Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 3 (3005–3024)

A SELECTIVE SURVEY OF SELECTIVE INFERENCE

Jonathan E. Taylor

The idea of a scientist, struck, as if by lightning with a question, is far from the truth. – Tukey [1980].

Abstract

It is not difficult to find stories of a crisis in modern science, either in the popular press or in the scientific literature. There are likely multiple sources for this crisis. It is also well documented that one source of this crisis is the misuse of statistical methods in science, with the P-value receiving its fair share of criticism. It could be argued that this misuse of statistical methods is caused by a shift in how data is used in 21st century science compared to its use in the mid-20th century which presumed scientists had formal statistical hypotheses before collecting data. With the advent of sophisticated statistical software available to anybody this paradigm has been shifted to one in which scientists *collect data first and ask questions later*.

1 The new (?) scientific paradigm

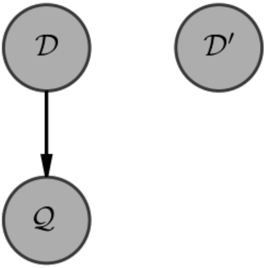


Figure 1: A simplified version of the scientific method.

We are all familiar with a paradigm that does allow scientists to collect data first and ask questions later: the classical scientific method illustrated in Figure 1. A scientist collects data $\mathfrak{D}$, generates questions of interest $\mathfrak{Q}(\mathfrak{D})$, then collects fresh data $\mathfrak{D}'$ for confirmation and perhaps to discover additional questions of interest. The problem with this new paradigm is that it seeks to use $\mathfrak{D}$ to answer these questions and may not have access to $\mathfrak{D}'$.

We pause here to note that Tukey used the term *question* rather than the more precise term *hypothesis* which statisticians might reasonably impute to be a statistical hypothesis. Given the computing capabilities of modern

MSC2010: primary 62-02; secondary 62J15, 62J05, 62-07.

statistical software, it is not really clear that data analysis produces statistical hypotheses:

> In practice, of course, hypotheses often emerge after the data have been examined; patterns seen in the data combine with subject-matter knowledge in a mix that has so far defied description. – P. Diaconis "Theories of Data Analysis" [n.d.]

We continue to distinguish a question Q from a *statistical object* such as a hypothesis test, i.e. a pair $(\mathfrak{M}, H_0)$ with $\mathfrak{M}$ a statistical model (a collection of distributions on some measurable space) and $H_0 \subset \mathfrak{M}$; or a pair $(\mathfrak{M}, \theta)$ with $\theta : \mathfrak{M} \rightarrow \mathbb{R}^k$ a parameter for which we might form a region or point estimate. An example of a Bayesian statistical might be a triple $(\pi, \ell, \mathcal{T})$ with π a prior, ℓ a likelihood and $\mathcal{T}$ some functional of the posterior. The transformation from questions to statistical objects is up to the scientist, perhaps in partnership with a statistician.

Returning to the new paradigm in science, whether the statistics community feel that this is the correct way to run experiments and advance a particular field of science, it is difficult to ignore the fact that it is how (at least some) modern science is practiced. We feel it is imperative to provide scientists with tools that provide some of the guarantees of the classical methods but are applicable in this new paradigm. These are the problems that the area of *selective inference* attempts to address. The term *selective* refers to the fact that the results reported in a scientific study (e.g. P-values, confidence intervals) are selected through some mechanism guided by the scientist. When the mechanism of selection is known, it is often possible to mitigate this selection bias.

1.1 Two prototypical settings with many questions. We describe two prototypical problems occurring in many modern scientific disciplines, from genomic studies to neuroscience and many others. Both involve a response $y \in \mathbb{R}$ (which we take to be real-valued simply for concreteness) and a set of features $X \in \mathbb{R}^p$.

1.1.1 Large scale inference. Often of interest are the p questions

$$Q_j^L : \text{Is feature } j \text{ associated with outcome } y? \qquad 1 \leq j \leq p \tag{1}$$

This problem is often referred to as *large-scale inference* Efron [2012] (L for large) and has brought about a renewed interest in empirical Bayes methodology and multiple comparisons in general.

A canonical experimental design in this problem samples n pairs IID from some law F in a statistical model $\mathfrak{M}$. Having these pre-determined set of questions allows the statistician, given the model $\mathfrak{M}$, to transform each question to a parameter in model $\mathfrak{M}$:

$Q_j^L \mapsto \theta_j^L \in \mathbb{R}^{\mathfrak{M}}$ where θ_j^L measures a marginal association between y and feature j on $\mathfrak{M}$. Parameters in hand, the statistician can then use the formal methods of statistical inference to "answer" these questions.

1.1.2 Feature selection in regression. The second problem also involves response y and features X, though in this case the scientist seeks to build a predictive model of y from X. At first glance, the natural questions are

(2) $\quad Q_j^R$: Is feature j important when trying to predict y from X? $\quad 1 \leq j \leq p.$

As is clear to any student after a course in linear regression (R for regression) the above questions are ill posed. This point is emphasized in Berk, Brown, Buja, K. Zhang, and Zhao [2013] which then posed the following questions

(3) $\quad \bar{Q}^R_{j|E}$: Is feature j correlated with the residual when trying predict y from $X_{E\setminus j}$?

Such questions are indexed by $j \in E, E \subset \{1, \dots, p\}$. These questions are also posed before data collection as soon as the scientist decided to collect these p features to build a predictive model for y from X.

These two problems have inspired much work in selective inference: with the large scale inference problem drawing intense focus in the early part of the 21st century Efron [2012] and Storey [2003] building on the seminal work of Benjamini and Hochberg [1995]. The regression problem is an area of more recent interest Berk, Brown, Buja, K. Zhang, and Zhao [2013], Hurvich and Tsai [1990], Lee, D. L. Sun, Y. Sun, and J. E. Taylor [2016], Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani [2014], and Wasserman and Roeder [2009] perhaps due to the added complexity of the set of possible questions under consideration.

2 Selective inference

We now describe some proposed methods to address this new scientific paradigm. The first set of methods, based on multiple comparisons, generally ignore the possible existence of $\mathfrak{D}'$ (formally equivalent to setting $\mathfrak{D}' = 0$, a constant random variable) while the conditional methods (of which the classical scientific method is one) do acknowledge that $\mathfrak{D}$ likely does not exist in a vacuum. A scientist probably will have run previous experiments and will (subject to restrictions) run future experiments.

2.1 Multiple comparisons and simultaneous inference. Some of the earliest references to selective inference Benjamini [2010] and Benjamini and Yekutieli [2005] come from the field of *multiple comparisons* in the large scale inference problem, particularly through the extraordinarily influential work of Benjamini and Hochberg [1995] and its introduction of the False Discovery Rate (FDR) as a more liberal error rate than the Family Wise Error Rate (FWER).

