Biometrika (2024), **111**, 2, *pp.* 393–416
https://doi.org/10.1093/biomet/asad078
Advance Access publication 22 December 2023

On selection and conditioning in multiple testing and selective inference

By JELLE J. GOEMAN

Department of Biomedical Data Sciences, Leiden University Medical Center,
Einthovenweg 20, 2333 ZC Leiden, The Netherlands
j.j.goeman@lumc.nl

AND ALDO SOLARI

Department of Economics, Management and Statistics, University of Milano-Bicocca,
Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy
aldo.solari@unimib.it

Summary

We investigate a class of methods for selective inference that condition on a selection event. Such methods follow a two-stage process. First, a data-driven collection of hypotheses is chosen from some large universe of hypotheses. Subsequently, inference takes place within this data-driven collection, conditioned on the information that was used for the selection. Examples of such methods include basic data splitting as well as modern data-carving methods and post-selection inference methods for lasso coefficients based on the polyhedral lemma. In this article, we take a holistic view of such methods, considering the selection, conditioning and final error control steps together as a single method. From this perspective, we demonstrate that multiple testing methods defined directly on the full universe of hypotheses are always at least as powerful as selective inference methods based on selection and conditioning. This result holds true even when the universe is potentially infinite and only implicitly defined, such as in the case of data splitting. We provide general theory and intuition before investigating in detail several case studies where a shift to a nonselective or unconditional perspective can yield a power gain.

Some key words: False discovery rate; Familywise error rate; Multiple testing; Simultaneous inference.

1. Introduction

When many potential research questions are considered simultaneously, researchers often report only a subset of the findings, typically the most striking, interesting or surprising ones. When interpreting results selected in this way, it is crucial to recognize that the evidence for the findings may be exaggerated because of the selection process. The field of selective inference, also known as multiple testing, aims to adjust inference for this data-driven selection of research questions. Selective inference methods ensure that the number or proportion of incorrect findings among the final reported findings remains small. The selective inference literature is large and well established (Benjamini, 2010; Dickhaus, 2014; Taylor & Tibshirani, 2015; Taylor, 2018; Benjamini et al., 2019a; Cui et al., 2021; Kuchibhotla et al., 2021;

Zhang et al., 2022). Classic approaches in the field control either the familywise error rate or the false discovery rate.

Recently, a two-step approach to selective inference has gained popularity (Lee et al., 2016; Tibshirani et al., 2016; Fithian et al., 2017; Charkhi & Claeskens, 2018; Bi et al., 2020). In this conditional approach, the data are first used to select a small set of hypotheses of interest from a large universe of hypotheses. Next, inference is conducted on the selected hypotheses using the same data, but conditional on the information used for the selection. The conditional approach can be seen as a sophisticated generalization of data splitting. In data splitting, a portion of the subjects are used to select hypotheses, and the rest are used for inference on them. Conditional approaches similarly use part of the information in the data for selection and the remainder for inference. Proponents of conditional selective inference often contrast their approach with classic methods, suggesting that the conditional way of thinking represents the most fitting philosophy for selective inference, addressing the problem of selection in the most effective way. For example, as stated by Kuffner & Young (2018): 'The appropriate conceptual framework for valid inference is that discussed in the statistical literature as "post-selection inference", which […] requires conditioning on the selection event and control of the error rate of the inference given it was actually performed.'

Conditional selective inference methods return a selection-adjusted p-value for each of the selected hypotheses, or a selection-adjusted confidence interval for each of the selected parameters. The key property of these selection-adjusted measures, i.e., uniformity under the null hypothesis for p-values, or coverage for confidence intervals, holds conditional on the selection event. In the situation where more than one such p-value or confidence interval is returned, some authors argue for a further round of adjustment for multiple testing (e.g., Benjamini et al., 2019b), while others consider it an option (e.g., Hyun et al., 2021) or do not perform any further correction (e.g., Lee et al., 2016). Even when further multiple testing is carried out, however, this is generally not considered part of the conditional selective inference method itself, but simply a post-processing of the selection-adjusted p-values or confidence intervals returned by the method. This detachment of the selection and inference steps has been criticized as being circular, because the interpretation of selected but not significant hypotheses is not always clear (Weinstein & Ramdas, 2020, § B.1, supplement).

In this article we adopt an alternative, holistic perspective on conditional selective inference. We argue that any follow-up, in terms of multiple testing or lack thereof, on the selection-adjusted p-values should be regarded as an integral component of the selective inference method. From this point of view, conceptual differences between conditional selective inference and classic methods largely vanish. We argue that for every conditional selective inference method, there exists a method that is not selective and not conditional which always rejects all the hypotheses the original method rejects, and possibly more. We give several general conditions under which unconditional and nonselective methods are truly superior to selective conditional methods, and present several examples. Our results hold for methods returning selection-adjusted p-values or selection-adjusted confidence intervals, and apply to a variety of error rates. Proofs of all the propositions and lemmas can be found in the Supplementary Material.

2. Conditional selective inference: basics

Let $\mathrm{P} \in M$ be a probability measure, where M, the model, is a collection of probability measures defined on a common outcome space Ω. We will first focus on hypothesis testing, addressing confidence intervals in § 10. A hypothesis is a subset $H \subseteq M$, and H is true if $\mathrm{P} \in H$ and false otherwise. We have data X distributed according to P.

Conditional selective inference procedures consider a random collection of hypotheses. Sometimes we assume that we know the distribution of S, for example when S consists of the null hypotheses corresponding to the active set of a lasso regression. In other cases we may have only a realization of S without knowledge of its distribution, such as when S was chosen freely by a user on the basis of the first half of the data. In both cases, however, we assume that we know what part of the information in X was used to select S. In the lasso example we know this information because we know how S was calculated. In the data-splitting example we know that the user saw only part of the data.

We will illustrate our general discussion with a recurring toy example. Assume that two p-values P_1 and P_2 are independent, and that $P_1 \sim \mathrm{Un}(0, 1)$ under hypothesis H_1 and $P_2 \sim \mathrm{Un}(0, 1)$ under H_2. A simple selective inference procedure could discard hypotheses for which the p-values are greater than some fixed λ. In this case we have $S = \{i : P_i \leqslant \lambda\}$. This is a situation considered by Zhao et al. (2019) and Ellis et al. (2020). A similar selection set would arise when performing inference based on the polyhedral lemma if the design is orthogonal (Reid et al., 2017).

The collection S is drawn from a larger universe of hypotheses, which often remains implicit in the selective inference literature. Let $\mathcal{S} = \{S(\omega) : \omega \in \Omega\}$ be the collection of all possible realizations of S. We define the universe U as all hypotheses that could have been in $\mathcal{S}$. Formally,

$$U = \bigcup_{\omega \in \Omega} S(\omega).$$

Unlike S, the universe U is fixed. It can be huge, or even infinite. For example, when S consists of null hypotheses for the regression coefficients of the active set of a lasso regression, then U contains all null hypotheses for all regression coefficients for all covariates adjusted for all possible sets of other covariates (cf. Berk et al., 2013; Bachoc et al., 2020). In other cases U is even unknown. For example, if S was chosen freely by the user using half the data, then U contains all hypotheses the user would have chosen if the data were different. In this case we know nothing about U except that it is a superset of S. To avoid trivial problems, we assume that $U \neq \emptyset$. In the toy example we have $U = \{1, 2\}$.

Conditional selective inference methods define selection-adjusted p-values $p_{H|S}$ for $H \in S$. These have the property that for every $\alpha \in [0, 1]$,

$$\sup_{\mathrm{P} \in H} \mathrm{P}(p_{H|S} \leqslant \alpha \mid S) \leqslant \alpha. \tag{1}$$

The selection-adjusted p-value differs from the usual definition of the p-value p_H, namely that for every $\alpha \in [0, 1]$, $\sup_{\mathrm{P} \in H} \mathrm{P}(p_H \leqslant \alpha) \leqslant \alpha$, because it conditions on S. By conditioning on the selection event S, the selection-adjusted p-value discards the information used for that selection. It uses as evidence against the selected hypothesis H only the remainder of the information in the data. Conditioning thus provides a neat separation between the information used for selecting S and that used for inferring on the hypotheses in S. Condition (1) remains valid if we condition on more than just S, but Fithian et al. (2017) argued that it is optimal to condition on the minimal amount of information under which S is measurable. Indeed, several authors have reported a gain in power by conditioning on less information (Jewell et al., 2022; Carrington & Fearnhead, 2023; Chen et al., 2023).

