T_i(y) + \zeta' U_i(y) - \psi_i(\theta, \zeta)\right\}\ f_{0,i}(y), \quad i = 1, 2, \tag{41}$$

with $\theta \in \mathbb{R}$ *and with the models for* Y_i *both satisfying Model 5.*

A *data-splitting procedure* is one for which $A \in \mathscr{F}(Y_1)$ and $\phi \in m\mathscr{F}(Y_2)$; that is, selection uses only Y_1 and inference uses only Y_2.

Our next theorem shows that if there is any information left over in Y_1 after conditioning on A and U_1, then data splitting is inadmissible as a result of discarding that information.

Theorem 10 (Inadmissibility of Data Splitting)**.** *Under Model 9, assume* $\operatorname{supp}(T_2 \,|\, U_2)$ *is a closed interval* $[a, b]$ *with* $a, b \in [-\infty, +\infty]$*. Consider testing the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{against} \quad H_1 : \theta \neq \theta_0 \tag{42}$$

at selective level $\alpha \in (0, 1)$ *on the event* A*. Any data splitting test* ϕ *is inadmissible unless* T_1 *is a function of* U_1 *on* A*.*

We defer the proof to Appendix B.

Remark Our condition on the support of $T_2 \,|\, U_2$ is stronger than it really needs to be. It serves to rule out certain perverse counterexamples, but weaker conditions could suffice at the cost of increased complication. What we really need is some guarantee that $\mathcal{L}_{\theta_0}(T_2 \,|\, U_2)$ puts some probability mass near one of ϕ's accept/reject cutoffs.

Example 3. To illustrate Theorem 10, consider a bivariate version of Example 2:

$$Y_i \sim N(\mu, 1), \ \ i = 1, 2, \quad \text{with } Y_1 \perp\!\!\!\perp Y_2, \tag{43}$$

in which we condition on the selection event $A = \{Y_1 > 3\}$.

With data splitting, we could construct a 95% confidence interval using only Y_2; namely, $Y_2 \pm 1.96$. This interval is valid but does not use all the information available. A more powerful alternative is to construct an interval based on the law

$$\mathcal{L}_\mu\left(Y_1 + Y_2 \ \middle|\ Y_1 > 3\right), \tag{44}$$

which uses the leftover information in Y_1.

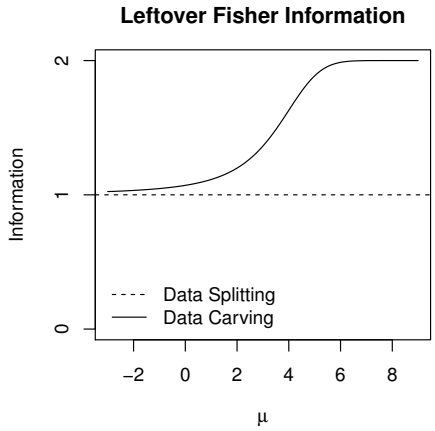
Figure 4a shows the Fisher information that is available to each test as a function of μ. The Fisher information of data splitting is exactly 1 no matter what μ is, whereas the optimal selective test has information approaching 2 as μ increases. Figure 4b shows the expected confidence interval length of the equal tailed interval as a function of μ. For $\mu \gg 3$, the data splitting interval is roughly 41% longer than it needs to be (in the limit, the factor is $\sqrt{2} - 1$).

Together, the plots tell a consistent story: when the selection event is not too unlikely, discarding the first data set exacts an unnecessary toll on the power of our second-stage procedure.

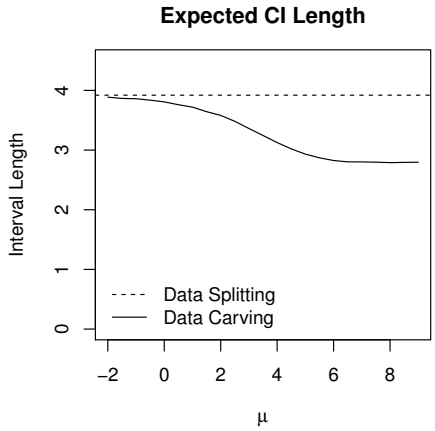
4 Computation

Corollary 7 gives us a principled way of constructing optimal hypothesis tests and confidence intervals in a selective exponential family setting, by reducing our problem to inference in the one-parameter exponential family $\mathcal{L}_\theta(T \,|\, U, A)$.

However, actually computing rejection cutoffs or interval endpoints requires access to that conditional law, either by an analytic form or by Monte Carlo. Depending on the model and the selection procedure, this computation may be trivial or it may be exceedingly difficult. It would be impossible to give a completely general treatment here, but we can suggest some general strategies.



(a) Fisher information available for second-stage inference.



(b) Expected confidence interval length.

Figure 4: Contrast between data splitting and data carving in Example 3, in which $Y_i \sim N(\mu, 1)$ independently for $i = 1, 2$. Data splitting discards Y_1 entirely, while data carving uses the leftover information in Y_1 for the second-stage inference. When $\mu \ll 3$, data carving also uses about one data point for inference since there is no information left over in Y_1. But when $\mu \gg 3$, conditioning barely effects the law of Y_1 and data carving has nearly two data points left over.

4.1 Monte Carlo Tests and Intervals

If we can obtain a stream of samples from $\mathcal{L}_\theta(T \,|\, U, A)$ for any value of θ, then we can carry out hypothesis tests and construct intervals. This can be done efficiently via rejection sampling if, for example, we can sample efficiently from $\mathcal{L}_\theta(T \,|\, U)$ and $\mathbb{P}_\theta(A \,|\, U)$ is not too small. Otherwise, more specialized approaches may be required.

Consider a test based on the statistic Z, which is distributed according to a one-parameter exponential family

$$Z \sim g_\theta(z) = e^{\theta z - \psi(\theta)}\, g_0(z) \tag{45}$$

Assume that in addition to Z we are given an independent sequence from the reference distribution

$$Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} g_0(z) \tag{46}$$

There is a Monte Carlo one-sided test of $H_0 : \theta \leq 0$, which rejects if the observed value Z is among the $(n+1)\,\alpha$ largest of $Z, Z_1, \ldots, Z_n$.

We can use the same sequence to test $H_0 : \theta \leq \theta_0$ for any other θ_0. Denote the importance-weighted empirical expectation as

$$\widehat{\mathbb{E}}_\theta\, h(Z) = \frac{\sum_{i=1}^n h(Z_i) e^{\theta Z_i}}{\sum_{i=1}^n e^{\theta Z_i}} \tag{47}$$

$$\xrightarrow{a.s.} \mathbb{E}_\theta\, h(Z) \quad \text{as } n \to \infty \text{ for integrable } h. \tag{48}$$

In effect, we have put an exponential family "through" the empirical distribution of the Z_i as suggested by Efron et al. (1996).

The Monte Carlo one-sided cutoff for a test of $H_0 : \theta \leq \theta_0$ is the smallest c_2 for which

$$\widehat{\mathbb{P}}_{\theta_0}(Z > c_2) \leq \alpha. \tag{49}$$

The test rejects for $Z > c_2$ and randomizes appropriately at $Z = c_2$.

The Monte Carlo UMPU two-sided test of $H_0 : \theta = \theta_0$ is a bit more involved, but similar in principle. We can solve for $c_1, \gamma_1, c_2, \gamma_2$ for which

$$\widehat{\mathbb{E}}_{\theta_0} \phi(Z) = \alpha \tag{50}$$

$$\widehat{\mathbb{E}}_{\theta_0} [Z\phi(Z)] = \alpha \, \widehat{\mathbb{E}}_{\theta_0} Z. \tag{51}$$

In Appendix C we discuss how (50–51) can be solved efficiently for fixed θ_0 and inverted to obtain a confidence interval. Monte Carlo inference as described above is computationally straightforward once $Z_1, \ldots, Z_n$ are obtained.