The goal in multiple comparisons procedures is to construct procedures T that control an error rate such as the FDR or FWER (defined below) over some statistical model. For concreteness, suppose that in the large scale inference problem we have access to an estimate Z_j of association θ_j between feature j and response y we might take $\mathfrak{D} = Z$ and our statistical model to be

$$\mathfrak{M}^L = \{N(\theta, \Sigma) : \theta \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p} \geq 0, \mathrm{diag}(\Sigma) = 1\} \tag{4}$$

with Σ known or unknown (but hopefully estimable). In this problem, a multiple comparisons procedure is a map $T : \mathbb{R}^p \to \{0,1\}^p$ that makes a decision whether each hypothesis $H_{0,j}$ is true or false. A procedure T that controls the FWER at level α satisfies

$$FWER(T, F) = \mathbb{P}_F\left(V(T, Z) > 0\right) \leq \alpha, \qquad \forall F \in \mathfrak{M}^L \tag{5}$$

with

$$V(T, Z) = \left\{j : \theta_j \neq 0, T_j(Z) = 1\right\} \tag{6}$$

the number of false positives the procedure T selected on outcomes Z where $T_j(Z) = 1$ signals a positive decision, i.e. that $H_{0,j}$ is false. When $p = 1$ and only one hypothesis is under consideration, $FWER$ reduces to Type I error, 0 for any $F \notin H_0$. A test that controls the Type I error at level α satisfies

$$\mathbb{P}_F(T(Z) = 1) \leq \alpha, \qquad \forall F \in H_0. \tag{7}$$

The FDR of procedure T is also expressible as an expectation under the law F. Note that controlling FDR or FWER are *marginal* properties of each $F \in \mathfrak{M}^L$. This will be contrasted below with *conditional* properties.

The prototypical example of a procedure that controls the FWER is the Bonferroni procedure which uses a simple bound on the law of the largest $(Z_j)_{1 \leq j \leq p}$:

$$\mathbb{P}_F\left(\max_{1 \leq j \leq p} |Z_j - \theta_j| > t\right) \leq p \cdot \bar{\Phi}(t), \qquad \forall F \in \mathfrak{M}^L \tag{8}$$

with $\bar{\Phi}$ the tail of a standard normal random variable. Tighter approximations of the left-hand valid over some $\mathfrak{M}$ can be used to get an improvement over Bonferroni. This area of research is sometimes referred to as *simultaneous inference*. The late 20th century saw its own golden era in research in this area Adler and J. E. Taylor [2007], Azaïs and Wschebor [2009], Siegmund and Worsley [1995], J. Sun [1993], and Takemura and Kuriki [2002] in which the feature space was modelled as approximating a continuum with Σ a representation of the covariance function of some Gaussian process.

It is well known that bounds for the left hand side of (8) translate to *(simultaneous) coverage guarantees* for confidence intervals for the θ_j. For example, suppose t is such that some bound for the left hand side of (8) is less than α. Then,

$$\mathbb{P}_F\left(\exists j : \theta_j \notin [Z_j - t, Z_j + t]\right) \leq \alpha, \qquad \forall F \in \mathfrak{M}^L. \tag{9}$$

This simultaneous approach was considered in Berk, Brown, Buja, K. Zhang, and Zhao [2013]. Formally (considering X fixed) the authors set

$$\mathfrak{M}^{\text{POSI}} = \{N(\mu, I_{n\times n}) : \mu \in \mathbb{R}^n\}. \tag{10}$$

and transform the questions $\bar{Q}^R_{j|E} \mapsto \theta^R_{j|E} \in (\mathbb{R}^n)^*$ to parameters taken to be the linear functionals

$$\theta^R_{j|E}(\mu) = e_j^T (X_E^T X_E)^{-1} X_E^T \mu \tag{11}$$

where $e_j : \mathbb{R}^E \to \mathbb{R}$ is projection onto the j coordinate. The authors of Berk, Brown, Buja, K. Zhang, and Zhao [ibid.] then note that $\mathfrak{M}^{\text{POSI}}$ can be embedded in a model of the form $\mathfrak{M}^L$ indexed by (j, E) with $j \in E$ and $E \subset \{1, \ldots, p\}$ and taking $\mathfrak{D}$ to be (some subset of) the collection of corresponding Z statistics. The authors propose finding a bound better than Bonferroni using simulation.

We end this section with *knockoffs* Barber and Candes [2014], a different approach to the regression problem within the framework of multiple comparisons. Being a regression problem, questions must specify E. The knockoff setting fixes $E = \{1, \ldots, p\}$ in which case the questions of interest are

$$\bar{Q}^F_j : \text{Is feature } j \text{ correlated with the residual when trying predict } y \text{ from } X_{-j}? \tag{12}$$

with F above standing for the *full* model. The authors consider X fixed and take $\mathfrak{D} = y$ and the statistical model to be

$$\mathfrak{M}^{\text{K}} = \left\{N(X\beta, \sigma^2 I) : \beta \in \mathbb{R}^p, \sigma^2 > 0\right\}. \tag{13}$$

There is again a natural transformation from questions to statistical hypotheses $\bar{Q}^F_j \mapsto H_{0,j} : \beta_j = 0$. The usual t or Z-statistics for the least-squares estimates in the above regression model could be used to test each of these hypotheses. Rather than use this embedding, the authors choose alternative statistics based on constructing a pseudo-feature for each feature j constructing a procedure that controls (a slight modification of) the FDR. Their demonstration of (modified) FDR control through counting processes has reinvigorated methodological work in FDR and has led to, among other things, work on other more adaptive procedures for FDR control Lei and Fithian [2016], Li and Barber [2015], Lei, Ramdas, and Fithian [2017], and Barber and Ramdas [2017]. Other interesting work in FDR control includes work on hierarchically arranged families of hypotheses Benjamini and Bogomolov [2014].

The authors of Barber and Candes [2014] demonstrate empirically that this construction can be more powerful than the natural embedding based on the usual t or Z statistics. The knockoffs framework has been extended Candes, Fan, Janson, and Lv [2016] to settings under which the law of X is assumed known and it is feasible to construct swap-exchangeable pseudo-features $\tilde{X}$, expanding the applicability of knockoffs when such assumptions are reasonable in which the questions $\bar{Q}_j^F$ are transformed to hypotheses of conditional independence.

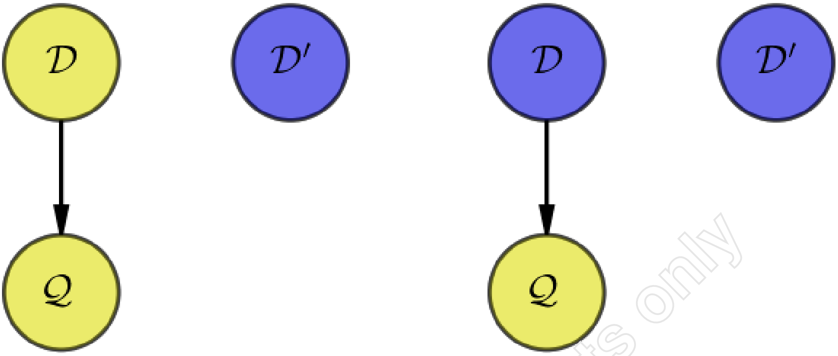
2.2 Does this match science's new paradigm? It seems that both the large-scale inference problem as well as the regression problem can be embedded into multiple comparisons problems (though technical considerations certainly still remain). We have also unfortunately ignored Bayesian methods up to this point, though we return to this briefly below.