There are many methods for calculating selection-adjusted p-values. The most straightforward way to achieve (1) is to separate the data into two independent components, $X = (X', X'')$, making sure that S is a function of X' only while $p_{H|S}$, for every $H \in S$, involves X'' only. This is the basic idea of data splitting (Moran, 1973; Cox, 1975; Rubin et al., 2006; Dahl et al., 2008; Wasserman & Roeder, 2009; Rinaldo et al., 2019). More sophisticated methods may use the data more efficiently by employing external randomization (Tian & Taylor, 2018; Panigrahi et al., 2023; Panigrahi & Taylor, 2023; Dharamshi et al., 2023; Leiner et al., 2023; Rasines & Young, 2023) or multiple data splits (Meinshausen et al., 2009; DiCiccio et al., 2020; Schultheiss et al., 2021). Some methods split the data adaptively, unmasking the data bit by bit until the user is ready to select the final set S and calculate the p-values conditional on that final S (Lei & Fithian, 2018; Duan et al., 2020). If an obvious split of the data is not available, the mathematics of the conditioning can become quite complex. The polyhedral lemma (Lee et al., 2016; Tibshirani et al., 2016), an important breakthrough, provides machinery to condition on selected sets arising in linear regression contexts, such as active sets from lasso regression. This result has been extended and applied in many contexts (Lee & Taylor, 2014; Yang et al., 2016; Tian & Taylor, 2017; Hyun et al., 2018; Liu et al., 2018; Taylor & Tibshirani, 2018; Heller et al., 2019; Panigrahi et al., 2021; Zhao et al., 2022; Garcia-Angulo & Claeskens, 2023).

In the toy example, we can calculate selection-adjusted p-values by looking at the conditional distribution of the p-values under the null hypothesis. If $i \in S$, we obtain $P_{i|S} = P_i/\lambda$. With a slight abuse of notation, we will write $i \in S$ instead of $H_i \in S$ and $P_{i|S}$ for $P_{H_i|S}$, which should cause no confusion. To adjust for the selection, the p-value has been multiplied by a factor of $1/\lambda$. It is easy to verify that whenever $i \in S$,

$$\mathrm{P}(P_{i|S} \leqslant t \mid S) = \mathrm{P}(P_i/\lambda \leqslant t \mid P_i \leqslant \lambda) = \lambda t/\lambda = t,$$

so that $P_{i|S}$ satisfies (1).

3. Multiple testing adjustment of selection-adjusted p-values

Having calculated selection-adjusted p-values, the usual next step is to decide which of the hypotheses in S can be rejected. A method must be decided for this, be it simply to reject all hypotheses with $p_{H|S} \leqslant \alpha$ for some α, or some more sophisticated multiple testing procedure. Whatever method is chosen, the end result is a random set $R \subseteq S$ of rejected hypotheses.

There are different opinions as to the properties the set R should have, but generally the focus is on avoiding false discoveries. Let

$$T_{\mathrm{P}} = \{H \in U : \mathrm{P} \in H\}$$

be the collection of all true hypotheses in U. Rejection of R induces $|R \cap T_{\mathrm{P}}|$ false discoveries, giving a false discovery proportion of

$$f_{\mathrm{P}}(R) = \frac{|R \cap T_{\mathrm{P}}|}{|R| \vee 1}.$$

To keep false discoveries in check, we can control the expectation of some error rate $e_{\mathrm{P}}(R)$, for which there are many choices (Benjamini, 2010; Benjamini et al., 2019a), such as $e_{\mathrm{P}}(R) =$

$f_{\mathrm{P}}(R)$ to control the false discovery rate FDR, $e_{\mathrm{P}}(R) = 1_{f_{\mathrm{P}}(R)>0}$ to control the familywise error rate FWER, or $e_{\mathrm{P}}(R) = 1_{f_{\mathrm{P}}(R)>\gamma}$ to control the false discovery exceedance rate FDX-γ. We assume that $0 \leqslant e_{\mathrm{P}}(R) \leqslant 1$ and that $e_{\mathrm{P}}(R) = 0$ whenever $R \cap T_{\mathrm{P}} = \emptyset$.

To control a chosen error rate, we bound its expectation by α. There are two flavours here. We can control the error rate conditional on S, requiring that for every $\mathrm{P} \in M$ and every $S \in \mathcal{S}$,

$$E_{\mathrm{P}}\{e_{\mathrm{P}}(R) \mid S\} \leqslant \alpha,$$

where $E_{\mathrm{P}}(\cdot) = \int_{\Omega} \cdot \, \mathrm{dP}$ is the expectation corresponding to P. Alternatively, we can aim for unconditional control, requiring that for every $\mathrm{P} \in M$,

$$E_{\mathrm{P}}\{e_{\mathrm{P}}(R)\} \leqslant \alpha.$$

Most researchers in conditional selective inference advocate control of the conditional error rate (Lee et al., 2016; Fithian et al., 2017; Kuffner & Young, 2018), though it has been shown that conditioning can sometimes be problematic (Kivaranovic & Leeb, 2021a,b). Other authors have argued for the unconditional error rate, sometimes finding that it leads to more power (Wu et al., 2010; Andrews et al., 2019, 2022). Indeed, the conditional error rate is the more stringent one, since conditional control implies unconditional control.

In the toy example, multiple testing is an issue only if $S = \{1, 2\}$. If we choose to control FWER at level α, we may use the methods of Hochberg (1988) and Hommel (1988), which are equivalent in the case of two hypotheses. This approach rejects each H_i if $P_{i|S} \leqslant \alpha/2$ and rejects both hypotheses if $P_{1|S}$ and $P_{2|S}$ are both at most α; the resulting procedure is displayed graphically in Fig. 1(a). Alternatively, we may choose to control FDR. With two hypotheses, the procedure of Benjamini & Hochberg (1995) is equivalent to the Hommel or Hochberg procedure just described and controls FWER as well as FDR. For controlling FDR we can do uniformly better with the minimally adaptive Benjamini–Hochberg procedure, MABH (Solari & Goeman, 2017). In the case of two hypotheses, this procedure also uniformly improves upon the adaptive procedure of Benjamini et al. (2006). MABH rejects each H_i if $P_{i|S} \leqslant \alpha/2$; it rejects both hypotheses if either $P_{1|S}$ and $P_{2|S}$ are both at most α, or the smaller is at most $\alpha/2$ and the larger at most 2α. The procedure is displayed graphically in Fig. 1(b).

So far we have assumed that the error rate depends only on R, but not on S. This assumption excludes the rate

$$e_{\mathrm{P}}(R, S) = \frac{|R \cap T_{\mathrm{P}}|}{|S| \vee 1} \tag{2}$$

that is implied by inference based on confidence intervals controlling the false coverage rate, FCR (Benjamini & Yekutieli, 2005). This is also the rate that is controlled if we perform no further multiple testing adjustment on the selection-adjusted p-values, but simply reject $R = \{i : P_{i|S} \leqslant \alpha\}$. This procedure is shown in Fig. 1(c). In the next few sections we will assume that the error rate is a function of R only, but we return to S-dependent error rates in § 12.

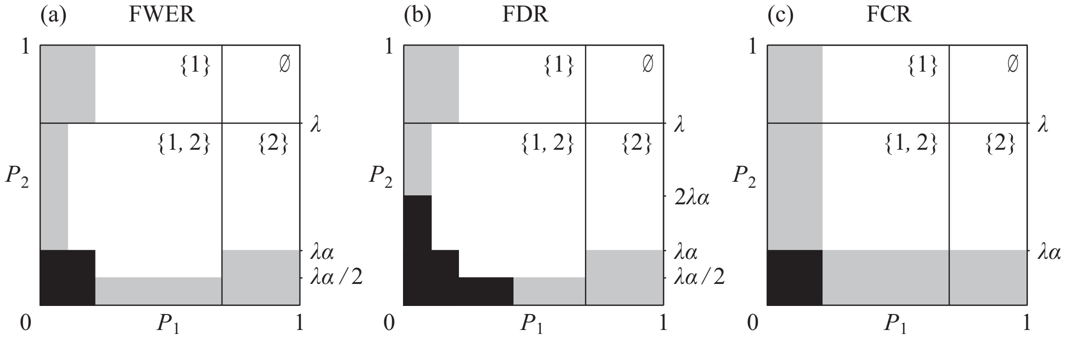

Fig. 1. A simple conditional selective inference procedure for two hypotheses inspired by Zhao et al. (2019) and Ellis et al. (2020). The procedure in (a) controls FWER, that in (b) FDR and that in (c) the FCR-inspired error rate (2). The set displayed in the upper right corner of each quadrant is the realization of S in that quadrant. Grey shading indicates areas in which one hypothesis is rejected; black indicates areas in which both hypotheses are rejected. The plot uses $\lambda = 0.7$ and $\alpha = 0.3$.