More generally, the Z_i could represent importance samples with weights W_i, or steps in a Markov chain with stationary distribution $g_{\theta_0}(z)$. The same methods apply as long as we still have

$$\widehat{\mathbb{E}}_\theta \, h(Z) = \frac{\sum_{i=1}^n W_i \, h(Z_i) \, e^{\theta Z_i}}{\sum_{i=1}^n W_i \, e^{\theta Z_i}} \tag{52}$$

$$\xrightarrow{a.s.} \mathbb{E}_\theta \, h(Z), \quad \text{for integrable } h. \tag{53}$$

Numerical problems may arise in solving (50–51) for θ_0 far away from the reference parameter used for sampling. Combining appropriately weighted samples from several reference values θ can help to keep the effective sample size from getting too small for any θ_0.

For further references on Monte Carlo inference see Besag and Clifford (1989); Forster et al. (1996); Mehta et al. (2000).

4.2 Sampling Gaussians with Affine and Quadratic Constraints

In the case where Y is Gaussian, several simplifications are possible. For one, there are many ways to sample from a truncated multivariate Gaussian distribution. In this paper, we use hit-and-run Gibbs sampling algorithms, while Pakman and Paninski (2014) suggest another approach based on Hamiltonian Monte Carlo.

The case where Y is Gaussian is especially important because it includes linear regression as a special case. The selection events of many popular selection procedures can be characterized by affine and quadratic inequality constraints on Y. As various authors have shown, such procedures include the lasso and elastic net (Lee et al., 2013), forward stepwise regression (Loftus and Taylor, 2014), and linear regression after marginal screening of parameters (Lee and Taylor, 2014). Furthermore, these papers derive closed-form distributions, obviating the need for sampling.

5 Selective Inference for Linear Regression

We now revisit an example that we encountered earlier—namely, linear regression—from the perspective of exponential families. Recall that the data were assumed to be generated from a multivariate normal distribution

$$Y \sim N_n(\mu, \, \sigma^2 I_n),$$

where μ is modeled as

$$\mu = X_M \beta^M. \tag{54}$$

To avoid trivialities, we will assume that X_M has full column rank for all M under consideration, so that β^M is well-defined. We call (6) with no constraints on μ the *saturated model*.

Even if we do not take the linear model (54) seriously, there is still a well-defined best linear predictor in the population for design matrix X_M:

$$\theta^M = \arg\min_\theta \mathbb{E}_\mu \left[\|Y - X_M \theta\|^2 \right] = X_M^\dagger \mu, \tag{55}$$

where $X_M^\dagger$ is the Moore-Penrose pseudoinverse. We call θ^M the *least squares coefficients* for M. According to this point of view, each θ_j^M corresponds to a particular linear functional $\eta'\mu$ for

$$\eta = \frac{X_{j\cdot M}}{\|X_{j\cdot M}\|^2}, \quad \text{where } X_{j\cdot M} = \mathcal{P}^\perp_{X_{M\setminus j}} X_j \tag{56}$$

is the remainder after adjusting X_j for the other columns of X_M, and $\mathcal{P}_{X_{M\setminus j}}$ denotes projection onto the column space of $X_{M\setminus j}$.

Several recent articles have tackled the problem of exact selective inference in linear regression after specific selection procedures (Lee et al., 2013; Loftus and Taylor, 2014; Lee and Taylor, 2014). These all use the saturated model as a way of avoiding the need to consider multiple candidate probabilistic models. They also assume the error variance is known, or that an estimate may be obtained from independent data, and target least-squares parameters in the saturated model.

Under the linear model with predictors X_M, there is no distinction between β^M and θ^M, whereas under the saturated model β^M may not exist (i.e., there is no β^M such that $\mu = X_M\beta^M$). For statistical inference, the main difference between the two is that the saturated model has $n - |M|$ additional nuisance parameters corresponding to $\mathcal{P}^\perp_{X_M}\mu$.

Notationally, instead of β_j^M denoting the jth coordinate of β^M, it will be more convenient to let it denote the coefficient of feature X_j if $j \in M$. For example, $\beta^{\{5,7\}} = \left(\beta_5^{\{5,7\}},\ \beta_7^{\{5,7\}}\right)$. Depending on whether σ^2 is assumed known or unknown, hypothesis tests for coordinates of β^M generalize either the z-test or the t-test.

5.1 Inference Under the Saturated Model

We can write the saturated model in exponential family form as

$$Y \sim \exp\left\{\frac{1}{\sigma^2}\mu' y + \frac{1}{2\sigma^2}\|y\|^2 - \psi(\mu, \sigma^2)\right\}, \tag{57}$$

which has $n+1$ natural paramaters if σ^2 is unknown (in which case $1/\sigma^2$ is a natural parameter) and n otherwise.

Consider inference with respect to some least-squares coefficient $\theta_j^M = \eta'\mu$ for some $\eta \in \mathbb{R}^n$. We can rewrite (57) in terms of η as

$$Y \sim \exp\left\{\frac{1}{\sigma^2\|\eta\|^2}\mu'\eta\ \eta' y + \frac{1}{\sigma^2}(\mathcal{P}_\eta^\perp\mu)'\ (\mathcal{P}_\eta^\perp y) - \frac{1}{2\sigma^2}\|y\|^2 - \psi(\mu,\sigma^2)\right\}. \tag{58}$$

If σ^2 is known, inference for θ_j^M after selection event A is based on the conditional law

$$\mathcal{L}_{\theta_j^M}\left(\eta' Y \ \middle|\ \mathcal{P}_\eta^\perp Y,\ A\right). \tag{59}$$

But for our conditioning on A, $\eta' Y$ would be independent of $\mathcal{P}_\eta^\perp Y$, so we would be performing the usual z-test with

$$Z = \frac{\eta' Y}{\sigma\|\eta\|} \sim N(0,1) \tag{60}$$

under the null hypothesis.

If σ^2 is unknown, inference is instead based on

$$\mathcal{L}_{\theta_j^M/\sigma^2}\left(\eta' Y \ \middle|\ \mathcal{P}_\eta^\perp Y,\ \|Y\|,\ A\right). \tag{61}$$

Unfortunately, the conditioning in (61) is too restrictive. The set

$$\left\{y :\ \mathcal{P}_\eta^\perp y = \mathcal{P}_\eta^\perp Y,\ \ \|y\| = \|Y\|\right\} \tag{62}$$

is a line intersected with the sphere $\|Y\|S^{n-1}$, and consists only of the two points $\{Y,\ Y - 2\eta'Y\}$, which are equally likely under the hypothesis $\theta_j^M = 0$. Under the saturated model, conditioning on $\|Y\|$ leaves insufficient information about θ_j^M to carry out a meaningful test.

5.2 Inference Under the Selected Linear Model

Under the selected model, $\beta_j^M = \theta_j^M = \eta'\mu$. Suppressing the superscript M in β^M, the selected model has the form

$$Y \sim \exp\left\{ \frac{1}{\sigma^2}\beta' {X_M}'y - \frac{1}{2\sigma^2}\|y\|^2 - \psi(X_M\beta, \sigma^2) \right\} \tag{63}$$

If σ^2 is known, then inference for β_j is based on

$$\mathcal{L}_{\beta_j}\left(X_j'Y \;\middle|\; {X_{M\setminus j}}'Y,\ A\right), \tag{64}$$

and otherwise on

$$\mathcal{L}_{\beta_j/\sigma^2}\left(X_j'Y \;\middle|\; {X_{M\setminus j}}'Y,\ \|Y\|,\ A\right). \tag{65}$$