Recall our original goal: to provide tools for inference in a paradigm where people collect data first and ask questions later. A quick look the examples show that the questions were actually pre-determined[1].

Not only were the questions pre-determined, the questions were formally transformed to statistical objects. This transformation is what allows statisticians to apply the formal methods of multiple comparisons to "answer" these questions. While these transformations of questions to statistical objects seem quite natural, they are not exhaustive. Why not just let the scientist look at the data to generate their own questions and have them pick the statistical objects for reporting?

2.3 Conditional inference. In this section, we describe a *conditional* approach to selective inference that allows the scientist to look at their data to generate questions of interest and corresponding statistical objects for final reporting. We do not have to look far for an example of the conditional approach. Indeed, our simple iteration of the scientific method as presented in Figure 1 provides an example.

[1] Arguably, one exception to this is the simultaneous approach of Berk, Brown, Buja, K. Zhang, and Zhao [2013] as it allows a researcher to choose which of some prespecified list of E to use in the report. But what is the scientist to do if she discovers her chosen E (after inspection) was not in the list specified before the analysis?



(a) The scientist conditions on $(\mathfrak{D}, \mathfrak{Q}(\mathfrak{D}))$ and must posit a model for the law of $\mathfrak{D}'|\mathfrak{D}$.

(b) The scientist conditions on $\mathfrak{Q}(\mathfrak{D})$, the minimal sigma algebra used to generate her statistical objects, and must posit a model for the joint law of $(\mathfrak{D}, \mathfrak{D}')$.

Figure 2: Two different conditional models.

2.3.1 The scientific method. Having collected data $\mathfrak{D}$, the scientist's data analysis can be represented as a function $\mathfrak{Q}(\mathfrak{D})$, quite literally the functions the scientist used in their exploratory data analysis, typically in some statistical software package. At this point, the scientist need not have attached any statistical model to the data as $\mathfrak{Q}(\mathfrak{D})$ are simply patterns.

Based on $\mathfrak{Q}(\mathfrak{D})$, the scientist is free to posit a statistical model $\mathfrak{M}$ (subject to defending this model to their peers) with corresponding statistical objects which will be used for formal statistical inference on $\mathfrak{D}'$. As the results are only evaluated on $\mathfrak{D}'$ we can view $\mathfrak{D}$ as fixed, fixing $\mathfrak{Q}(\mathfrak{D})$ as well. This fixing of $\mathfrak{Q}(\mathfrak{D})$ allows the scientist to transform these patterns into statistical objects such as hypothesis tests, point estimators or confidence intervals.

Fixing $\mathfrak{D}$ is equivalent to conditioning on it – any honest accounting of how $\mathfrak{D}$ and $\mathfrak{D}'$ came to be must acknowledge that $\mathfrak{D}$ is random so there certainly exists some joint distribution for $(\mathfrak{D}, \mathfrak{D}')$. Formal inference is applied only to $\mathfrak{D}'$ hence the scientist's model $\mathfrak{M}$ is really a model for the law of $\mathfrak{D}'|\mathfrak{D}$. This fixing of $\mathfrak{D}$ is illustrated in Figure 2a, denoting fixed variables by yellow and variables modeled by the scientist in blue.

2.3.2 Conditional approach in general. If $\mathfrak{Q}(\mathfrak{D})$ is enough information for the scientist to posit a model $\mathfrak{M}$ for the law $\mathfrak{D}'|\mathfrak{D}$, it is often reasonable to assume it is enough information for the scientist to posit a model for the joint law of $(\mathfrak{D}, \mathfrak{D}')$. For example, when both $\mathfrak{D}$ and $\mathfrak{D}'$ are IID samples from some population then any model for $\mathfrak{D}'$ is similarly a model for $\mathfrak{D}$. In this setting, if a model for $\mathfrak{D}'$ is defensible to their peers, the same model must be certainly defensible for $\mathfrak{D}$.

Since it is $\mathfrak{Q}(\mathfrak{D})$ that led the scientist to the questions that were then transformed into statistical objects, it is sufficient to only fix $\mathfrak{Q}(\mathfrak{D})$. This is the basis of the conditional approach to selective inference Bi, Markovic, Xia, and J. Taylor [2017], Fithian, D. Sun, and J. Taylor [2014], and Lee, D. L. Sun, Y. Sun, and J. E. Taylor [2016]. Conditioning only on $\mathfrak{Q}(\mathfrak{D})$ we apply the formal tools of statistical inference to the remaining randomness in $(\mathfrak{D}, \mathfrak{D}')$. In turn, this means that the appropriate Type I error to consider is the *selective Type I error*, requiring the *conditional guarantee*

$$\mathbb{P}_F(T(\mathfrak{D}, \mathfrak{D}') = 1 | \mathfrak{Q}(\mathfrak{D}) = q) \leq \alpha, \qquad \forall F \in H_0. \tag{14}$$

Coverage for an interval estimate is replaced with the notion of *selective coverage*. This setting is depicted in Figure 2b, in which only $\mathfrak{Q}(\mathfrak{D})$ is conditioned on. If $\mathfrak{D}'$ is unavailable, it is still possible to use these tools as long as the scientist is able to defend a model for the law of $\mathfrak{D}$ to their peers.

Conditioning on the event $\{\mathfrak{Q}(\mathfrak{D}) = q\}$ transforms any model $\mathfrak{M}$ for the joint law of $(\mathfrak{D}, \mathfrak{D}')$ to a new model

$$\mathfrak{M}_q^* = \left\{ F^* : \frac{dF^*}{dF}(d, d') \propto 1_{\{\mathfrak{Q}^{-1}(q)\}}(d), F \in \mathfrak{M} \right\} \tag{15}$$

where q is the value of $\mathfrak{Q}(\mathfrak{D})$ observed by the scientist. We should note that, as in Figure 2a, the model itself has been selected *after* the scientist has observed the patterns $\mathfrak{Q}(\mathfrak{D})$.

What has this approach bought us? For one thing, we have freed the scientist from the "natural" transformations of questions to statistical objects we saw in our discussion of simultaneous methods. The scientist is free to transform the observed patterns into statistical objects how they see fit.

At what cost has this benefit come? The first cost is that conditional rather than marginal guarantees are required. Conditional guarantees are generally stronger than marginal guarantees, though they may need stricter assumptions to hold. Exploration of the gap between assumptions required for selective and marginal guarantees is certainly an interesting problem.

A second cost is the cost of exploration itself. In the classical scientific method, the scientist is faced with the cost of collecting a second data set $\mathfrak{D}'$ in order to apply the formal methods of statistical inference. In Figure 2b the scientist is able to reuse some of the data for inference. How much is available for reuse will depend very much on $\mathfrak{Q}$ – if this is the identity map, then clearly fixing $\mathfrak{Q}(\mathfrak{D})$ is equivalent to fixing $\mathfrak{D}$ and no data remains after exploration. If $\mathfrak{M} = \{f_\theta : \theta \in \Theta\}$ is a parametric model, then the amount of information for estimating θ can be quantified by the Fisher information. The model $\mathfrak{M}^*_q$ will have its own Fisher information, referred to as *leftover Fisher information* in Fithian, D. Sun, and J. Taylor [2014]. Ideally, the scientist is able to explore their data in such a way that they can discovering interesting questions while preserving leftover information.