4. A holistic perspective and main observation

The approaches described in the previous two sections can be seen as two-stage methods. First, from a universe U of hypotheses a selection $S \subseteq U$ is made. Next, within that selection some hypotheses are rejected while others are not, and $R \subseteq S$ is returned. The set R is the final result of any method; it is the set we make inferential claims about.

Rather than analysing the two steps $U \to S$ and $S \to R$ separately, here we take a holistic perspective, viewing the two steps together as a single method $U \to S \to R$, or briefly $U \to R$. By viewing the two steps together we stress that the selection step $U \to S$ and the rejection step $S \to R$ are in the hands of the same analyst. The analyst chooses a method for the selection step $U \to S$ and a method for the inference step $S \to R$. The analyst also chooses what part of the information in the data to use for the selection step and what part of the data to reserve for the inference step.

From this holistic perspective, the choice of S in a procedure $U \to S \to R$ is, therefore, part of the method, and this part may be optimized. The holistic perspective implies that such optimization should be focused on obtaining a larger or more useful set R, since R, not S, represents the final inference of the method. In general, we would like to have as many rejections as possible, while keeping the chosen error rate under control. Moreover, from the holistic perspective all rejections of hypotheses in U are welcome, since every hypothesis in U could have been in S.

In the toy example, we can visualize the holistic view of the three procedures simply by removing all reference to S in Fig. 1, as shown in Fig. 2. This now displays three single-step procedures, defined directly on the universe $U = \{1, 2\}$ and based on the nonselection-unadjusted P_1 and P_2. The rejected sets R for the procedures in Fig. 2 are trivially identical to those of their counterparts in Fig. 1. However, in the holistic perspective of Fig. 2, the λ that previously determined S now becomes a tuning parameter, to be chosen freely by the analyst before seeing the data. The holistic perspective de-emphasizes the importance of S.

From the holistic perspective, we see that S plays two distinct roles in conditional selective inference. Firstly, S focuses the attention of the multiple testing procedure on hypotheses in S, restricting R to be a subset of S; this is the selective property of the procedure. Secondly, by conditioning on S the procedure ignores the information used to find S for the final inference; this is the conditional property of the procedure. We see both roles of S in the

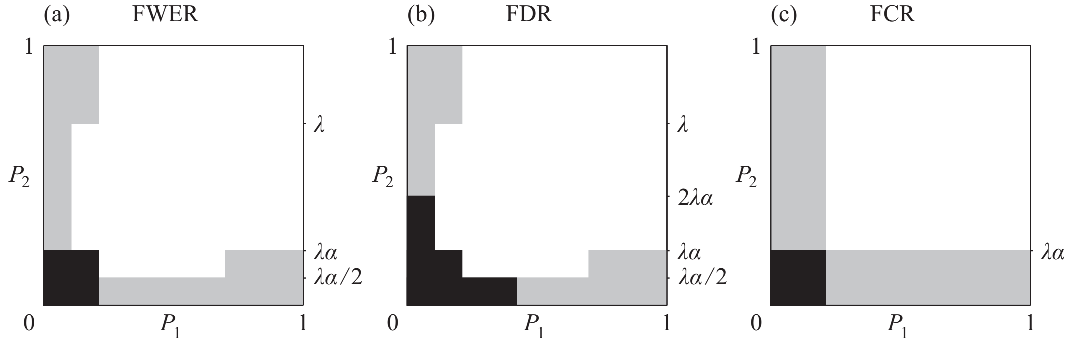

Fig. 2. Holistic perspective of the procedures in Fig. 1. Grey shading indicates areas in which one hypothesis is rejected; black indicates areas in which both hypotheses are rejected.

procedures of the toy example in Fig. 1. The procedure never rejects hypotheses outside S, so it is selective. We can see that the procedures are conditional, because the procedure in each S-defined quadrant is a valid multiple testing procedure by itself: if we were to stretch any quadrant so that it covers the entire unit square, we would obtain a method with valid FWER, FDR or FCR control, respectively.

The holistic perspective allows us to decouple the selective and conditional properties of conditional selective inference. We say that a procedure $U \to R$ is selective on S' if surely for all $\mathrm{P} \in M$, $R \subseteq S'$. We say that $U \to R$ is conditional on S'' if it controls its error rate conditionally on S'', i.e., if surely $E_{\mathrm{P}}\{e_{\mathrm{P}}(R, S'') \mid S''\} \leqslant \alpha$. By design, a conditional selective procedure $U \to S \to R$ is selective on S and conditional on S. However, the same procedure may be selective or conditional on sets it was not constructed around. Procedures are always selective on sets that are surely larger than S, and every procedure is, trivially, selective on R. Every procedure that is conditional on S is also conditional on $U \setminus S$, since S and $U \setminus S$ carry the same information. In Fig. 2 one can verify that all three procedures are conditional and selective on, for example, $S' = \{i \colon P_i \leqslant (1+\lambda)/2\}$.

In an important special case, every procedure is selective on U, since $R \subseteq U$ by definition. Moreover, every procedure is conditional on U, since the conditional error rate for U is the unconditional error rate, and control of any conditional error rate implies control of the unconditional error rate. This brings us to our first main observation: for every conditional selective multiple testing procedure on S there exists a conditional selective procedure on U, i.e., an unconditional, nonselective procedure that always rejects at least as many hypotheses.

Observation 1. *Let $U \to S \to R$ be a conditional selective inference procedure with the properties that $R \subseteq S$ surely and that $E_{\mathrm{P}}\{e_{\mathrm{P}}(R) \mid S\} \leqslant \alpha$ surely for all $\mathrm{P} \in M$. Then there exists a procedure $U \to R'$ such that $R' \supseteq R$ surely and $E_{\mathrm{P}}\{e_{\mathrm{P}}(R')\} = E_{\mathrm{P}}\{e_{\mathrm{P}}(R') \mid U\} \leqslant \alpha$ for all $\mathrm{P} \in M$.*

To prove Observation 1, simply take $R' = R$ and observe that $E_{\mathrm{P}}\{e_{\mathrm{P}}(R)\} = E_{\mathrm{P}}[E_{\mathrm{P}}\{e_{\mathrm{P}}(R) \mid S\}]$. We call Observation 1 an observation rather than a theorem or proposition, because as a mathematical result it is completely trivial: if we do not restrict to $R \subseteq S$, but allow the method also to reject hypotheses in $U \setminus S$, it may achieve more rejections that way; if we do not condition on S, we retain more information for finding a possibly larger R. Observation 1 is merely an immediate consequence of the holistic perspective we have adopted.