Because

$$\begin{aligned} X_j'y &= X_j'\mathcal{P}_{X_{M\setminus j}}y + X_j'\mathcal{P}^{\perp}_{X_{M\setminus j}}y && (66)\\ &= X_j'\mathcal{P}_{X_{M\setminus j}}y + \|X_{j.M}\|^2\eta'y, && (67) \end{aligned}$$

we can equivalently base the tests on

$$\mathcal{L}_{\beta_j}\left(\eta'Y \;\middle|\; {X_{M\setminus j}}'Y,\ A\right) \quad \text{and} \quad \mathcal{L}_{\beta_j/\sigma^2}\left(\eta'Y \;\middle|\; {X_{M\setminus j}}'Y,\ \|Y\|,\ A\right), \tag{68}$$

respectively. If σ^2 is known, $Z = \eta'Y/\sigma\|\eta\|$ is the usual z-test statistic. If σ^2 is unknown, we can equivalently base the test on the usual t-test statistic

$$\widetilde{T} = \frac{\eta'Y}{\hat{\sigma}\|\eta\|},$$

where the usual χ^2 estimate of variance,

$$\hat{\sigma}^2 = \frac{\|\mathcal{P}^{\perp}_{X_M}Y\|^2}{n - |M|} = \frac{\|Y\|^2 - \|\mathcal{P}_{X_M}Y\|^2}{n - |M|}, \tag{69}$$

depends only on quantities that we have already conditioned on.[5]

In the case where there is no selection, $Z \sim N(0,1)$ and $\widetilde{T} \sim t_{n-|M|}$, giving the usual z- and t-tests. In the selective case (A is a proper subset of $\mathbb{R}^n$), we use the same test statistics, but their null distributions are different depending on the form of A.

Constructing a selective t-interval is slightly less straightforward than the general case described in Section 4.1 because β_j is not a natural parameter of the selected model; rather, β_j/σ^2 is. Testing $\beta_j = 0$ is equivalent to testing $\beta_j/\sigma^2 = 0$, but testing $\beta_j = c$ for $c \neq 0$ does not correspond to any point null hypothesis about β_j/σ^2. However, we can define

$$\tilde{y} = y - bX_j \sim N(X\beta - bX_j, \sigma^2 I). \tag{70}$$

Because $(\beta_j - b)/\sigma^2$ is a natural parameter for $\tilde{y}$, we can carry out a UMPU selective t-test for $H_0 : \beta_j = b \iff (\beta_j - b)/\sigma^2 = 0$.

[5]Note that, given A, $\hat{\sigma}^2$ in (69) is neither unbiased for σ^2 nor χ^2-distributed. We recommend against viewing it as a serious estimate of σ^2 in the selective setting.

5.3 Saturated Model or Selected Model?

When σ^2 is known, we have a choice whether to carry out the z-test with test statistic $Z = \eta' Y / \sigma \|\eta\|$ in the saturated or the selected model. In other words, we must choose either to assume that $\mathcal{P}_{X_M}^{\perp} \mu = 0$ or to treat it as an unknown nuisance parameter. Writing

$$U = X_{M \setminus j}{}' Y, \quad \text{and} \quad V = \mathcal{P}_{X_M}^{\perp} Y, \tag{71}$$

we must choose whether to condition on U and V (saturated model) or only U (selected model). Conditioning on both U and V can never increase our power relative to conditioning only on U, and will in most cases reduce it per Corollary 8.

In the non-selective case, this choice makes no difference at all since T, U, and V are mutually independent. In the selective case, however, the choice may be of major consequence as it can lead to very different tests. In general, T, U, and V are not conditionally independent given A, and $\mathcal{P}_{X_M}^{\perp} \mu$ may play an important role in determining the conditional distribution of T. If we needlessly condition on V, we may lose a great deal of power, whereas failing to condition on V could lead us astray if $\mathcal{P}_{X_M}^{\perp} \mu$ is large. A simple example can elucidate this contrast.

Example 4. Suppose that $y \sim N_2(\mu, I_2)$, with design matrix $X = I_2$, and we choose the best one-sparse model. Our selection procedure chooses $M = \{1\}$ if $|Y_1| > |Y_2|$, and $M = \{2\}$ otherwise.

Figure 5 shows one realization of this process with $Y = (2.9, 2.5)$. $|Y_1|$ is a little larger than $|Y_2|$, so we choose $M = \{1\}$. The yellow highlighted region $A = \{|Y_1| > |Y_2|\}$ is the chosen selection event, and the selected model is

$$Y \sim N_2\left((\mu_1, 0), I_2\right). \tag{72}$$

In this case, $T = Y_1$, $V = Y_2$, and there is no U since X_M has only one column. The selected-model test is based on $\mathcal{L}(Y_1 \,|\, A)$, whereas the saturated-model test is based on $\mathcal{L}(Y_1 \,|\, Y_2, A)$. The second conditioning set, a union of two rays, is plotted in brown. Under the hypothesis $\mu = 0$, the realized $|Y_1|$ is quite large given A, giving p-value 0.015. By contrast, $|Y_1|$ is not terribly large given $\{Y_2 = 2.5\} \cap A = \{Y_2 = 2.5, |Y_1| > 2.5\}$, leading to p-value 0.30.

The difference between the saturated and selected models is especially important in early steps of sequential model-selection procedures that use a form of the saturated-model z-test. It has been observed in several cases that if there are two strong variables with similar effect sizes, the p-value in the first step may not be very small (Taylor et al., 2014; G'Sell et al., 2013). We will explore this subtle issue further in a forthcoming companion paper dealing with sequential model selection.

5.4 Connections to Related Work

The works cited earlier use the saturated model with known error variance, testing the hypothesis

$$H_0 : \eta' \mu = \theta_0 \quad \text{against} \quad \eta' \mu \neq \theta_0. \tag{73}$$

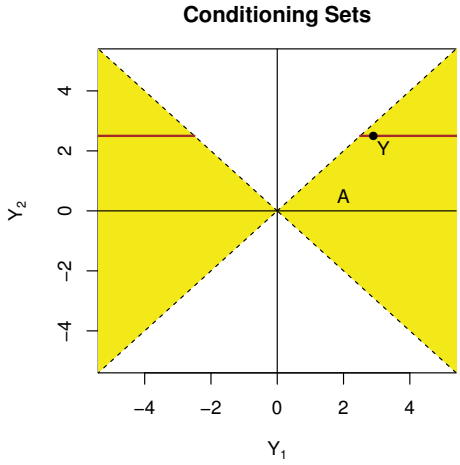
Assuming $\theta_0 = 0$ and $\sigma^2 = 1$, these tests base inference on a pivotal uniform statistic of the form

$$W(Y) = \frac{\Phi(\eta' Y) - \Phi(\mathcal{V}^-)}{\Phi(\mathcal{V}^+) - \Phi(\mathcal{V}^-)}, \tag{74}$$

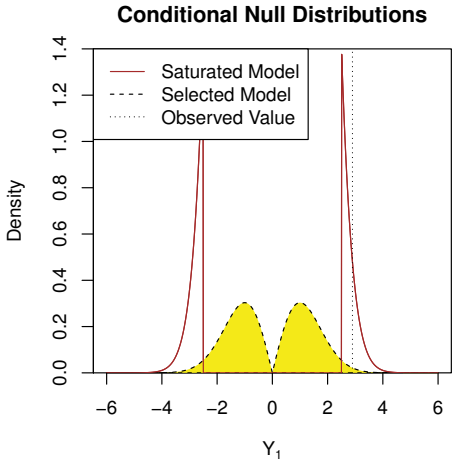
where A is some convex region of $\mathbb{R}^n$, often polyhedral, and

$$\mathcal{V}^+(Y) = \sup_{\{t:\, Y + t\eta \in A\}} \eta'(Y + t\eta) \tag{75}$$

$$\mathcal{V}^-(Y) = \inf_{\{t:\, Y + t\eta \in A\}} \eta'(Y + t\eta) \tag{76}$$



(a) For $Y = (2.9, 2.5)$, the selected-model conditioning set is $A = \{y : |y_1| > |y_2|\}$, a union of quadrants, plotted in yellow. The saturated-model conditioning set is $\{y : y_2 = 2.5\} \cap A = \{y : y_2 = 2.5, |y_1| > 2.5\}$, a union of rays, plotted in brown.