3 Examples and implications of the conditional approach

In the setting of Figure 2b we allow the scientist to posit a model $\mathfrak{M}$ after having observed $\mathfrak{Q}(\mathfrak{D})$, though we require that the scientist posit a model for the joint law of $(\mathfrak{D}, \mathfrak{D}')$ rather than the usual $\mathfrak{D}'|\mathfrak{D}$. The scientist, perhaps with the assistance of a statistician, will then declare some statistical objects of interest defined on $\mathfrak{M}$: hypothesis tests, point estimates, confidence intervals, etc.

To simplify our presentation we make the assumption that $(\mathfrak{D}, \mathfrak{D}')$ are (perhaps asymptotically) jointly Gaussian, implying that the patterns the scientist sees are formed by inspecting some approximately linear statistic. As the Gaussian family is parametric, this also implies there is a well-defined notion of leftover information. Of course, many statistical models (and selection procedures) can be reduced (asymptotically) to this setting. We acknowledge that allowing the scientist to view more complicated statistics, such as scatterplots does not obviously fit into this framework and certainly this is worthy of further study. These two observations bring us to the first of several challenges in the conditional setting.

Challenge 1 (Selective Central Limit Theorem)**.** Without the effect of selection, there is an extensive literature on uses of the CLT to justify approximations in statistical inference. Sequences of models such as $\mathfrak{M}^*_{q,n}$ do not fit into this classical setting, though sometimes uniform consistency in L^p and in an appropriate weak sense (i.e. avoiding the impossibility results of Leeb and Pötscher [2006]) along sequences of models $\mathfrak{M}_n$ can be transferred over to sequences $\mathfrak{M}^*_{q,n}$ Tian and J. E. Taylor [2015] and Markovic and J. Taylor [2016]. We acknowledge that these results are likely suboptimal.

Challenge 2 (Rich Interactive Selection)**.** Scatterplots are standard tools in exploratory analyses, as are other summaries. Are there realistic mechanisms to release similar information to scientists that are not wasteful in leftover information?

3.1 The scientific method is inadmissible. Our first example makes a rather bold claim. In this setting, the scientist has access to $\mathfrak{D}'$ and the mechanism by which $\mathfrak{Q}(\mathfrak{D})$ is fixed is by fixing (or conditioning) on all of $\mathfrak{D}$. This is a finer sigma algebra than that of $\mathfrak{Q}(\mathfrak{D})$, which means we are conditioning on more than we need to. Statistical objects constructed conditional on $\mathfrak{D}$ are often inadmissible with concrete dominating procedures. A more precise statement in terms of hypothesis tests can be found in Theorem 9 of Fithian, D. Sun, and J. Taylor [2014]. Though this theorem is stated in terms of data splitting Cox [1975] in which $\mathfrak{D}$ is part of a full data set and $\mathfrak{D}'$ the remaining data, it is clearly applicable to the setting where a scientist collects fresh data $\mathfrak{D}'$.

For concreteness, consider a simple model for the *file-drawer* filter (in which only large positive Z statistics are reported) along with a replication study. We can model this as

$$(\mathfrak{D}, \mathfrak{D}') \sim N\left(\begin{pmatrix}\mu \\ \mu\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right)$$

being suitably normalized sample means from some population and

$$\mathfrak{Q}(\mathfrak{D}) = \begin{cases} 1 & \mathfrak{D} > 2 \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

and the scientist has observed $\mathfrak{Q}(\mathfrak{D}) = 1$. The natural replication estimate is

$$\hat{\mu}(\mathfrak{D}, \mathfrak{D}') = \mathfrak{D}'.$$

It is evident that the sufficient statistic for μ in $\mathfrak{M}$ is $(\mathfrak{D} + \mathfrak{D}')/2$. It is also clear that this holds conditional on $\mathfrak{Q}(\mathfrak{D})$ as well. Hence, one can Rao-Blackwellize $\hat{\mu}$

$$\hat{\mu}_{RB}(\mathfrak{D} + \mathfrak{D}') = \mathbb{E}_F\left(\mathfrak{D}' \middle| \mathfrak{Q}(\mathfrak{D}), \mathfrak{D} + \mathfrak{D}'\right).$$

Simple calculations show that for $\mu \gg 2$ the Rao-Blackwellized estimator is essentially $(\mathfrak{D}+\mathfrak{D}')/2$ which has variance $1/2$ compared to 1, the variance of $\hat{\mu}$. Confidence intervals for μ and tests of hypotheses of the form $H_0 : \mu = \mu_0$ are also relatively straightforward to construct in the conditional model

$$\mathfrak{M}_1^* = \left\{ F^* : \frac{dF^*}{dF}(d, d') \propto 1_{(2,\infty)}(d), F \in \mathfrak{M} \right\}$$

Such procedures have been proposed in similar but not identical settings Cohen and Sackrowitz [1989] and Sampson and Sill [2005] in which follow-up data $\mathfrak{D}'$ is available through a designed experiment.

We see a clear gap in performance between the simple estimator arrived at following the classical scientific method and one which conditions on less. We note that this statement of inadmissibility is relative to the sigma algebra we condition on. If it is asserted that the correct sigma algebra conditions on $\mathfrak{D}$ then $\hat{\mu}$ is in fact the UMVU and this domination disappears. It is argued in Bi, Markovic, Xia, and J. Taylor [2017] that the minimal sigma algebra to condition is that generated by the patterns the scientist observed as this is the sigma algebra with which the statistical objects is determined. This is very mildly in contrast to the formal theory laid out in Fithian, D. Sun, and J. Taylor [2014] which presumes that $\mathfrak{Q}(\mathfrak{D})$ is already in the form of statistical objects, though the theory of itself needs no essential change.

Often, it will be convenient to condition on more, perhaps to ensure that computations are feasible. This is allowed for in the formal theory of Fithian, D. Sun, and J. Taylor [ibid.] through a selection variable, and was used in the first (non-data splitting) conditional approach to the regression problem in Lee, D. L. Sun, Y. Sun, and J. E. Taylor [2016]. It is not hard to imagine exploratory analyses $\mathfrak{Q}$ whose sigma algebra is simply too complex to describe so that restricting models to events $\mathfrak{Q}^{-1}(q)$ may be computationally infeasible resulting in a costly waste of leftover information. This suggests the seemingly uncontroversial principle of not using too complex an exploratory analysis to generate questions. We see no reason a priori that more complex analyses will lead the scientist to more interesting questions.