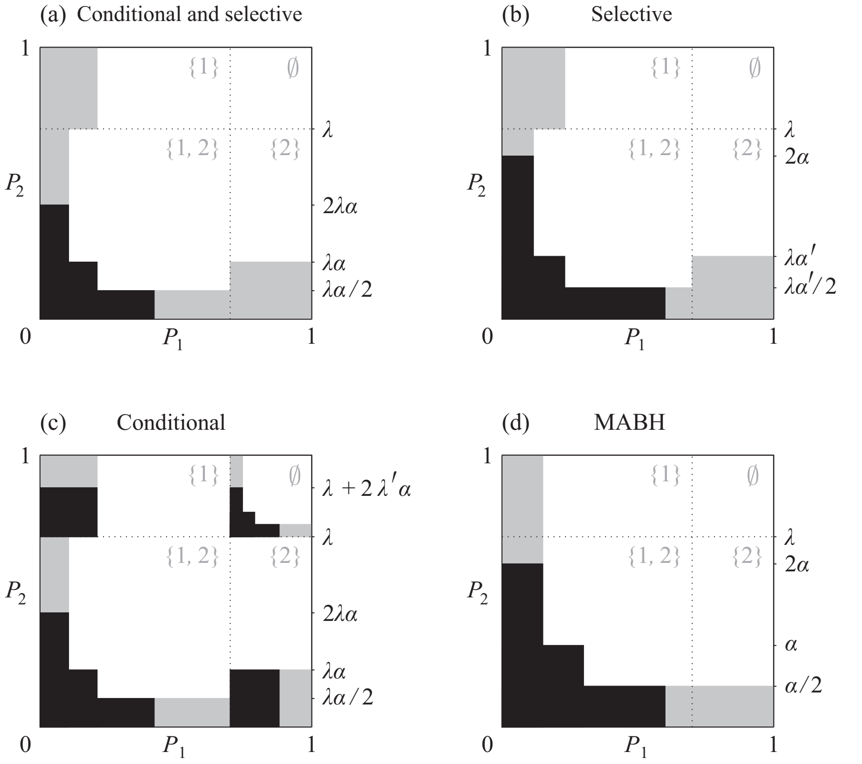

Fig. 3. (a) The conditional selective procedure of the toy example, controlling FDR, with its (b) selective and (c) conditional improvements, as well as (d) the MABH procedure shown as a reference. Grey shading indicates areas in which one hypothesis is rejected; black indicates areas in which both hypotheses are rejected. Here $\lambda' = 1 - \lambda$ and $\alpha' = \alpha/(2\lambda - \lambda^2)$.

However, Observation 1 answers the important question of how much of the information in the data to allocate to the selection step $U \to S$ and how much to the rejection step $S \to R$. According to Observation 1, the optimal choice is always simply to take $S = U$. Without losing power, we can allocate zero information to the selection step and retain all the information for the rejection step. This is an important insight.

5. First example: the toy example

Observation 1 says that a holistic method $U \to R'$ always exists that is at least as powerful, in the sense that $R' \supseteq R$, as a conditional selective procedure $U \to S \to R$. However, it does not show that achieving a true improvement is always possible; nor does it show how to find such an improvement if it exists. Nevertheless, there are many cases in which substantial improvement over a conditional selective procedure is possible.

In this section we illustrate this with the toy example of Fig. 1, focusing on its FDR-controlling variant. The toy example will help us gain intuition for the general case. As a preview, Fig. 3 displays in (a) the FDR-controlling conditional selective procedure, with two uniform improvements in (b) and (c). The procedure in (c) is not selective on S, sometimes rejecting hypotheses outside S, but it still controls FDR conditional on S. The procedure in (b) is still selective on S, guaranteeing $R \subseteq S$, but only has unconditional FDR control. The standard MABH procedure is given in (d) for comparison.

How did we arrive at these improvements? For the conditional improvement in (c), we continue to aim for control of FDR conditional on S, but we allow the procedure to reject hypotheses in $U \setminus S$. To do this, we also calculate selection-adjusted p-values $P_{i|S}$ for $i \notin S$, obtaining

$$P_{i|S} = \begin{cases} P_i/\lambda, & i \in S, \\ (P_i - \lambda)/(1-\lambda), & i \notin S. \end{cases}$$

While the selection-adjusted p-values are larger than the original ones for $i \in S$, the reverse is true when $i \notin S$. Next, we extend the procedure by continuing to test hypotheses in $U \setminus S$ after all hypotheses in S are rejected. If $S = \{1, 2\}$, the procedure is not changed. If $S = \{1\}$ and H_1 was rejected, we may continue to test H_2, rejecting when $P_{2|S=\{1\}} \leqslant 2\alpha$, and analogously for $S = \{2\}$. This fixed-sequence procedure, conditional on $S = \{1\}$, is easily seen to be valid for FDR control and is related to fixed-sequence FDR-controlling procedures proposed by Farcomeni & Finos (2013) and Lynch et al. (2017). If $S = \emptyset$, rather than rejecting nothing, we may use a MABH procedure on $P_{1|S=\emptyset}$ and $P_{2|S=\emptyset}$.

The resulting procedure, quite a strange one, is given in Fig. 3(c). It consists of four miniature multiple testing procedures, applied to conditional p-values, that are valid conditional on S for the four realizations of S. For $S = \{1, 2\}$ and $S = \emptyset$ we have a conditional MABH; for $S = \{1\}$ and $S = \{2\}$ we have a fixed-sequence FDR-controlling procedure, prioritizing the hypothesis in S. The resulting procedure clearly uniformly improves upon the procedure in Fig. 1. It does so by also considering hypotheses outside S for rejection. However, the improved procedure retains the property that it controls FDR conditional on S, since each of the miniature procedures is valid for FDR control.

A different type of improvement may be achieved if we are willing to give up on conditional FDR control. This is shown in Fig. 3(b). The improvement comes in two parts. First, we remark that the original procedure does not exhaust the α-level under the global null hypothesis: if $H_1 \cap H_2$ is true, FDR is controlled at level $(2\lambda - \lambda^2)\alpha$. We can therefore gain power by starting the procedure at level $\alpha' = \alpha/(2\lambda - \lambda^2)$ instead of at α. Secondly, after the original procedure has rejected H_1, it rejects H_2 if $P_{2|S=\{1,2\}} \leqslant 2\alpha$, i.e., when $P_2 \leqslant 2\lambda\alpha$. If we are not performing conditional control, however, there is no need to use the conditional p-value, and we may alternatively reject H_2, after we have rejected H_1, simply if $P_2 \leqslant 2\alpha$. The procedure resulting from these two improvements is given in Fig. 3(b). This procedure's FDR control is not conditional on S anymore, but it remains selective on S, assuming $\lambda \geqslant 2\alpha$. The validity of this new procedure may not be immediately obvious and is stated in the following lemma.

Lemma 1. *Suppose that P_1 and P_2 are independent and standard uniform under H_1 and H_2, respectively. Without loss of generality, assume that $P_1 \leqslant P_2$. Let $0 \leqslant \lambda \leqslant 1$ and $\alpha' = \alpha/(2\lambda - \lambda^2)$. Define a procedure that rejects H_1 when $P_1 \leqslant \lambda\alpha'/2$, or when $P_1 \leqslant \lambda\alpha'$ and $P_2 \leqslant \lambda\alpha'$, or when $P_1 \leqslant \lambda\alpha'$ and $P_2 > \lambda$, and that rejects H_2 when H_1 is rejected and $P_2 \leqslant 2\alpha$. This procedure controls* FDR *at level α.*

We have constructed two improvements of the conditional selective procedure we started with. One of these procedures retains the property of the original procedure that it controls FDR conditional on S, while the second retains the property that it only rejects hypotheses in S. The holistic perspective, however, does not care about S or about properties relating to S. It sees these two new methods simply as uniform improvements of the original that never reject fewer hypotheses and sometimes reject more. One of these, that in Fig. 3(c), is arguably somewhat strange and difficult to motivate from a holistic perspective; compare with the test of Berger (1989) improving the likelihood ratio test and the discussion in Perlman & Wu (1999). The procedure in Fig. 3(b) seems more reasonable.

As a fourth procedure, in Fig. 3(d) we show the regular MABH procedure, which does not attempt to be conditional or selective on S. This might be the procedure we would have chosen if we had adopted a holistic perspective from the beginning. In this particular case, MABH actually happens to be selective on S, as long as $\lambda \geqslant 2\alpha$. Comparing the conditional procedure in Fig. 3(c) with MABH, we see a massive shift of power away from $S = \{1, 2\}$ towards $S = \{1\}$, $S = \{2\}$ and $S = \emptyset$. Comparing the selective procedure in Fig. 3(b) with MABH, we see that while both procedures are selective, the original MABH still focuses relatively more power on $S = \{1, 2\}$, whereas the procedure in 3(b) still has a relatively strong focus on small sets S. This focus actually chimes with the motivation of the procedure we started from: Zhao et al. (2019) and Ellis et al. (2020) advocated their method for an application context in which null p-values tend to be near 1, so that $S = \{1\}$ and $S = \{2\}$ are relatively likely.

The comparison with MABH also serves to illustrate that uniformly improving a method $U \to S \to R$ by $U \to U \to R'$, with the requirement that $R' \supseteq R$, is not usually a question of simply adjusting the tuning parameter λ in such a way that S becomes U. The MABH procedure, resulting from the choice of $\lambda = 1$ in the conditional selective method, will be a more powerful method in many situations, but it is not a uniform improvement of the original method unless $\lambda \leqslant 1/2$. Generally, finding a true uniform improvement, in the sense that $R' \supseteq R$ surely for all $\mathrm{P} \in M$, involves much more work than merely adjusting a tuning parameter.