(b) Conditional distributions of Y_1 under $H_0 : \mu_1 = 0$. Under the hypothesis $\mu = 0$, the realized $|Y_1|$ is quite large given A, giving p-value 0.015. By contrast, $|Y_1|$ is not too large given $A \cap \{y : y_2 = Y_2\}$, giving p-value 0.3.

Figure 5: Contrast between the saturated-model and selected-model tests in Example 4, in which we fit a one-sparse model with design matrix $X = I_2$. The selected-model test is based on $\mathcal{L}_0(Y_1 \,|\, A)$, whereas the saturated-model test is based on $\mathcal{L}_0(Y_1 \,|\, Y_2, A)$.

Under H_0, the test statistic $\eta' Y$ takes on the distribution of a standard Gaussian random variable truncated to the interval $[\mathcal{V}^-, \mathcal{V}^+]$. As a result, $W(Y)$ is exactly the cumulative distribution function of $\mathcal{L}_0(\eta' Y \,|\, \mathcal{P}_\eta^\perp Y,\ A)$. In other words, $W(Y)$ is the observed quantile of $T = \eta' Y$ under its null distribution. Figure 6 illustrates the logic of this procedure.

These procedures are selective z-tests under the saturated model. The selected-model approach allows us to perform more powerful inference when σ^2 is known, and to use the selective t-test when it is unknown.

6 Other Exponential Families

In this section we describe tests in two simple non-Gaussian settings, selective inference in a binomial problem, and tests involving a scan statistic in Poisson process models. More generally, we address the question of selective inference in generalized linear models.

6.1 Selective Clinical Trial

Consider a clinical trial with m candidate treatments for heart disease. We give treatment j to n_j patients for $0 \le j \le m$, with $j = 0$ corresponding to the placebo. The number of patients on treatment j to suffer a heart attack during the trial is

$$Y_j \overset{\text{ind.}}{\sim} \text{Binom}(p_j, n_j), \quad \text{with } \log \frac{p_j}{1 - p_j} = \begin{cases} \theta & j = 0 \\ \theta - \beta_j & j > 0 \end{cases}, \tag{77}$$

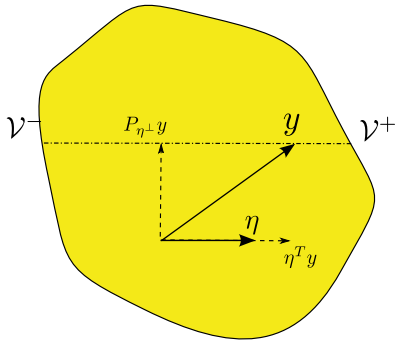


Figure 6: Saturated-model inference for a generic convex selection set for $Y \sim N(\mu, I_n)$. After conditioning on the yellow set A, $\mathcal{V}^+$ is the largest $\eta' Y$ can get while $\mathcal{V}^-$ is the smallest it can get. Under $H_0 : \eta'\mu = 0$, the test statistic $\eta' Y$ takes on the distribution of a standard Gaussian random variable truncated to the interval $[\mathcal{V}^-, \mathcal{V}^+]$. As a result, $W(Y) = \frac{\Phi(\eta' Y) - \Phi(\mathcal{V}^-)}{\Phi(\mathcal{V}^+) - \Phi(\mathcal{V}^-)}$ is uniformly distributed.

so β_j measures the efficacy of treatment j. The likelihood for Y is

$$Y \sim \exp\left\{\theta \sum_{j=0}^{m} y_j - \sum_{j=1}^{m} \beta_j y_j - \psi(\theta, \beta)\right\} \prod_{j=0}^{m} \binom{Y_j}{n_j}, \tag{78}$$

an exponential family with $m+1$ sufficient statistics.

After observing the data, we select the best treatment in-sample, then construct a confidence interval for its odds ratio relative to placebo. Defining $\hat{p}_j = Y_j/n_j$, let $\widehat{J}(Y) = \arg\min_j \hat{p}_j$, and assume for simplicity the n_j are relatively prime, ruling out ties. If the first treatment is selected, we base our inference for β_1 on the conditional law

$$\mathcal{L}_{\beta_1}\left(Y_1 \;\middle|\; \sum_{j=0}^{m} Y_j,\; Y_2, \ldots, Y_J,\; \widehat{J}(Y) = 1\right) \tag{79}$$

Under this law, $Y_2, \ldots, Y_m$ are fixed, as is $Y_0 + Y_1$, with Y_0 and Y_1 the only remaining unknowns. Thus, we have the two-by-two multinomial table

	Control	Treatment
Heart attack	Y_0	Y_1
No heart attack	$n_0 - Y_0$	$n_1 - Y_1$

The margins are fixed and we have the constraint that $Y_1 < n_1 \min_{j>1} \hat{p}_j$. This amounts to a selective Fisher's exact test. Aside from the constraint on its support, the distribution of Y_1 is hypergeometric if $\beta_1 = 0$ and otherwise noncentral hypergeometric with noncentrality parameter β_1; we can use this distribution to construct an interval for β_1.

6.2 Poisson Scan Statistic

As a second simple example, consider observing a Poisson process $Y = \{Y_1, \ldots, Y_{N(Y)}\}$ on the interval $[0, 1]$ with piecewise-constant intensity, possibly elevated in some unknown window $[a, b]$. That is, $Y \sim \text{Poisson}(\lambda(t))$ with

$$\lambda(t) = \begin{cases} e^{\alpha+\beta} & t \in [a, b] \\ e^{\alpha} & \text{otherwise.} \end{cases} \tag{80}$$

Our goal is to locate $[a, b]$ by maximizing some scan statistic, then test whether $\beta > 0$ or construct a confidence interval for it. Assume we always have $[\hat{a}, \hat{b}] = [Y_i, Y_j]$ for some i, j; this is true, for example, if we use the multi-scale-adjusted likelihood ratio statistic proposed in Rivera and Walther (2013).

The density of Y can be written in exponential family form as

$$Y \sim \exp\left\{ \sum_{i=1}^{N(y)} \log \lambda(y_i) - \int_0^1 \lambda(s)\, ds \right\} \tag{81}$$

$$= \exp\big\{ \alpha\, N(Y) + \beta\, T(y) - \psi(\alpha, \beta) \big\}, \tag{82}$$

where

$$T(y) = \sum_{i=1}^{N(y)} \mathbf{1}\{y_i \in [a, b]\} \quad \text{and} \quad \psi(\alpha, \beta) = e^{\alpha}(1 - b + a) + e^{\alpha+\beta}(b - a). \tag{83}$$

If A is the event that $[a, b]$ is chosen, we carry out inference with respect to $\mathcal{L}_\beta\,(T \,|\, N, A)$. Note that under $\beta = 0$ and conditional on N, Y is an i.i.d. uniform random sample on $[0, 1]$.

Once we condition on the event $\{a, b \in Y\}$, the other $N - 2$ values are uniform. Thus, we can sample from $\mathcal{L}_\beta\,(T \,|\, N, A)$ with $\beta = 0$ by taking Y to include a, b, and $N - 2$ uniformly random points, then rejecting samples for which $[a, b]$ is not the selected window.

6.3 Generalized Linear Models

Our framework extends to logistic regression, Poisson regression, or other generalized linear model (GLM) with response Y and design matrix X, since the GLM model may be represented as an exponential family of the form

$$Y \sim \exp\left\{\beta'\ X'y - \psi(X\beta)\right\}\ f_0(y). \tag{84}$$

As a result, we can proceed just as we did in the case of linear regression in the reduced model, conditioning on $U = X_{M\setminus j}{}'Y$ and basing inference on $\mathcal{L}_{\beta_j^M}(X_j'Y \,|\, U, A)$.

A difficulty may arise for logistic or Poisson regression due to the discreteness of the response distribution Y. If some control variable X_1 is continuous, then for almost every realization of X, all configurations of Y yield unique values of $U = X_1'Y$. In that case, conditioning on $X_1'Y$ means conditioning on Y itself. No information is left over for inference, so that the best (and only) exact level-α selective test is the trivial one $\phi(Y) \equiv \alpha$. By contrast, if all of the control variables are discrete variables like gender or ethnicity, then conditioning on U may not constrain Y too much.