3.1.1 Dominating the scientific method in the regression problem. Figure 3 from Tian and J. E. Taylor [2015] (which reproduces and extends an example in Fithian, D. Sun, and J. Taylor [2014]) demonstrates inadmissibility in the regression setting in which the LASSO R. Tibshirani [1996] is used to select variables. In this example, the total number of samples is held fixed and the portion allocated to $\mathfrak{D}$ and $\mathfrak{D}'$ varies. The vertical axis is Type II error, the complement of statistical power. The horizontal axis is the probability of the selection mechanism discovering all of the true effects in this regression problem, meant to be a proxy for the quality of the patterns revealed to the scientist. We carried out inference for partial correlations in the Gaussian model with features E selected by $\mathfrak{D}$. This seems the natural model a scientist would use for $\mathfrak{D}'$ if they decided to replicate the study, collecting only features E discovered in the pilot study. Other statistical objects are certainly reasonable, such as the E coordinates of the full model (a subset of the targets in model (13)) or the debiased LASSO targets Javanmard and Montanari [2013], T. Sun and C.-H. Zhang [2012], and Dezeure, Bühlmann, Meier, and Meinshausen [2015] if $n < p$. The curve labelled *data splitting* follows the classical scientific method, conditioning on $\mathfrak{D}$. The curve labelled *data carving* conditions only on $\mathfrak{Q}(\mathfrak{D})$ but makes decisions about exactly the same statistical objects as data splitting. It is clear that data carving dominates data splitting with data splitting having Type II error of different magnitude to the data carving curve and red curve. This red curve brings us to our next example.

3.2 Noisier is better? Randomized selection algorithms. The initial description of Figure 1 had $\mathfrak{D}'$ as fresh data while pilot $\mathfrak{D}$ was used to discover patterns $\mathfrak{Q}(\mathfrak{D})$. This is an artificial constraint: the scientist uses $\mathfrak{D}$ to discover patterns and simply must posit a model for the joint law of $(\mathfrak{D}, \mathfrak{D}')$. Inference is then carried out after restricting this model to the event $\mathfrak{Q}^{-1}(q)$. In particular, $\mathfrak{D}$ could be a randomized version of $\mathfrak{D}'$ with the randomization chosen by the scientist perhaps with the assistance of a statistician Tian and J. E. Taylor [2015]. Indeed, our file drawer replication study can easily be seen in this light: let Z_1 denote the pilot data and Z_2 the replication data. We take $\mathfrak{D}' = (Z_1 + Z_2)/2 \sim N\left(\mu, \frac{1}{2}\right)$ and

$$\mathfrak{D}|\mathfrak{D}' \sim N\left(\mathfrak{D}', \frac{1}{2}\right).$$

Unsurprisingly, by sufficiency, the law of $\mathfrak{D}|\mathfrak{D}'$ does not depend on the unknown μ and it is enough to consider the law of $\mathfrak{D}'$ after marginalizing over $\mathfrak{D}$. The statistical problem can be reduced to inference in the model

$$\mathfrak{M}_q^* = \left\{F^* : \frac{dF^*}{dF}(d') \propto \int 1_{\{\mathfrak{Q}(\cdot)=q\}}(u) G(du|d'), F \in \mathfrak{M}\right\} \tag{17}$$

with $G(\cdot|d')$ the kernel representing the conditional law of $\mathfrak{D}$ given $\mathfrak{D}'$.

It is apparent that the Radon-Nikodym derivative or likelihood ratio relating F^* to its corresponding F will often be a smooth function. In the case that $G(\cdot|d') = \delta_{d'}$, in which case $\mathfrak{D} \overset{\text{a.s.}}{=} \mathfrak{D}'$, this will in fact be an indicator function. In the setting of Gaussian randomization, the smoothness of this function can be directly related to the leftover information, explaining why both data carving and the additive noise model in Figure 3 show an improvement in power after addition of at least some randomness into the pattern generation stage. With no randomization, the (rescaled) leftover Fisher information can rapidly approach 0 for some parameter values while the corresponding information after randomization can be bounded below. In this sense some noise is better than no noise, though Figure 3 demonstrates there is a tradeoff between quality of patterns and statistical power.

The red curve in Figure 3 uses the same data generating mechanism as the LASSO in the regression problem with $\mathfrak{D}' = y$ and $\mathfrak{D} = y + \omega$ with $\omega \sim N(0, \tau^2 I)$. The curve traces out the Type II error and probability of screening as τ varies. It seems as if this particular randomization does better than data carving in this figure. However, as $\mathfrak{D}$ and hence $\mathbb{Q}(\mathfrak{D})$ differ between the two curves, a direct comparison of data carving to the randomization above is somewhat difficult. In practice, there is no guarantee that any two scientists given access to $\mathfrak{D}$ will use the same $\mathbb{Q}$ or construct the same statistical objects having observed the same $\mathbb{Q}(\mathfrak{D})$. It is not hard to imagine settings where some scientists know the "right" $\mathbb{Q}$ to use based on domain experience, or perhaps know the "right" statistical objects to report having observed $\mathbb{Q}(\mathfrak{D})$. Such scientists will likely be able to extract more interesting answers from $\mathfrak{D}'$ and likely more money from funding agencies than others. Identifying such scientists and / or modelling their behavior, or even identifying the "right" statistical objects seems a daunting task which we decline to pursue.

3.3 Patterns divined from convex programs. The LASSO R. Tibshirani [1996] is a popular algorithm used to discover important features in the regression context. Let us remind the readers of the LASSO optimization problem

$$\hat{\beta}_\lambda(y, X) = \text{argmin}_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{18}$$

It is well known that for large enough values of λ, the solution will often be sparse. The non-zero entries of $\hat{\beta}(y, X, \lambda)$ are a natural candidate for the "important" variables in predicting y from X in a linear model. Further, the event

$$\left\{y : \text{sign}(\hat{\beta}_\lambda(y, X)) = s\right\}$$

can be described in terms of a set of affine inequalities Lee, D. L. Sun, Y. Sun, and J. E. Taylor [2016]. This observation demonstrated that conditional inference in the regression problem was feasible, yielding the *polyhedral lemma* subsequently used in Heller, Meir, and Chatterjee [2017], R. J. Tibshirani, J. Taylor, Lockhart, and R. Tibshirani [2016], and R. J. Tibshirani, Rinaldo, R. Tibshirani, and Wasserman [2015] among other places. The assumption that X be fixed is not strictly necessary with suitable modification of the covariance estimate in the polyhedral lemma J. Taylor and R. Tibshirani [2017].

Challenge 3 (High Dimensional Selective Inference)**.** High dimensional inference Bühlmann and Geer [2011] is a very important topic given the sheer size of p in modern science. Rigorously addressing the conditional approach in this setting is certainly challenging. While some results are available Markovic, Xia, and J. Taylor [2017] and Wasserman and Roeder [2009] much work remains.

The LASSO has inspired many other convex optimization algorithms meant to elucidate interesting structure or patterns in $\mathfrak{D}$. A very short list might include Becker, Candès, and Grant [2011], Chen, Donoho, and Saunders [1998], Ming and Lin [2005], and Yuan and Lin [2007]. Many natural $\mathcal{Q}$ suggest themselves from such convex programs. Convex programs also lend themselves naturally to randomization by perturbation of the objective function. While Section 2.3 describes the reason *why* we condition on $\mathcal{Q}(\mathfrak{D})$, it is important to describe *how* we achieve this. Here we describe some of the approach of Tian, Panigrahi, Markovic, Bi, and J. Taylor [2016] which gives a general idea of the *how*.