Comparing the conditional selective procedure and its two improvements, we see that the conditional selective procedure is exactly the intersection of its conditional and its selective improvements: it rejects either of H_1 and H_2 if and only if both the selective and the conditional improvements do. Compared with the conditional selective procedure, the selective improvement may have additional rejections if $S = \{1, 2\}$, while the conditional improvement cannot. On the other hand, the conditional improvement may have more rejections if $S = \emptyset$, while the selective procedure remains powerless there. If $S = \{1\}$ or $S = \{2\}$, both procedures may have additional rejections over the conditional selective procedure. However, the selective procedure has more chance of rejecting the hypothesis in S, while the conditional procedure may additionally reject a hypothesis outside S. The two improvements are, in this sense, disjoint.

The two improvements in Fig. 3 are easy to generalize to the case of more than two null hypotheses. They illustrate an important general principle about selection and conditioning in multiple testing. This principle says that selection and conditioning pull a procedure in opposite directions. Conditioning forces a procedure to distribute its power evenly over the outcome space, since the procedure must have proper error control on all realizations of S, conditional on S. Selection, on the other hand, focuses the power of a procedure away from hypotheses in $U \setminus S$, since it restricts rejections to S. A procedure that is both selective and conditional must therefore necessarily focus power both away from S and away from $U \setminus S$. Since there is nowhere for the power to go, it vanishes. The conditional selective procedure in Fig. 3(a), being the intersection of a conditional and a selective procedure, is therefore suboptimal as either. It is definitely suboptimal from the holistic perspective.

6. (In)admissibility conditions

Having looked in detail at a small example, we now come back to the general case. We will give some sufficient conditions under which uniform improvements exist.

We say that a conditional selective inference procedure $U \to S \to R$ is inadmissible if $U \to R'$ exists that uniformly improves upon $U \to S \to R$ in the sense that $R \subseteq R'$ surely for all $\mathrm{P} \in M$ and $\mathrm{P}(R \subset R') > 0$ for at least one $\mathrm{P} \in M$, while still controlling the error rate, i.e., $E_{\mathrm{P}}\{e_{\mathrm{P}}(R')\} \leqslant \alpha$. We will be a bit more precise and say that $U \to S \to R$ is inadmissible as a selective method on S if the uniform improvement still satisfies $R' \subseteq S$ surely. Similarly, $U \to S \to R$ is said to be inadmissible as a conditional method on S if the uniform improvement still controls its error rate conditional on S. Remember, however, that from the holistic perspective we do not care too much about S or about these subclasses of inadmissibility.

Our definition of a uniform improvement is very strict, as in Goeman et al. (2021), requiring that $R' \subseteq R$ for every outcome $\omega \in \Omega$. A uniform improvement, therefore, can never fail to reject a hypothesis that the method it improves upon does reject. This requirement makes admissibility a very low bar to achieve. For example, a FWER-controlling method that rejects all hypotheses in U with probability α, independently of the data, and rejects nothing with probability $1-\alpha$ is admissible according to our definition. Since admissibility is so easy to achieve, inadmissibility is particularly bad news.

We will give several sufficient conditions for inadmissibility of conditional selective methods. Propositions 1–3 apply to any error rate. Proposition 4 is only for FWER control.

Proposition 1. *If $\delta > 0$ is known such that $\mathrm{P}(S \cap T_{\mathrm{P}} = \emptyset) \geqslant \delta$ for all $\mathrm{P} \in M$, then $U \to S \to R$ is inadmissible as a selective procedure on S, unless $R = S$ surely for all $\mathrm{P} \in M$.*

In other words, Proposition 1 says that any conditional selective procedure is inadmissible if, with positive probability, the selection step results in a set S without true hypotheses; for examples see Al Mohamad et al. (2020), Ellis et al. (2020) and Heller & Solari (2023). In this case, it is impossible to make false discoveries, and the α for such S can be better spent elsewhere. The condition of the proposition implies that S has FWER control at level δ, but allows $\delta > \alpha$. The proposition does not apply when $R = S$ surely, but we will come back to that case in Observation 4 in § 12.

Proposition 2. *If $\mathrm{P}(S = \emptyset) > 0$ for some $\mathrm{P} \in M$, then $U \to S \to R$ is inadmissible as a conditional procedure on S. It is inadmissible as a selective procedure on any S' for which $S' \supseteq S$ surely for all $\mathrm{P} \in M$ and $S' \neq \emptyset$ surely for all $\mathrm{P} \in M$.*

Proposition 2 says that a conditional selective procedure may be improved if it sometimes selects $S = \emptyset$. There is a subtle but important difference from Proposition 1: if $\mathrm{P}(S = \emptyset) > 0$ for all $\mathrm{P} \in M$, then we would fulfil the conditions for Proposition 1, but Proposition 2 requires only that this happens for at least one $\mathrm{P} \in M$. Intuitively, if $S = \emptyset$ sometimes, we can make no errors in that case, and we can spend the α allocated to that case elsewhere.

Proposition 3. *If α' is known such that*

$$\alpha' = \sup_{\mathrm{P} \in M} E_{\mathrm{P}}\{e_{\mathrm{P}}(R)\} < \sup_{\mathrm{P} \in M} E_{\mathrm{P}}\left[\sup_{\mathrm{P} \in M} E_{\mathrm{P}}\{e_{\mathrm{P}}(R) \mid S)\}\right] \tag{3}$$

and $\mathrm{P}(R = S) < 1$ for at least one $\mathrm{P} \in M$, then $U \to S \to R$ is inadmissible as a selective method.

To understand Proposition 3, note that the left-hand side of (3) is equal to

$$\sup_{\mathrm{P}\in M} E_{\mathrm{P}}\big[E_{\mathrm{P}}\{e_{\mathrm{P}}(R) \mid S)\}\big],$$

so that (3) holds with $\leqslant$ by definition. Unconditional control bounds the left-hand side of (3) by α, while conditional control implies that the right-hand side of (3) is bounded by α. Any gap between the two can be exploited by an unconditional test to gain power. Such a gap may arise if the 'worst case' P, for which the conditional α-level is exhausted, depends on S. We give an example in the Supplementary Material.

Proposition 4. *If $U \to S \to R$ controls* FWER *conditional on S and there exists* $\mathrm{P} \in M$ *such that* $\mathrm{P}(R = S \mid S) > 0$ *for some $S \subset U$, then $U \to S \to R$ is inadmissible as a conditional procedure on S.*

Proposition 4 exploits the sequential rejection principle (Goeman & Solari, 2010), which says that if we reject all hypotheses under consideration, we may recycle the α and continue testing with a new batch. For a conditional selective procedure, this means that if we have exhausted all hypotheses in S, we may continue testing hypotheses in $U \setminus S$.

In the toy example, we see that the conditions of Propositions 1, 2 and 4 are all fulfilled, provided that $\lambda < 1$. The probability that we select only false null hypotheses is $(1-\lambda)^2$, $1-\lambda$ or 1 in the situations where 2, 1 or 0 hypotheses are true, respectively, so the condition of Proposition 1 is fulfilled with $\delta = (1-\lambda)^2$. Under $\mathrm{P} \in H_1 \cap H_2$ we have $\mathrm{P}(S = \emptyset) = (1-\lambda)^2 > 0$, so the condition of Proposition 2 is also fulfilled. Finally, if FWER was controlled, take $S = \emptyset$; then all hypotheses in S are rejected with positive probability for every $\mathrm{P} \in M$, conditional on $S = \emptyset$. It may seem from this checking of the conditions that the crucial characteristic that makes the procedure in the toy example inadmissible is the fact that it selects $S = \emptyset$ with positive probability. However, this is not the only driving factor. For example, perhaps the most important improvement of the procedure in Fig. 3(a) over that in Fig. 3(b) is the increase of the critical value from $2\lambda\alpha$ to 2α for rejecting the second hypothesis after rejecting the first. This change is not tied to the selection of $S = \emptyset$ in any way. The propositions of this section are sufficient conditions for inadmissibility, but they are by no means necessary. We will see examples of improvements of procedures that never select $S = \emptyset$ in § 7 and § 8.

The propositions in this section should be seen as examples of classes of procedures that could be improved by letting go of selection and conditioning. The emphasis was on uniform improvements. Often, procedures may be constructed that do not necessarily uniformly improve upon the original, but are substantially more powerful for relevant alternatives. An example is the standard MABH in the toy example, which, although not a uniform improvement over the original, has much larger rejection regions for both H_1 and H_2.