Because $X'Y$ is approximately a multivariate Gaussian random variable, a more promising approach may to base inference on the asymptotic Gaussian approximation, though we will not pursue that here. Tian and Taylor (2014) discuss selective inference in certain non-Gaussian problems.

7 Simulation: High-Dimensional Regression

As a simple illustration, we compare selective inference in linear regression after the lasso for $n = 100, p = 200$. Here, the rows of the design matrix X are drawn from an equicorrelated multivariate Gaussian distribution with pairwise correlation $\rho = 0.3$ between the variables. The columns are normalized to have length 1.

We simulate from the model

$$Y \sim N(X\beta, I_n), \tag{85}$$

with β 7-sparse and its non-zero entries set to 7. The signal to noise ratio (SNR) (magnitude of β) was chosen so that data splitting with half the data yielded a superset of the true variables on roughly 20% of instances. For data splitting and carving, Y is partitioned into selection and inference data sets Y_1 and Y_2, containing n_1 and $n_2 = n - n_1$ data points respectively.

We compare two procedures:

Data Splitting after Lasso on Y_1 (Split_{n_1}): Use the lasso on Y_1 to select the model, and use Y_2 for inference.

Data Carving after Lasso on Y_1 (Carve_{n_1}): Use the lasso on Y_1 to select the model, and use Y_2 and whatever is left over of Y_1 for inference.

Procedure Carve_{100} is inference after the using the lasso on the full data set Y.

For the data carving procedures, we use the selected-model z-test of Section 5.2 after lasso variable selection using Lagrange parameter

$$\lambda = 2\mathbb{E}(\|X^T\epsilon\|_\infty), \quad \epsilon \sim N(0, \sigma^2 I)$$

as described in (Negahban et al., 2012). In addition, we condition on the signs of the active lasso coefficients, so that procedure Carve_{100} is the inference-after-lasso test considered in Lee et al. (2013).[6]

We know from Theorem 10 that procedure Carve_{n_1} strictly dominates procedure Split_{n_1} for any n_1, but there is a selection–inference tradeoff between data-carving procedures Carve_n and Carve_{n_1} for $n_1 < n$. Carve_n uses all of the data for selection, and is therefore likely to select a superior model, whereas procedure Carve_{n_1} reserves more power for the second stage.

Let R be the size of the model selected and V the number of noise variables included. We compare the procedures with respect to aspects of their selection performance:

- chance of screening, i.e. obtaining a correct model ($\mathbb{P}(R - V = 7)$ or p_{screen}).
- expected number of noise variables selected ($\mathbb{E}[V]$),
- expected number of true variables selected ($\mathbb{E}[R - V]$),
- false discovery rate of true variables selected ($\mathbb{E}[V/\max(R, 1)]$ or FDR),

Conditional on having obtained a correct model, we also compare them on aspects of their second stage performance:

- probability of correctly rejecting the null for one of the true variables (Power),
- probability of incorrectly rejecting the null for a noise variable (Level).

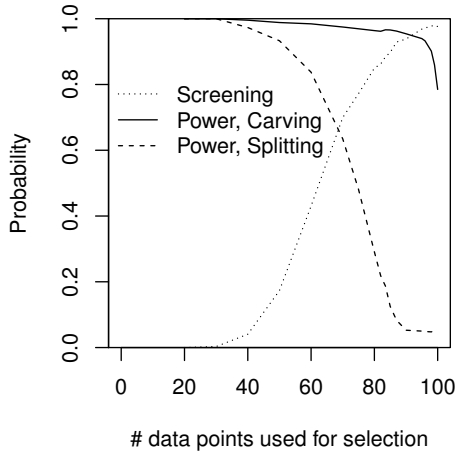
[6]Because of the form of the selection event when we use the lasso after n data points, the test statistic is conditionally independent of $\mathcal{P}^{\perp}_{X_M} Y$. Thus, there is no distinction between the saturated- and selected-model z-tests after the lasso on all n data points.

Algorithm	p_{screen}	$\mathbb{E}[R-V]$	$\mathbb{E}[R]$	FDR	Power	Level
Carve_{100}	0.98	6.97	15.30	0.52	0.79	0.05
Split_{50}	0.17	5.21	14.19	0.62	0.93	0.05
Carve_{50}	0.17	5.21	14.19	0.62	0.99	0.05
Split_{75}	0.77	6.72	15.58	0.54	0.48	0.05
Carve_{75}	0.77	6.72	15.58	0.54	0.97	0.05

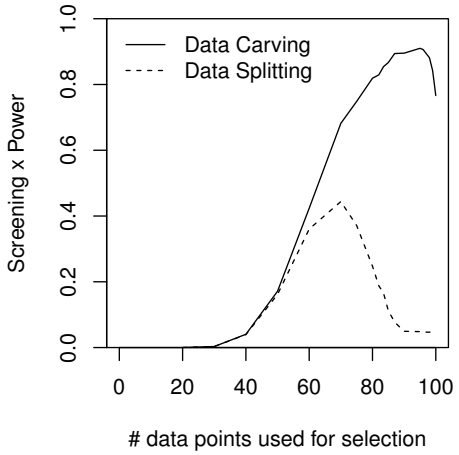
Table 1: Simulation results. p_{screen} is the probability of successfully selecting all 7 true variables, and Power is the power, conditional on successful screening, of tests on the true variables. The more data we use for selection, the better the selected model's quality is, but there is a cost in second-stage power. Carve_{75} appears to be finding a good tradeoff between these competing goals. Carve_{n_1} always outperforms Split_{n_1}, as predicted by Theorem 10.

Algorithm	p_{screen}	$\mathbb{E}[R-V]$	$\mathbb{E}[R]$	FDR	Power	Level
Carve_{100}	0.98	6.97	15.21	0.51	0.79	0.05
Split_{50}	0.18	5.25	14.19	0.61	0.93	0.05
Carve_{50}	0.18	5.25	14.19	0.61	0.99	0.06

Table 2: Simulation results under misspecification. Here, errors ϵ are drawn independently from Student's t_5. Our conclusions are identical to Table 1.



(a) Probability of successful screening, and power conditional on screening, for Split_{n_1} and Carve_{n_1}.



(b) Probability of successful screening times power conditional on screening, for Split_{n_1} and Carve_{n_1}.

Figure 7: Tradeoff between power and model selection. As n_1 increases and more data is used in the first stage, we have a better chance of successful screening (picking all the true nonzero variables). However, increasing n_1 also leads to reduced power in the second stage. Data splitting suffers much more than data carving, though both are affected.

The results, shown in Table 1, bear out the intuition of Section 3.2. Because procedure Carve_{100} uses the most information in the first stage, it performs best in terms of model selection, but pays a price in lower second-stage power relative to Split_{50} or Carve_{50}. The procedure Carve_{50} clearly dominates Split_{50}, as expected. Increasing n_1 from 50 to 75 improves p_{screen} for Split_{75}, but Split_{75} suffers a drop in power, placing it roughly equivalent to Carve_{100} in the second stage. Procedure Carve_{75} seems to strike a better compromise.

Figure 7 shows the tradeoff curve of model selection success (as measured by the probability of successful screening) against second-stage power conditional on successful screening. As n_1 increases, stage-one performance improves while stage-two performance declines, but the decline is much slower for data carving. Surprisingly, Carve_{98} and Carve_{99} have much higher power than Carve_{100}: 90%, 86%, and 79% respectively. We cannot explain why holding out just one or two data points in the first stage improves power so dramatically. Better understanding this tradeoff is an interesting topic of further work.

Finally, to check the robustness of data carving, we replace the Gaussian errors with independent errors drawn from Student's t distribution with five degrees of freedom. The numbers barely change at all; see Table 2. Tian and Taylor (2014) rigorously analyze the case of non-Gaussian errors.

8 Selective Inference and Multiple Inference

A common approach to the problem of inference after selection is to replace the nominal error rate with an alternative joint error rate that is deemed appropriate to the scientific setting.