Consider the problem

$$\hat{\beta}(\mathfrak{D}, \omega) = \text{argmin}_{\beta \in \mathbb{R}^k} \ell(\beta; \mathfrak{D}) + \mathcal{P}(\beta) - \omega^T \beta + \frac{\epsilon}{2}\|\beta\|_2^2 \tag{19}$$

where ℓ is some smooth loss involving the data (not necessarily a negative log-likelihood), $\mathcal{P}$ is some structure inducing convex function, $\epsilon > 0$ is some small parameter that is sometimes necessary in order to assure the program has a solution and $\omega \sim G$ is a randomization with G (typically having a smooth density g) chosen by the scientist. In terms of Figure 2b the randomization ω can be modelled as part of the function $\mathcal{Q}$ and we are free to take $\mathfrak{D} = \mathfrak{D}'$.

The KKT conditions of such a problem can be written as

$$\omega = \nabla \ell(\beta; \mathfrak{D}) + u + \epsilon \cdot \beta. \tag{20}$$

with (β, u) required to satisfy

$$u \in \partial \mathcal{P}(\beta). \tag{21}$$

Suppose that the scientist seeks for patterns in the pair (β, u) so that $\mathcal{Q} = \mathcal{Q}(\mathfrak{D}, \omega) = \bar{\mathcal{Q}}(\hat{\beta}(\mathfrak{D}, \omega), \hat{u}(\mathfrak{D}, \omega))$. It turns out that, in wide generality, there is a natural mechanism through which we can condition on events expressed in terms of (β, u). Geometrically, if $\mathcal{P}$ is a seminorm given by the support function of convex set K, then the condition condition is equivalent to $(u, \beta) \in N(K)$ where $N(K)$ is the normal bundle of K and the integral necessary to restrict to the event of interest can be expressed through a change of variables as

$$\mathbb{P}(\mathcal{Q}(\mathfrak{D}, \omega) = q | \mathfrak{D}) = \int_{N(K)} 1_{\{\bar{\mathcal{Q}}^{-1}(q)\}}(u, \beta)\; g(\phi(u, \beta; \mathfrak{D}))\; J_\phi(u, \beta; \mathfrak{D})\; \mathcal{H}_k(d\beta\; du). \tag{22}$$

where g is the density of ω,

$$\phi(\beta, u; \mathfrak{D}) = \nabla \ell(\beta; \mathfrak{D}) + u + \epsilon \cdot \beta$$

is the change of variables introduced by inverting the KKT conditions above and $\mathcal{H}_k$ is k-dimensional Hausdorff measure on $N(K) \subset \mathbb{R}^{2k}$. See Tian, Panigrahi, Markovic, Bi, and

J. Taylor [ibid.] for further details and examples beyond the LASSO. With a little work, expanding the Jacobian in the integrals in (22) yield objects closely related to integrals against the generalized curvature measures of K Schneider [1993]. Much of the work cited above on Gaussian processes and simultaneous inference also involve such geometric objects through Weyl and Steiner's tube formulae Adler and J. E. Taylor [2007].

In many cases of interest the rather complicated looking (22) can be expressed as

$$\int_{\bar{K}_q} g\left(A\mathfrak{D} + Bo + \eta\right)\, do \tag{23}$$

for some η measurable with respect to $\mathfrak{Q}(\mathfrak{D}, \omega)$ and some nice (often convex and polyhedral) $\bar{K}_q \subset \mathbb{R}^k$ where the variable of integration o is meant to stand for "optimization variables", i.e. the pair (u, β) in (22). If we presume that the randomization used by the scientist is Gaussian (e.g. data carving can be represented as asymptotically adding Gaussian randomization Markovic and J. Taylor [2016]), then the likelihood ratio in the model $\mathfrak{M}^*_q$ can be expressed as

$$\mathbb{P}(o \in \bar{K}_q | \mathfrak{D}) \tag{24}$$

for some η measurable with respect to $\mathfrak{Q}(\mathfrak{D}, \omega)$ where the pair $(\mathfrak{D}, o)$ are jointly Gaussian with mean and covariance determined by the mean and covariance of $\mathfrak{D}$, the pair (A, B) and the covariance matrix of the randomization ω.

3.4 Benjamini-Hochberg given access to replication data.

Some selection algorithms do not involve directly solving a (randomized) convex program such as (19) yet the appropriate likelihood ratio can be described similarly. For instance, suppose $\mathfrak{D} = Z \sim N(\mu, \Sigma)$ and the selection algorithm involves taking the top k of the Z-statistics. A randomized version might add $\omega \sim N(0, \tau^2 I)$ to Z before ranking. We can take the map $\mathfrak{Q}(Z, \omega)$ to return the identity of the top k and perhaps their signs. Formally this is an example of (19) involved in finding the maximizer of the convex function that returns the sum of the top k order statistics, an example of SLOPE Bogdan, Berg, Sabatti, Su, and Candès [2015]. Let E_k denote the identity of these variables and s_k their signs. The selection probability can be expressed as

$$\mathbb{P}(o \in \mathfrak{D} + \bar{C}(E_k, s_k) + \eta | \mathfrak{D}) \tag{25}$$

where $\bar{C}(E_k, s_k)$ is the convex cone identifying the top k coordinates on $\mathbb{R}^p$ and their signs.

Another example with such a representation is a version the Benjamini-Hochberg algorithm in which $\mathfrak{D} = Z \sim N(\mu, \Sigma)$ and $\mathfrak{Q}(\mathfrak{D}, \omega)$ identifies which effects are selected by BH using suitably normalized "new" Z-statistics $Z + \omega$ as well as the ordering of the non-rejected null Z-statistics Reid, J. Taylor, and R. Tibshirani [2017]. This allows a scientist to run the BH algorithm on a randomized version of their data, preserving some information for a point estimate or perhaps an interval estimate. Recalling our file drawer replication study, we see that this is mathematically equivalent to setting $\mathfrak{D}$ to be pilot data and $\mathfrak{D}'$ to be a replication study. While it is possible to construct intervals for the effects of the variables selected by BH using only $\mathfrak{D}$ Benjamini and Yekutieli [2005], it is not immediately obvious how one might improve these intervals given replication data. Nor is it clear how one might arrive at unbiased estimates of such parameters using only $\mathfrak{D}$, though a simple generalization of our file drawer examples illustrates how to Rao-Blackwellize the replicate unbiased estimate, or hypothesis tests and confidence intervals.

We mention this example to correct what seems to be a misconception in some of the selective inference literature. The conditional approach is sometimes presented as subordinate in a hierarchy of types of simultaneous guarantees Benjamini [2010] and Benjamini and Bogomolov [2014]. This is not really the case, the algorithm we just described would have a marginal FDR-control property as well as unbiased estimators of the selected effects arrived at through the conditional approach. In other words, the notion of simultaneous inference in the conditional approach is certainly a well-defined topic of research, see Hung and Fithian [2016] for another example of use of the conditional approach in the simultaneous setting.

Challenge 4 (Simultaneous Selective Inference)**.** What kind of finite sample procedures can be used to control FDR for a collection of hypotheses generated from $\mathfrak{Q}(\mathfrak{D})$? If the selection step has produced a small set of questions, is multiplicity correction still needed?