7. Second example: conditioning on the winner

The toy example considered thus far may seem to hinge much on the property that it selected $S = \emptyset$ with positive probability. In this section we look at a situation where $\mathrm{P}(S = \emptyset) = 0$ for all $\mathrm{P} \in M$.

The hypotheses that attract the most attention in the literature are generally those with the smallest p-values. It is of interest, therefore, to consider selection rules based on ranks. Selective inference for such selections, so-called 'inference on winners', has been considered by Zhong & Prentice (2008), Reid et al. (2017), Fuentes et al. (2018), Andrews et al. (2022), Zrnic & Fithian (2023) and Zrnic & Jordan (2023). We consider the simplest set-up here, where we select only a single 'winner'. In this set-up, we consider the question of whether the winner is truly nonnull.

Let $P_1, \dots, P_n$ be independent p-values, standard uniform under their respective null hypotheses $H_1, \dots, H_n$, so that $U = \{1, \dots, n\}$. We consider the selection rule that selects the single hypothesis for which the p-value is smallest, with ties broken arbitrarily, so that $|S| = 1$ always.

If we want to condition on the selection event $S = \{i\}$, we cannot simply reject for small values of P_i, adjusting the critical value for the selection event as we did in the toy example of Fig. 1. To see why this would be problematic, consider a set-up with $n = 2$ in which H_1 is null, but H_2 is not. Then

$$\mathrm{P}(P_1 \leqslant t \mid S = \{1\}) = \frac{\mathrm{P}(P_1 \leqslant t, P_1 \leqslant P_2)}{\mathrm{P}(P_1 \leqslant P_2)} = \frac{\mathrm{P}(P_1 \leqslant P_2 \wedge t)}{\mathrm{P}(P_1 \leqslant P_2)} = \frac{E_{\mathrm{P}}(P_2 \wedge t)}{E_{\mathrm{P}}(P_2)}. \tag{4}$$

Since P_2 is under the alternative, its distribution is arbitrary, so it could be uniform on $[0, t]$. In that case, (4) evaluates to 1. Hence, for every $t > 0$ there exists a $\mathrm{P} \in M$ such that $\mathrm{P}(P_1 \leqslant t \mid S = \{1\}) = 1$. Therefore, it is impossible to bound (4), in supremum over $\mathrm{P} \in M$, by α. Consequently, it is impossible to construct a conditional selective procedure that rejects for small values of P_i.

A way around this conundrum was offered by Reid et al. (2017), who proposed using an alternative test statistic $P_{i|S=\{i\}} = P_i / \min_{j \neq i} P_j$. Conditional on $S = \{i\}$, we have that $P_i / \min_{j \neq i} P_j$ is standard uniform for all $\mathrm{P} \in H_i$, as Lemma 2 states. Based on this lemma we can construct a conditional selective inference procedure. It rejects H_i $(i \in S)$ when $P_i / \min_{j \neq i} P_j \leqslant \alpha$. We call this Procedure A.

Lemma 2. *If* $n \geqslant 2$ *and* $\mathrm{P} \in H_i$, *then* $P_i / \min_{j \neq i} P_j \sim \mathrm{Un}(0, 1)$ *given* $S = \{i\}$.

What error rate does this conditional procedure on S control? On a family S of only one hypothesis, unadjusted testing, FCR, FWER and FDR control are all identical; Procedure A, therefore, controls all these error rates simultaneously. To construct potential improvements of the method, we must therefore decide which error rate to retain control of. We choose FDR for this example.

As in § 5, we construct three alternative procedures. The first, Procedure B, retains validity conditional on S, but possibly rejects hypotheses outside S. The second, Procedure C, will have unconditional FWER control, but still rejects only hypotheses within S. The third, Procedure D, will be fully unconditional and defined on U.

To construct procedure B, we must extend the notion of conditional p-values for H_j with $j \notin S$. We need the following lemma.

Lemma 3. *If* $n \geqslant 2$ *and* $\mathrm{P} \in H_j$ *for* $j \neq i$, *then* $(P_j - P_i)/(1 - P_i) \sim \mathrm{Un}(0, 1)$, *independent of* $(P_k)_{k \neq j}$, *given* $S = \{i\}$.

We will use $P_{i|S} = (P_j - P_i)/(1 - P_i)$ for $j \neq i$. As in § 5, we see that adjustment for nonselection results in p-values that are smaller than their unadjusted counterparts, rather than larger. Procedure B will be a two-step method based on these selection-adjusted p-values. Let i be such that $S = \{i\}$. Then the procedure first tests H_i, rejecting if $P_i / \min_{j \neq i} P_j \leqslant \alpha$. If it fails to reject H_i, the procedure stops. Otherwise it continues with a Benjamini–Hochberg

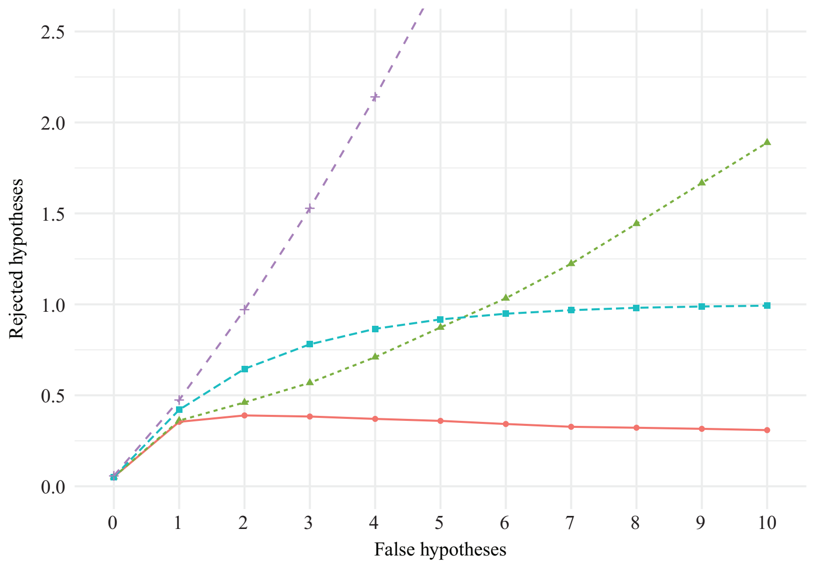

Fig. 4. Expected number of rejections for the four methods defined in § 7, Procedures A (red solid line with circles), B (green dotted line with triangles), C (blue dashed line with squares) and D (purple dashed line with crosses), based on $n = 100$ hypotheses, $\alpha = 0.05$ and 10^4 simulations.

procedure at level $\alpha' = n\alpha/(n-1)$ on the $n-1$ hypotheses H_j $(j \neq i)$, using $(P_j - P_i)/(1 - P_i)$ as p-values. This procedure clearly uniformly improves upon Procedure A if $n > 1$; its validity is stated in Lemma 4.

LEMMA 4. *Procedure B controls* FDR *given* $S = \{i\}$.

For Procedure C, we ignore the conditioning on $S = \{i\}$, but still restrict rejection to S only. This means that we can simply reject H_i for small P_i. By independence of the p-values, we may reject H_i when $P_i \leqslant 1 - (1 - \alpha)^{1/n}$. This is Procedure C. For Procedure D, the fully unconditional procedure, we simply choose the familiar Benjamini–Hochberg procedure.

While Procedure B uniformly improves upon Procedure A, the unconditional Procedures C and D do not. To see this, consider the situation where $P_2, \ldots, P_n$ are always equal to 1, which they could be under the alternative, or if null p-values are allowed to be stochastically larger than uniform. In that case, Procedures A and B reject H_1 if $P_1 \leqslant \alpha$, while Procedures C and D need $P_1 \leqslant 1 - (1 - \alpha)^{1/n}$ and α/n, respectively.

We compared the four procedures in a simple simulation. Out of 100 hypotheses, from 0 to 10 were considered to be under the alternative, yielding a p-value based on a one-sided normal test with a mean shift of 3; the remaining p-values were standard uniform. Figure 4 reports the expected number of rejected hypotheses for each of Procedures A, B, C and D. We see that the original conditional method, Procedure A, is very much directed towards sparse alternatives, even losing power as the density of the signal increases. In contrast, all the other methods gain power with increasing signal. The unconditional Procedure C, which, like Procedure A, only ever rejects the winner, rejects it with greater probability than Procedure A for all scenarios. The fully unconditional Benjamini–Hochberg method, although not a uniform improvement, is the clear overall winner, rejecting most hypotheses on average even in the sparse scenarios.