For example, suppose that $\theta_q,\ q = 1, \ldots, m$ correspond to parameters of a common model M. We designate a small number $R(Y) = |\widehat{\mathcal{Q}}(Y)|$ of them as interesting and construct a confidence interval $C_q(Y)$ for each $q \in \widehat{\mathcal{Q}}$. Benjamini and Yekutieli (2005) propose controlling the *false*

coverage-statement rate (FCR)

$$\mathbb{E}\left[\frac{V}{\max(R,1)}\right], \quad \text{where} \quad V(Y) = \left|\left\{q:\, q \in \widehat{\mathcal{Q}},\ \theta_q(F) \notin C_q(Y)\right\}\right| \tag{86}$$

is the number of non-covering intervals constructed.

FCR control is closely related to selective coverage. By choosing appropriate selection variables S_q, we can adapt selective coverage to control the FCR.

Proposition 11 (FCR Control via Selective Error Control)**.** *Assume $\mathcal{Q}$ is countable with each $q \in \mathcal{Q}$ corresponding to a different parameter θ_q for the same model M. Let $R(Y) = |\widehat{\mathcal{Q}}(Y)|$ with $R(Y) < \infty$ a.s., and define $V(Y)$ as in (86).*

If each C_q enjoys coverage at level $1-\alpha$ given $S_q = \big(\mathbf{1}_{A_q}(Y), R(Y)\big)$, then the collection of intervals $(C_q, q \in \widehat{\mathcal{Q}})$ controls the FCR at level α:

$$\mathbb{E}\left[\frac{V}{\max(R,1)}\right] \;\leq\; \mathbb{E}\left[\frac{V}{R} \;\middle|\; R \geq 1\right] \;\leq\; \alpha. \tag{87}$$

Proof. Let $V_q(Y) = \mathbf{1}\left\{q \in \widehat{\mathcal{Q}}(Y),\, \theta_q(F) \notin C_q(Y)\right\}$, so that $V = \sum_{q\in\mathcal{Q}} V_q$. For $R \geq 1$, and for any $F \in M$,

$$\mathbb{E}_F\left[V \,|\, R\right] \;=\; \sum_{q\in\mathcal{Q}} \mathbb{E}_F\left[V_q \,|\, R\right] \;\leq\; \sum_{q\in\mathcal{Q}} \alpha\, \mathbb{E}_F\left[\mathbf{1}_{A_q}(Y) \,|\, R\right] \;=\; \alpha R, \tag{88}$$

hence $\mathbb{E}\left[V/R \,|\, R\right] = \alpha$ for each $R \geq 1$. □

Other authors have addressed inference after selection by proposing to control the FWER, the chance that any selected test incorrectly rejects the null or any constructed confidence interval fails to cover its parameter. For example, the "post-selection inference" (PoSI) method of Berk et al. (2013) constructs simultaneous $(1-\alpha)$ confidence intervals for all parameters of all linear regression models that were ever under consideration. As a result, no matter how we choose the model, the overall probability of constructing any non-covering interval is controlled at α.[7]

Selective error control can be adapted to control the FWER as well. If our selection rule always chooses a single hypothesis to test, then we have overall FWER control.

Proposition 12 (FWER Control for Singleton $\widehat{\mathcal{Q}}$)**.** *Assume that $|\widehat{\mathcal{Q}}(Y)| \overset{a.s.}{=} 1$, and let $Q(Y) = (M(Y), H_0(Y))$ denote the single (random) selected model–hypothesis pair. If each ϕ_q controls the selective error at level α, then the test $\phi(y) = \phi_{Q(y)}(y)$ controls the FWER at level α:*

$$\mathbb{E}_F\left[\phi(Y);\, F \in H_0(Y)\right] \leq \alpha \tag{89}$$

Proof. Condition on Q:

$$\mathbb{E}_F\left[\phi(Y)\mathbf{1}\{F \in H_0(Y)\}\right] = \mathbb{E}_F\left[\,\mathbb{E}_F\left[\phi(Y) \,|\, Q(Y)\right]\mathbf{1}\{F \in H_0(Y)\}\,\right] \tag{90}$$

$$\leq \alpha\, \mathbb{P}_F\left[F \in H_0(Y)\right] \tag{91}$$

□

More generally, it is clear that if we construct $(\phi_q, q \in \mathcal{Q})$ to control any joint error rate conditional on the entire selected set $\widehat{\mathcal{Q}}$, we will also have marginal control of the same joint error rate.

[7] Technically, PoSI constructs simultaneous intervals for *least-squares parameters* as defined in Section 5, which correspond to linear functionals of $\mu = \mathbb{E}Y$.

However, the converse of Proposition 12 is not true: FWER control does *not* in general guarantee control of relevant selective error rates. For example, suppose $Q(Y) = 1$ with probability 0.9 and $Q(Y) = 2$ otherwise. If ϕ_1 and ϕ_2 have selective error rates $\alpha_1 = 0.02$ and $\alpha_2 = 0.3$ respectively, the overall FWER is still controlled at $\alpha = 0.05$.

Does our conservatism when asking question 1 compensate for our anti-conservatism when asking question 2? To answer this question we must consider not only mathematics but also the relevant scientific context. If the different questions represent a bag of anonymous, *a priori* undifferentiated hypotheses, then a joint error rate like the FWER or FDR may be a good proxy for our scientific goals. For example, when performing a large-scale genome-wide "fishing expedition" for loci associated with type II diabetes, the fraction of null genes among all purported discoveries is a very relevant quantity: it measures what fraction of our time and money will be wasted following up on false leads.

In other scientific applications, however, different hypotheses have quite distinct identities and may vary greatly in their importance and interpretation. For example, a confidence interval for the effect of gender on salary after controlling for one's job title may be much more socially consequential than an interval for the effect of job title after controlling for gender. As such, averaging our error rates across the two questions is inappropriate.

9 Discussion

Selective inference concerns the properties of inference carried out after using a data-dependent procedure to select which questions to ask. We can recover the same long-run frequency properties among answers to *selected* questions that we would obtain in the classical non-adaptive setting, if we follow the guiding principle of selective error control:

The answer must be valid, given that the question was asked.

Happily, living up to this principle can be a simple matter in exponential family models including linear regression, due to the rich classical theory of optimal testing in exponential family models. Even if we are possibly selecting from a large menu of diverse and incompatible models, we can still design tests one model at a time and control the selective error using the test designed for the selected model. We generally pay a price for conditioning, so it is desirable to condition on as little as possible. Data carving can dramatically improve on data splitting by using the leftover information in Y_1, the data set initially designated for selection.

Many challenges remain. Deriving the cutoffs for sample carving tests can be computationally difficult in general. In addition, the entire development of this article takes the model selection procedure $\widehat{\mathcal{Q}}$ as given, when in reality we can choose $\widehat{\mathcal{Q}}$. More work is needed to learn what model selection procedures lead to favorable second-stage properties.

As data sets and research questions become more and more complex, we have less and less hope of specifying adequate probabilistic models ahead of time. As such, a key challenge of complex research is to balance the goal of choosing a realistic model against the goal of inference once we have chosen it. We hope that the ideas in this article represent a step in the right direction.

Acknowledgements

William Fithian was supported by National Science Foundation VIGRE grant DMS-0502385 and the Gerald J. Lieberman Fellowship. Dennis Sun was supported in part by the Stanford Genome Training Program (NIH/NHGRI T32 HG000044) and the Ric Weiland Graduate Fellowship. Jonathan Taylor was supported in part by National Science Foundation grant DMS-1208857

and Air Force Office of Sponsored Research grant 113039. We would like to thank Stefan Wager, Trevor Hastie, Rob Tibshirani, Brad Efron, Maxwell Grazier G'sell, and Yuval Benjamini for helpful discussions.

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

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.

Julian Besag and Peter Clifford. Generalized monte carlo significance tests. *Biometrika*, 76(4): 633–642, 1989.