3.5 A tractable pseudo-likelihood. Let us restrict our attention to the case $\mathfrak{D} = Z \sim N(\mu, \Sigma)$ when the appropriate appropriate can be expressed as in (23). In this setting, $\mathfrak{M}_q^*$ can be viewed as the marginal law of $\mathfrak{D}$ under some joint Gaussian law for $(\mathfrak{D}, o)$ truncated to the event (23) with implied mean of $(\mathfrak{D}, o)$ some affine function of μ. This truncated Gaussian law has normalizing constant

$$\mathbb{P}_\mu(o \in \bar{K}_q). \tag{26}$$

If this normalizing constant were known, the rich toolbox of exponential families would be at our disposal. In Panigrahi, J. Taylor, and Weinstein [2016] we propose using a smoothed version of a Chernoff or large deviations estimate of (26). As the sets $\bar{K}_q$ are often simple, it is possible to solve this optimization problem quickly. This optimization procedure yields a composite or pseudo-MLE estimate that yields (approximately) conditionally unbiased estimates of μ in this setting. Investigation of the performance of this estimator is a topic of ongoing research.

Having (approximately) normalized the likelihood ratio in model $\mathfrak{M}_q^*$ it is apparent that one may put a prior on μ itself. This approach was developed in the univariate setting in Yekutieli [2012] in which there is often no need to approximate the normalizing constant. Use of this approximation is considered in Panigrahi, J. Taylor, and Weinstein [2016]. One may then use all of the advantages of the Bayesian paradigm in this conditional approach, modulo the fact that the likelihood is only approximately normalized.

3.6 Combining queries: inferactive data analysis. We have alluded to scientists posing questions or queries of $\mathfrak{D}$ in terms of the solution to a randomized convex program. Of course, a more realistic data analysis paradigm allows the scientist more complex queries. Indeed, the result of one query may inspire a scientist to pose a new query, or perhaps to spend some grant money to collect fresh data. A satisfactory theory of data analysis should be able to handle such situations. Suppose we limit each query to those similar to (23), allowing the results of one query to influence the following queries. Then, it is not hard to see that, after two queries, the (unnormalized) appropriate likelihood ratio takes the form

$$\mathbb{P}(o_1 \in \bar{K}_{q_1} | \mathfrak{D}_1) \cdot \mathbb{P}(o_2 \in \bar{K}_{(q_1,q_2)} | (\mathfrak{D}_1, \mathfrak{D}_2)) \tag{27}$$

where each o_i represent the optimization variables in each query and $\mathbb{P}$ is some implied joint Gaussian law for the triple $(\mathfrak{D}_1, \mathfrak{D}_2, o_1, o_2)$ with $\mathfrak{D}_1$ the data available at time 1, $(\mathfrak{D}_1, \mathfrak{D}_2)$ at time 2. Generalizing this to m queries is straightforward. We have named this resulting formalism for inference after allowing a scientist to interact with their data *inferactive data analysis* Bi, Markovic, Xia, and J. Taylor [2017].

Challenge 5 (In Silico Implementation of Inferactive Data Analysis)**.** The approximate pseudo-MLE is seen to be a separable convex optimization problem, yielding hope for scaling up to a reasonable number of queries. In the limiting Gaussian model, the relevant reference measures can formally be represented via generalizations of *estimator augmentation* Tian, Panigrahi, Markovic, Bi, and J. Taylor [2016] and Zhou [2014]. Some small steps have been taken in this direction but more work is definitely needed.

Other approaches to adaptive data analysis include Berk, Brown, Buja, K. Zhang, and Zhao [2013] in the regression problem, as well as some very interesting work in applications of differential privacy to data analysis Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth [2014].

Acknowledgments. All of our work presented here is joint work, particularly with current and former students and colleagues from the Statistical Learning Group in the Department of Statistics at Stanford run jointly with Trevor Hastie and Rob Tibshirani. The author has also profited greatly from discussions with many other colleagues having been given the chance to present this work at numerous conferences.

References

Robert J. Adler and Jonathan E. Taylor (2007). *Random fields and geometry*. Springer Monographs in Mathematics. New York: Springer (cit. on pp. 3009, 3019).

J-M. Azaïs and M. Wschebor (2009). *Level sets and extrema of random processes and fields*. Wiley (cit. on p. 3009).

Rina Foygel Barber and Emmanuel Candes (Apr. 2014). "Controlling the False Discovery Rate via Knockoffs". arXiv: 1404.5609 (cit. on pp. 3009, 3010).

Rina Foygel Barber and Aaditya Ramdas (Sept. 2017). "The p-filter: multilayer false discovery rate control for grouped hypotheses". en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 1247–1268 (cit. on p. 3010).

Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant (2011). "Templates for convex cone problems with applications to sparse signal recovery". English. *Mathematical Programming Computation* 3.3, pp. 165–218 (cit. on p. 3017).

Yoav Benjamini (Dec. 2010). "Simultaneous and selective inference: Current successes and future challenges". eng. *Biometrical Journal. Biometrische Zeitschrift* 52.6, pp. 708–721. PMID: 21154895 (cit. on pp. 3007, 3020).

Yoav Benjamini and Marina Bogomolov (Jan. 2014). "Selective inference on multiple families of hypotheses". en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 297–318 (cit. on pp. 3010, 3020).

Yoav Benjamini and Yosef Hochberg (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing". *J R Statist Soc B* 57.1, pp. 289–300 (cit. on pp. 3007, 3008).

Yoav Benjamini and Daniel Yekutieli (Mar. 2005). "False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters". *Journal of the American Statistical Association* 100.469, pp. 71–81 (cit. on pp. 3007, 3020).

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao (Apr. 2013). "Valid post-selection inference". EN. *The Annals of Statistics* 41.2, pp. 802–837. MR: MR3099122 (cit. on pp. 3007, 3009, 3010, 3021).

Nan Bi, Jelena Markovic, Lucy Xia, and Jonathan Taylor (July 2017). "Inferactive data analysis". arXiv: 1707.06692 (cit. on pp. 3012, 3014, 3021).

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès (Sept. 2015). "SLOPE—Adaptive variable selection via convex optimization". EN. *The Annals of Applied Statistics* 9.3, pp. 1103–1140. MR: MR3418717 (cit. on p. 3019).

Peter Bühlmann and Sara van de Geer (June 2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 1st Edition. Springer (cit. on p. 3017).

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv (Oct. 2016). "Panning for Gold: Model-free Knockoffs for High-dimensional Controlled Variable Selection". arXiv: 1610.02351 (cit. on p. 3010).

Scott Chen, David Donoho, and Michael Saunders (1998). "Atomic decomposition for basis pursuit". *SIAM Journal on Scientific Computing* 20.1, pp. 33–61 (cit. on pp. 3017, 3018).

Arthur Cohen and Harold B Sackrowitz (1989). "Two stage conditionally unbiased estimators of the selected mean". *Statistics & Probability Letters* 8.3, pp. 273–278 (cit. on p. 3014).

DR Cox (1975). "A note on data-splitting for the evaluation of significance levels". *Biometrika* 62.2, pp. 441–444 (cit. on p. 3014).

Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen (Nov. 2015). "High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi". EN. *Statistical Science* 30.4, pp. 533–558. MR: MR3432840 (cit. on p. 3015).

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth (Nov. 2014). "Preserving Statistical Validity in Adaptive Data Analysis". arXiv: 1411.2664 (cit. on p. 3021).