8. Data splitting and carving

Data splitting is perhaps the archetypal conditional selective inference method. It splits the data into two parts, using one part for selecting S and the other part for inference. Standard data splitting splits the data by subjects. Data carving is a more advanced version of data splitting (Fithian et al., 2017; Schultheiss et al., 2021; Panigrahi, 2023) that uses alternative ways of splitting the information in the data into independent parts, thus making more efficient use of the data. We show that data splitting and carving are inadmissible in general, at least for FWER control.

A special feature of data splitting is that the selection step that results in S is completely unconstrained, as long as the selection remains independent of the second part of the data. This implies that the universe U from which S was chosen is in principle infinite. The inadmissibility conditions in § 4 still apply, however. We have a simple corollary to Proposition 4, owing to the infinite nature of U. The inefficiency of data splitting has been noted by other authors. Jacobovic (2022) established inadmissibility of Moran's (1973) data-split test, and Fithian et al. (2017) have shown that data splitting yields inadmissible selective tests in exponential family models.

Proposition 5. *Data splitting is inadmissible as a selective method for* FWER *control if U is infinite and S is almost surely finite.*

Proposition 5 says that a data-splitting procedure is inadmissible because the analyst always runs the risk of selecting too few hypotheses for S. If all hypotheses in S are rejected, the classic data-splitting procedure must stop and loses out on some rejections it could have made. A uniform improvement would be a procedure that selects not just S, but an infinite sequence of pairwise-disjoint continuations $S_1, S_2, \ldots$. This procedure would always continue testing the next selected set after the previous one has been completely rejected. All of $S_1, S_2, \ldots$ must still be chosen using the first part of the data only. Control is therefore still conditional on the first part of the data.

Proposition 5 pertains to FWER control only. We conjecture that the same result holds for FDR, since FDR is by nature more lenient than FWER for making further rejections in $S_2, S_3, \ldots$ if it has already made many rejections, i.e., all of S_1. We do not have a general proof for this, but as an example consider FDR-controlling methods of the type discussed by Li & Barber (2017). These estimate FDR along an incremental sequence of potential rejection sets, rejecting the largest set for which the FDR estimate is less than α. Such procedures would gain power if the sequence is continued beyond S into $S_1, S_2, \ldots$.

With data splitting, the splitting of the data into two parts is arbitrary by nature, and the question of how much of the data to use for the selection and inference steps arises naturally. Some authors have proposed repeated splitting (Meinshausen et al., 2009; DiCiccio et al., 2020). Such methods are unconditional: while inference in each random split is conditional on the S from that split, control in the final analysis is unconditional. Multiple data splitting can, therefore, also be seen as an unconditional improvement of a conditional method.

9. Third example: data splitting

In § 8 we showed that data splitting is inadmissible as a conditional method for FWER control. If we are prepared to move away from conditional control, we can often improve methods further, although not always uniformly. We investigate a specific simple case in more detail.

Let $U = \{1, \ldots, n\}$ be finite, and suppose that the analysis on the two parts of the data results in pairs of independent p-values $\{P_{1,i}, P_{2,i}\}$ for H_i $(i = 1, \ldots, n)$. A natural choice for S is $S = \{i\colon P_{1,i} \leqslant \lambda\}$ for some fixed $0 \leqslant \lambda \leqslant 1$. With this choice, a conditional Bonferroni procedure would reject

$$R = \{i \in S\colon P_{2,i} \leqslant \alpha/|S|\}. \tag{5}$$

We can rewrite this as $R = \{i \in U\colon Q_i \leqslant \lambda\alpha/|S|\}$, with $Q_i = \lambda P_{2,i}$ if $P_{1,i} \leqslant \lambda$ and $Q_i = 1$ otherwise. Here, Q_i is a valid unconditional p-value, since $\mathrm{P}(Q_i \leqslant t) = \mathrm{P}(P_{1,i} \leqslant \lambda)\mathrm{P}(\lambda P_{2,i} \leqslant t) = \lambda \min(t/\lambda, 1) \leqslant t$. We could also have constructed an unconditional procedure on U based on the same Q_i. This would reject

$$R' = \{i \in U\colon Q_i \leqslant \alpha/n\}. \tag{6}$$

Comparing the conditional and unconditional procedures (5) and (6), we see that $R' \subseteq R$ whenever $|S| \leqslant \lambda n$, and $R' \supseteq R$ otherwise. The conditional procedure seemingly only has a chance to reject more than the unconditional if $|S|$ is smaller than its expectation under the complete null hypothesis with uniform p-values. The more signal in the data, the larger we would expect S to be, and the smaller the conditional R becomes relative to the unconditional R'. The conditional procedure has a chance to be better only if null p-values are stochastically larger than uniform. This argument generalizes immediately beyond Bonferroni to other symmetric monotone procedures. For example, the unconditional procedure of Benjamini & Hochberg (1995) on Q_i for $i \in U$ dominates its conditional equivalent on Q_i for $i \in S$ if $|S| > \lambda n$.

In the example just discussed, with $S = \{i\colon P_{1,i} \leqslant \lambda\}$, if λ was fixed a priori and $P_{1,1}, \ldots, P_{1,n}$ are independent, then we are not using all the information remaining after selecting S. Rather than splitting the data into $P_{1,1}, \ldots, P_{1,n}$ used for finding S and $P_{2,1}, \ldots, P_{2,n}$ used for testing, the data could be split into $1_{\{P_{1,1} \leqslant \lambda\}}, \ldots, 1_{\{P_{1,n} \leqslant \lambda\}}$ used for finding S and $P_{1,1|S}, \ldots, P_{1,n|S}$ and $P_{2,1}, \ldots, P_{2,n}$ used for testing. Such alternative splits are known as data carving. They tune the amount of information that is allocated to the selection and testing steps more efficiently. However, from the perspective of unconditional procedures, this still seems a rather convoluted way of combining the information from $P_{1,i}$ and $P_{2,i}$. A natural and more powerful choice would be, for instance, a Fisher combination, equivalent to rejecting for low values of $P_{1,i} \times P_{2,i}$, or, even more naturally, a single p-value calculated from a direct analysis of the combined data. Such analyses also obviate the need for choosing λ.

10. Selective confidence intervals and the false coverage rate

So far we have discussed mostly rejection of hypotheses based on p-values. However, a large part of the selective inference literature focuses on selection-adjusted confidence intervals, controlling the conditional FCR. In this section we apply the holistic perspective to selective inference based on confidence intervals.

A confidence interval is a random subset $C \subseteq M$ of the model space M. A confidence interval is said to have $(1 - \alpha)$-coverage if for all $\mathrm{P} \in M$,

$$\mathrm{P}(\mathrm{P} \in C) \geqslant 1 - \alpha.$$

We always define a confidence interval as a subset of the full parameter space. We can do this without loss of generality. For example, if our parameter space for $\theta = (\theta_1, \theta_2)$ is $\mathbb{R}^2$, we can write the confidence interval $[a, b]$ for θ_1 as the 'interval' $C = [a, b] \times \mathbb{R}$ for θ. This greatly simplifies notation. We will keep using the word interval, even though C can be any region.

In the selective inference context, we let $S \subseteq U$ be a random set of confidence intervals of interest, where U, as before, is the universe from which we are selecting. The collection of confidence intervals depends on S, and we write $C_{i|S}$ for $i \in S$. The confidence intervals should have conditional $(1-\alpha)$-coverage if for all $\mathrm{P} \in M$ and for $i \in S$,

$$\mathrm{P}(\mathrm{P} \in C_{i|S} \mid S) \geqslant 1 - \alpha. \tag{7}$$

If we report more than one confidence interval, we must account for multiplicity. We can demand that the confidence intervals be (conditionally) simultaneous over the selected event, i.e., surely for all $\mathrm{P} \in M$,

$$\mathrm{P}\Big(\mathrm{P} \in \bigcap_{i \in S} C_{i|S} \,\Big|\, S\Big) \geqslant 1 - \alpha, \tag{8}$$

where the unconditional variant drops the conditioning on S. Similarly, we can control FCR. The unconditional variant demands that for all $\mathrm{P} \in M$,

$$E_P\left[\frac{|\{i \in S\colon \mathrm{P} \in C_{i|S}\}|}{|S| \vee 1}\right] \geqslant 1 - \alpha. \tag{9}$$

Conditional on S, this simplifies to the requirement that surely for all $\mathrm{P} \in M$,

$$\frac{1}{|S| \vee 1} \sum_{i \in S} \mathrm{P}\big(\mathrm{P} \in C_{i|S} \,\big|\, S\big) \geqslant 1 - \alpha. \tag{10}$$

An attractive property of selection-adjusted confidence intervals is that they control FCR without further adjustment, since (7) implies (10); see also Weinstein et al. (2013, Theorem 2), Fithian et al. (2017, Proposition 11) and Lee et al. (2016, Lemma 2.1).