Allan Birnbaum. Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. *The Annals of Mathematical Statistics*, pages 21–36, 1955.

Lawrence D Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages i–279, 1986.

DR Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional Inference: Confidence intervals, p-values and R-Software hdi. *ArXiv e-prints*, August 2014.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron, Robert Tibshirani, et al. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

Jonathan J Forster, John W McDonald, and Peter WF Smith. Monte carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 445–453, 1996.

R. Foygel Barber and E. Candes. Controlling the False Discovery Rate via Knockoffs. *ArXiv e-prints*, April 2014.

Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: unlocking the file drawer. *Science*, 2014.

Max Grazier G'Sell, Jonathan Taylor, and Robert Tibshirani. Adaptive testing for the graphical lasso. *arXiv preprint arXiv:1307.4765*, 2013.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv preprint arXiv:1301.4240*, 2013.

George Johnson. New truths that only one can see. *The New York Times*, 2014.

Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. *arXiv preprint arXiv:1402.5596*, 2014.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference with the lasso. *arXiv preprint arXiv:1311.6238*, 2013.

EL Lehmann and Joseph P Romano. *Testing statistical hypotheses*. New York:. Springer, 2005.

EL Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(3):219–236, 1955.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso (with discussion). *The Annals of Statistics*, 42(2):413–468, 2014.

Joshua R Loftus and Jonathan E Taylor. A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*, 2014.

Ted K Matthes and Donald R Truax. Tests of composite hypotheses for the multivariate exponential family. *The Annals of Mathematical Statistics*, pages 681–697, 1967.

Cyrus R Mehta, Nitin R Patel, and Pralay Senchaudhuri. Efficient monte carlo methods for conditional logistic regression. *Journal of The American Statistical Association*, 95(449):99–108, 2000.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of MM-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, November 2012. ISSN 0883-4237. doi: 10.1214/12-STS400. URL `http://projecteuclid.org/euclid.ss/1356098555`.

Richard A Olshen. The conditional level of the ftest. *Journal of the American Statistical Association*, 68(343):692–698, 1973.

Ari Pakman and Liam Paninski. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.

Camilo Rivera and Guenther Walther. Optimal detection of a jump in the intensity of a poisson process or in a density with likelihood ratio statistics. *Scandinavian Journal of Statistics*, 40(4): 752–769, 2013.

Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, 2014.

Xiaoying Tian and Jonathan E Taylor. Affine selection procedures with non-gaussian errors. *in preparation*, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Asaf Weinstein, William Fithian, and Yoav Benjamini. Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108(501): 165–176, 2013.

Ed Yong. Replication studies: Bad copy. *Nature*, 485(7398):298–300, 2012.

A Proof of Proposition 1

Proof. For group i, let R_i be the number of true nulls selected and let V_i denote the number of false rejections. If $Z_n^V = \sum_{i=1}^n V_i$ and $Z_n^R = \sum_{i=1}^n R_i$, then we need to show $\limsup_{n\to\infty} Z_n^V/Z_n^R \le \alpha$.

By design, $0 \le V_i \le R_i$ and $\mathbb{E}(V_i) \le \alpha\,\mathbb{E}(R_i)$. As a result, $\mathbb{E}[Z_n^V]/\mathbb{E}[Z_n^R] \le \alpha$ for every n, so we just need to show that the two sums are not far from their expectations. Because

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(R_i)}{i^2} \le B \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty,$$

we can apply Kolmogorov's strong law of large numbers to the independent but non-identical sequence $R_1, R_2, \ldots$ to obtain

$$\frac{1}{n}(Z_n^R - \mathbb{E}Z_n^R) \xrightarrow{a.s.} 0, \quad \text{so} \quad \left|\frac{Z_n^R}{\mathbb{E}Z_n^R} - 1\right| \le \left|\frac{\delta}{n}(Z_n^R - \mathbb{E}Z_n^R)\right| \xrightarrow{a.s.} 0.$$

As for Z_n^V, we have

$$\frac{1}{n}(Z_n^V - \mathbb{E}Z_n^V) \xrightarrow{a.s.} 0, \quad \text{so} \quad \frac{Z_n^V}{\mathbb{E}Z_n^R} - \alpha \le \frac{\delta}{n}(Z_n^V - \mathbb{E}Z_n^V) \xrightarrow{a.s.} 0;$$

in other words, $Z_n^R/\mathbb{E}Z_n^R \xrightarrow{a.s.} 1$ and $\limsup_n Z_n^V/\mathbb{E}Z_n^R \overset{\text{a.s.}}{\le} \alpha$. □

B Proof of Theorem 10

Proof. Let $T = T_1 + T_2$, $U = U_1 + U_2$. If ϕ is admissible then

$$\phi(Y) \in m\mathscr{F}(T, U, A) \subseteq m\mathscr{F}(T, U_1, U_2, A). \tag{92}$$

Given $(A, U_1 = u_1, U_2 = u_2)$, the density of (Y_1, Y_2) simplifies to

$$Y \sim \exp\left\{\theta\,\big(T_1(y_1) + T_2(y_2)\big) - \psi(\theta)\right\}\ \mathbf{1}_A(y_1) \prod_{i=1}^{2} f_{0,i}(y_i \,|\, u_i). \tag{93}$$

T_1 and T_2 are conditionally independent given (U_1, U_2, A) because Y_1 and Y_2 are.

If T_1 is not a function of U_1 on A then there exist $\delta > 0$ and real-valued τ for which $\min \mathbb{P}(A^-), \mathbb{P}(A^+) > 0$ where

$$A^- = \{T_1 < \tau(U_1) - \delta\} \cap A, \quad \text{and} \tag{94}$$

$$A^+ = \{T_1 > \tau(U_1) + \delta\} \cap A. \tag{95}$$

This probability depends on (θ, ζ) but if it is positive for one (θ, ζ) it is positive for all.

If ϕ is admissible among level α tests given Y_1, it must have acceptance regions of the form

$$1 - \phi(Y) = \mathbf{1}\left\{c_1(U_2) \le T_2 \le c_2(U_2)\right\} \tag{96}$$

for some (possibly infinite) cutoffs c_1, c_2. Because $\alpha > 0$, we must have at least one of the $c_i \in (a, b)$ with positive probability. Note we can always replace T_i with $-T_i$; thus, without loss of generality assume

$$\mathbb{P}\left(c_2(U_2) < b\right) > 0 \tag{97}$$

Define the event

$$B = \{c_2(U_2) < b\} \bigcap \{c_2(U_2) - c_1(U_2) > \varepsilon\} \tag{98}$$

which has positive probability for some $\varepsilon \in (0, \delta)$.

Moreover, T_2 must place some mass near the cutoff on each side, so that $\min \mathbb{P}(B^-), \mathbb{P}(B^+) > 0$ where

$$B^- = \{c_2(U_2) - \varepsilon < T_2 < c_2(U_2)\} \cap B, \quad \text{and} \tag{99}$$

$$B^+ = \{c_2(U_2) < T_2 < c_2(U_2) + \varepsilon\} \cap B. \tag{100}$$

Notice that

$$A^- \cap B^- \Rightarrow T - c_2 - \tau \ \in \ (-\delta - \varepsilon, \ -\delta), \tag{101}$$

$$A^- \cap B^+ \Rightarrow T - c_2 - \tau \ \in \ (-\delta, \ -\delta + \varepsilon), \tag{102}$$

$$A^+ \cap B^- \Rightarrow T - c_2 - \tau \ \in \ (\delta - \varepsilon, \ \delta), \tag{103}$$

$$A^+ \cap B^+ \Rightarrow T - c_2 - \tau \ \in \ (\delta, \ \delta + \varepsilon), \tag{104}$$

corresponding to four disjoint intervals of increasing value of T given (U_1, U_2). Furthermore, note that $\phi(Y)$ takes values $0, 1, 0, 1$ on the four respective events. This fact rules out the possibility that the acceptance region of ϕ has convex (U_1, U_2)-sections. □

C Monte Carlo Tests and Confidence Intervals: Details

Assume Z arises from a one-parameter exponential family

$$Z \sim g_\theta(z) = e^{\theta z - \psi(\theta)}\, g_0(z). \tag{105}$$

We wish to compute (by Monte Carlo) the UMPU two-sided rejection region for the hypothesis $H_0 : \theta = \theta_0$. Let $U \sim \text{Unif}[0, 1]$ be an auxiliary randomization variable.