Bradley Efron (Nov. 2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. English. Reprint edition. Cambridge, UK; New York: Cambridge University Press (cit. on pp. 3006, 3007).

William Fithian, Dennis Sun, and Jonathan Taylor (Oct. 2014). "Optimal Inference After Model Selection". arXiv: 1410.2597 (cit. on pp. 3012–3015).

Ruth Heller, Amit Meir, and Nilanjan Chatterjee (Nov. 2017). "Post-selection estimation and testing following aggregated association tests". arXiv: 1711.00497 (cit. on p. 3017).

Kenneth Hung and William Fithian (Oct. 2016). "Rank Verification for Exponential Families". arXiv: 1610.03944 (cit. on p. 3020).

Clifford M Hurvich and Chih—Ling Tsai (1990). "The impact of model selection on inference in linear regression". *The American Statistician* 44.3, pp. 214–217 (cit. on p. 3007).

Adel Javanmard and Andrea Montanari (2013). "Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory". arXiv: 1301.4240 (cit. on p. 3015).

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor (June 2016). "Exact post-selection inference, with application to the lasso". EN. *The Annals of Statistics* 44.3, pp. 907–927. MR: MR3485948 (cit. on pp. 3007, 3012, 3015, 3017).

Hannes Leeb and Benedikt M. Pötscher (Oct. 2006). "Can one estimate the conditional distribution of post-model-selection estimators?" *The Annals of Statistics* 34.5, pp. 2554–2591. MR: MR2291510 (cit. on p. 3013).

Lihua Lei and William Fithian (Sept. 2016). "AdaPT: An interactive procedure for multiple testing with side information". arXiv: 1609.06035 (cit. on p. 3010).

Lihua Lei, Aaditya Ramdas, and William Fithian (Oct. 2017). "STAR: A general interactive framework for FDR control under structural constraints". arXiv: 1710.02776 (cit. on p. 3010).

Ang Li and Rina Foygel Barber (May 2015). "Accumulation tests for FDR control in ordered hypothesis testing". arXiv: 1505.07352 (cit. on p. 3010).

Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani (Apr. 2014). "A significance test for the lasso". EN. *The Annals of Statistics* 42.2, pp. 413–468. MR: MR3210970 (cit. on p. 3007).

Jelena Markovic and Jonathan Taylor (Dec. 2016). "Bootstrap inference after using multiple queries for model selection". arXiv: 1612.07811 (cit. on pp. 3013, 3019).

Jelena Markovic, Lucy Xia, and Jonathan Taylor (Mar. 2017). "Comparison of prediction errors: Adaptive p-values after cross-validation". arXiv: 1703.06559 (cit. on p. 3017).

Yuan Ming and Yi Lin (2005). "Model selection and estimation in regression with grouped variables". *Journal of the Royal Statistical Society: Series B* 68.1, pp. 49–67 (cit. on pp. 3017, 3018).

Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein (2016). "Bayesian Post-Selection Inference in the Linear Model". arXiv: 1605.08824 (cit. on pp. 3020, 3021).

Stephen Reid, Jonathan Taylor, and Robert Tibshirani (June 2017). "Post-selection point and interval estimation of signal sizes in Gaussian samples". en. *Canadian Journal of Statistics* 45.2, pp. 128–148 (cit. on p. 3020).

Allan R Sampson and Michael W Sill (2005). "Drop-the-Losers Design: Normal Case". *Biometrical Journal* 47.3, pp. 257–268 (cit. on p. 3014).

Rolf Schneider (1993). *Convex bodies: the Brunn-Minkowski theory*. Vol. 44. Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press (cit. on p. 3019).

D. O Siegmund and K. J Worsley (1995). "Testing for a signal with unknown location and scale in a stationary Gaussian random field". *The Annals of Statistics* 23.2, pp. 608–639 (cit. on p. 3009).

John D. Storey (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value". *Ann Statist* 31.6, pp. 2013–2035 (cit. on p. 3007).

Jiayang Sun (Jan. 1993). "Tail Probabilities of the Maxima of Gaussian Random Fields". *The Annals of Probability* 21.1. ArticleType: research-article / Full publication date: Jan., 1993 / Copyright © 1993 Institute of Mathematical Statistics, pp. 34–71 (cit. on p. 3009).

Tingni Sun and Cun-Hui Zhang (Dec. 2012). “Scaled sparse linear regression”. *Biometrika* 99.4, pp. 879–898 (cit. on p. 3015).

A. Takemura and S. Kuriki (2002). “Maximum of Gaussian field on piecewise smooth domain: Equivalence of tube method and Euler characteristic method.” *Ann. of Appl. Prob.* 12.2, pp. 768–796 (cit. on p. 3009).

Jonathan Taylor and Robert Tibshirani (Mar. 2017). “Post-selection inference for ℓ1-penalized likelihood models”. en. *Canadian Journal of Statistics* (cit. on p. 3017).

“Theories of Data Analysis” (n.d.). “Theories of Data Analysis: From Magical Thinking Through Classical Statistics”. In: (cit. on p. 3006).

Xiaoying Tian, Snigdha Panigrahi, Jelena Markovic, Nan Bi, and Jonathan Taylor (2016). “Selective sampling after solving a convex problem”. arXiv: 1609.05609 (cit. on pp. 3018, 3021).

Xiaoying Tian and Jonathan E. Taylor (July 2015). “Selective inference with a randomized response”. arXiv: 1507.06739 (cit. on pp. 3013, 3015, 3016).

Robert Tibshirani (1996). “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B* 58.1, pp. 267–288 (cit. on pp. 3015, 3017).

Ryan J. Tibshirani, Alessandro Rinaldo, Robert Tibshirani, and Larry Wasserman (June 2015). “Uniform Asymptotic Inference and the Bootstrap After Model Selection”. arXiv: 1506.06266 (cit. on p. 3017).

Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani (Apr. 2016). “Exact Post-Selection Inference for Sequential Regression Procedures”. *Journal of the American Statistical Association* 111.514, pp. 600–620 (cit. on p. 3017).

John W. Tukey (1980). “We Need Both Exploratory and Confirmatory”. *The American Statistician* 34.1, pp. 23–25 (cit. on p. 3005).

Larry Wasserman and Kathryn Roeder (Oct. 2009). “High-dimensional variable selection”. EN. *The Annals of Statistics* 37.5. Zentralblatt MATH identifier: 05596898; Mathematical Reviews number (MathSciNet): MR2543689, pp. 2178–2201 (cit. on pp. 3007, 3017).

Daniel Yekutieli (June 2012). “Adjusted Bayesian inference for selected parameters”. en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, pp. 515–541 (cit. on p. 3021).

Ming Yuan and Yi Lin (2007). “Model selection and estimation in the Gaussian graphical model”. *Biometrika* 94.1, pp. 19–35 (cit. on pp. 3017, 3018).

Qing Zhou (Oct. 2014). “Monte Carlo Simulation for Lasso-Type Problems by Estimator Augmentation”. *Journal of the American Statistical Association* 109.508, pp. 1495–1516. arXiv: 1401.4425 (cit. on p. 3021).

Received 2017-12-01.

Jonathan E. Taylor

jonathan.taylor@stanford.edu