Define the dictionary ordering on $[0, 1]$:

$$(z_1, u_1) \prec (z_2, u_2) \iff z_1 < z_2 \text{ or } (z_1 = z_2 \text{ and } u_1 < u_2). \tag{106}$$

If $\Gamma_1 = (c_1, \gamma_1)$ and $\Gamma_2 = (c_2, 1 - \gamma_2)$, then the region

$$R_{\Gamma_1, \Gamma_2} = \{(z, u) : (z, u) \prec \Gamma_1 \text{ or } (z, u) \succ \Gamma_2\} \tag{107}$$

implements the rejection region for the test with cutoffs c_1, c_2 and boundary randomization parameters γ_1, γ_2.

For $\Gamma_1 \prec \Gamma_2$, write

$$K_1(\Gamma_1, \Gamma_2; \theta) = \mathbb{P}_\theta(R_{\Gamma_1, \Gamma_2}) - \alpha \tag{108}$$

$$K_2(\Gamma_1, \Gamma_2; \theta) = \mathbb{E}_\theta(Z \,|\, (Z, U) \in R^C_{\Gamma_1, \Gamma_2}) - \mathbb{E}_\theta(Z), \tag{109}$$

so that the correct cutoffs Γ_i are those for which $K_1(\Gamma_1, \Gamma_2; \theta) = K_2(\Gamma_1, \Gamma_2; \theta) = 0$. For fixed θ, K_1 is decreasing in Γ_1 and increasing in Γ_2, while K_2 is increasing in both Γ_1 and Γ_2.

Let $(Z_1, W_1), (Z_2, W_2), \ldots$ be a sequence of random variables for which

$$\widehat{\mathbb{E}}_\theta^n h(Z) = \frac{\sum_{i=1}^n W_i h(Z_i) e^{\theta Z_i}}{\sum_{i=1}^n W_i e^{\theta Z_i}} \tag{110}$$

$$\xrightarrow{a.s.} \mathbb{E}_\theta h(Z). \tag{111}$$

for all integrable h. This would be true if (Z_i, W_i) are a valid i.i.d. sample or i.i.d. importance sample from g_0, or if they come from a valid Markov Chain Monte Carlo algorithm.

If $\widehat{K}_i^n$ are defined analogously to K_i for $i = 1, 2$, with $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ replaced with their importance-weighted empirical versions $\widehat{\mathbb{E}}_\theta^n$ and $\widehat{\mathbb{P}}_\theta^n$, then $\widehat{K}_i^n \xrightarrow{a.s.} K$ pointwise as $n \to \infty$, and $\widehat{K}_i^n$ satisfy the same monotonicity properties almost surely for each n. As a result, we have almost sure convergence on compacta for $(\widehat{K}_1^n, \widehat{K}_2^n)$:

$$\sup_{(\Gamma_1, \Gamma_2) \in G} \max_i \left\| \widehat{K}_i^n(\Gamma_1, \Gamma_2; \theta) - K_i(\Gamma_1, \Gamma_2; \theta) \right\| \tag{112}$$

for each θ, for compact $G \in (\mathbb{R} \times [0, 1])^2$.

We carry out our tests by solving for Γ_1 and Γ_2 which solve $\widehat{K}_1^n$ and $\widehat{K}_2^n$, in effect defining the UMPU tests for a one-parameter exponential family through the approximating empirical measure. Specifically, we can define

$$\widehat{\Gamma}_2(\Gamma_1; \theta) = \inf \left\{ \Gamma_2 : \widehat{K}_1^n(\Gamma_1, \Gamma_2; \theta) = 0 \right\}, \tag{113}$$

with $\widehat{\Gamma}_2 = \infty$ if the set is empty. That is, for a given lower cutoff we define the upper cutoff to obtain a level-α acceptance region if that is possible. Then, $\widehat{K}_2^n \left(\Gamma_1, \widehat{\Gamma}_2(\Gamma_1; \theta); \theta \right)$ is an increasing function and we can solve it using binary search. Let $\widehat{R}_\theta$ denote the rejection region so obtained.

Note that (z, u) is in the left-tail of $\widehat{R}_\theta$ if and only if $\widehat{K}_2^n \left((z, u), \widehat{\Gamma}_2((z, u)); \theta \right) < 0$. This fact, paired with an analogous test for whether (z, u) is in the right tail, gives us a quick way to carry out the test. It also allows us to quickly find the upper and lower confidence bounds for the approximating empirical family, via binary search.

D Sampling for the Selective t-Test: Details

Let $C \subseteq \mathbb{R}^k$ denote a set with nonempty interior and consider the problem of integrating some integrable function $h(y)$ against the uniform probability measure on $C \cap S^{k-1}$, where S^{k-1} is the unit sphere of dimension $k - 1$, assuming the intersection is non-empty. Assume we are given an i.i.d. sequence of uniform samples $Y_1, Y_2, \ldots$ from $C \cap B^k$, where B^k is the unit ball.

Let $R \sim \frac{r^{k-1}}{k}$, so that if $Z \sim \text{Unif}(S^{k-1})$, then $Y = RZ \sim \text{Unif}(B^k)$. Let

$$W(Z) = \left(\int_0^1 \mathbf{1}\{rZ \in C\} \, \frac{r^{k-1}}{k} \, dr \right)^{-1} \tag{114}$$

We can use the Y_i for which $Z_i = Y_i / \|Y_i\| \in C$ as a sequence of importance samples with

weights $W(Z_i)$, since

$$\mathbb{E}(h\,(Z)\mathbf{1}\,\{Y, Z \in C\}\,W(Z)) \tag{115}$$

$$= \int_{S^{k-1}} \int_0^1 h(z)\mathbf{1}\,\{z, rz \in C\}\,W(z)\frac{r^{k-1}}{k}\,dr\,dz \tag{116}$$

$$= \int_{S^{k-1}} h(z)\mathbf{1}\,\{z \in C\}\;dz \tag{117}$$

$$= \mathbb{E}\,(h(Z)\mathbf{1}\,\{Z \in C\})\,. \tag{118}$$

To carry out the selective t-test of $H_0 :\ \beta_j = 0$, we need to sample from

$$\mathcal{L}\left(\eta' Y \;\middle|\; \mathcal{P}_{X_{M\setminus j}} Y,\ \ \|Y\|,\ \ A\right). \tag{119}$$

Let $U = \mathcal{P}_{X_{M\setminus j}} Y$, and let $Q \in \mathbb{R}^{n\times(n-|M|-1)}$ be such that $QQ' = \mathcal{P}^{\perp}_{X_{M\setminus j}}$. Then $L^2 \triangleq \|Q'Y\|^2 = \|Y\|^2 - \|U\|^2$ is fixed under the selection event. Let

$$C = \{v :\ U + Qv \in A\}, \tag{120}$$

so that $A_U = U + QC$, an $(n - |M| - 1)$-dimensional hyperplane intersected with A, is the event we would sample from for the selective z-test.

Under H_0, Y is uniformly distributed on

$$(U + QC) \cap \|Y\|S^{n-1} = U + Q\left(C \cap LS^{n-|M|-2}\right). \tag{121}$$

Assume we can resample Y^* uniformly from $A_U \cap (U + LB^{n-|M|-1})$, which is just sampling from A_U with an additional quadratic constraint. Then $V^* = Q'(Y^* - U)$ is a sample from the ball of radius L, intersected with C. We can turn V^* into an importance-weighted sample from the sphere via the scheme outlined above; then, the same importance weight suffices to turn Y^* into a sample from the selective t-test conditioning set.