For confidence intervals we have the following analogue of Observation 1.

Observation 2. *If C_i with $i \in S$ control* (8) *or* (10) *conditionally on S, then there exist C_i' with $i \in U$ such that $C_i' \subseteq C_i$ for $i \in S$ surely and which control* (8) *or* (10)*, respectively, with $S = U$.*

This observation is, again, trivial. We simply take $C_i' = C_i$ if $i \in S$ and $C_i' = M$ otherwise. Like Observation 1, Observation 2 answers the question of what the optimal choice of S is if we are interested in confidence intervals that are as narrow as possible. The answer is that $S = U$ is the optimal choice.

Like Observation 1, Observation 2 does not say whether taking $S = U$ can actually help to shorten the confidence intervals. However, it is easy to find examples in which this is possible, certainly for FCR control. Take, for example, the original FCR-controlling method of Benjamini & Yekutieli (2005), which constructs marginal confidence intervals of level $1 - |S|\alpha/|U|$. For this method, taking $S = U$ clearly results in the narrowest confidence intervals. This observation holds generally for FCR control: as confidence intervals tend to become narrower as S becomes larger, there is every incentive for the analyst to choose S as large as possible, since this would yield both more and narrower confidence intervals. In the extreme case where $S = U$, FCR control reduces to average marginal coverage, an even weaker criterion than marginal coverage, which is achieved by uncorrected confidence intervals.

Specifically for the property of simultaneous over the selected event, we have the following additional observation.

Observation 3. *If C_i with $i \in S$ are unconditionally simultaneous over the selected S, then for every $S' \subseteq U$ there exist C_i' with $i \in S'$ which are unconditionally simultaneous over the selected S' and such that $C_i' \subseteq C_i$ surely for all $i \in S \cap S'$.*

To see that this observation is true, simply take $C_i' = C_i$ for $i \in S \cap S'$ and $C_i' = M$ for $i \in S' \setminus S$.

The observation says that any unconditional method that is simultaneous over the selected for some $S \subseteq U$ is also simultaneous over the selected on any other $S' \subseteq U$. This suggests, at least for unconditional methods, that simultaneous over the selected is not a different concept from just simultaneous over U, i.e., simultaneous.

11. Fourth example: post-selection inference for the lasso

One of the major application areas of conditional selective inference is post-selection inference on the parameters of a lasso model. A major breakthrough in this area has been the polyhedral lemma (Lee et al., 2016), which allows calculation of p-values and confidence intervals for regression coefficients, conditional on their selection by a lasso algorithm. The toy example of § 5 is in fact a special case of the approach of Lee et al. (2016), and we will not discuss it again. In this section we consider a variant due to Liu et al. (2018) of lasso-based selective inference, in which additional interesting issues arise.

The set-up is as follows. We consider the usual linear model setting, in which we have a fixed $n \times m$ design matrix X and assume that $Y = X\beta + \epsilon$, where the m-vector β is unknown and $\epsilon \sim N(0, \sigma^2 I_n)$, where σ^2 is assumed to be known. In this model we fit a lasso regression with a fixed penalty parameter λ. Let $\tilde{\beta}_i$ $(i = 1, \ldots, m)$ be the resulting coefficient estimates. We define the selected set as $S = \{i : \tilde{\beta}_i \neq 0\}$.

Liu et al. (2018) define selection-adjusted confidence intervals by conditioning, not on the full selected set S, but only on the selection of the confidence interval of interest. They require that for all $\mathrm{P} \in M$ and for $i \in S$,

$$\mathrm{P}(\mathrm{P} \in C_{i|i \in S} \mid i \in S) \geqslant 1 - \alpha. \tag{11}$$

Condition (11), although implied by (7), is substantially weaker, because it conditions on less information. In a part of their paper, Fithian et al. (2017) considered conditioning on $i \in S$ rather than on the full S for testing, recognizing that less conditioning leads to more information for inference. Liu et al. (2018) adopted this viewpoint for confidence intervals, arguing that by conditioning on this minimal event, more variation remains in the data for determining the precise value of β_i. The methodology of Jewell et al. (2022) and Neufeld et al. (2022) shares the 'general recipe' of Liu et al. (2018), stating that the ultimate goal is to satisfy equation (11) rather than (7) when it comes to selective inference.

Indeed, the conceptual difference between the two properties (11) and (7) is huge, but there is a steep price to pay for conditioning only on $i \in S$. Complications arise in subsequent

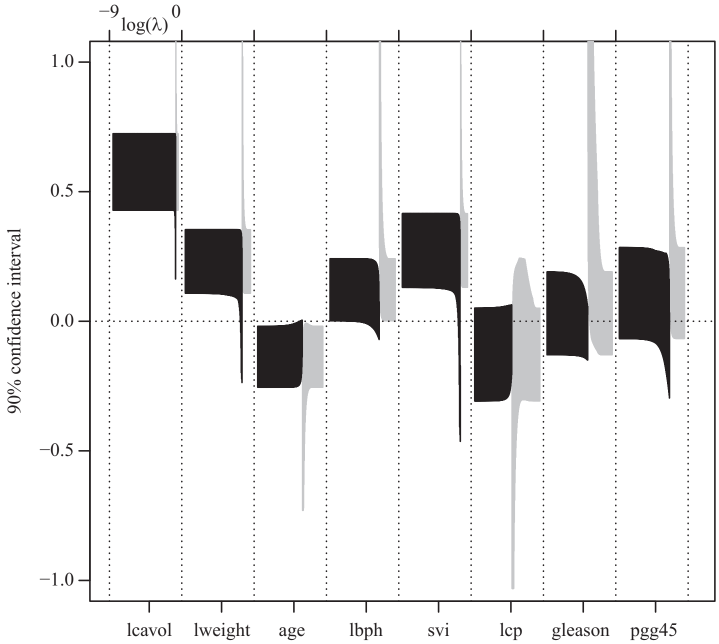

Fig. 5. Selective conditional confidence intervals obtained using the method of Liu et al. (2018) applied to the variables of the prostate data (Stamey et al., 1989), as a function of λ. Black intervals are conditional on selection by the lasso, while grey ones are conditional on nonselection.

error rate control because the coverage of each $C_{i|i\in S}$ is conditional on a different event for every $i \in S$. Because of this, the property, mentioned in § 10, that selection-adjusted coverage (7) implies FCR control (10) is lost: (11) does not imply (10) or even (9). Without a common conditioning event, there is no hope of combining the confidence intervals into any combined conditional error rate. For example, constructing $|S|$ confidence intervals, each conditional on $j \in S$, at level $1-\alpha/|S|$ does not guarantee simultaneous coverage, even unconditionally; we need confidence intervals at level $1-\alpha/m$ for that. In the Supplementary Material we give a numerical example showing lack of conditional and unconditional FCR control of the confidence intervals of Liu et al. (2018) at confidence level $1-\alpha$, and lack of conditional and unconditional simultaneous control at confidence level $1-\alpha/|S|$. Lack of FCR control of the method of Liu et al. (2018) was also observed in Panigrahi & Taylor (2023, Table 1), but without explanation.

By Observation 2, there is no reason to be selective and report confidence intervals for $i \in S$ only. Indeed, the premise of restricting attention to the selection of S is often that variables not in S are not important for the outcome. Confidence intervals or p-values for nonselected variables are an important instrument for checking this. It is straightforward to extend the theory of Liu et al. (2018) to calculate $C_{i|i\notin S}$ with $i \notin S$ for the nonselected regression coefficients, and we give the mathematical details in the Supplementary Material. Figure 5 displays 90% confidence intervals for all eight variables of the famous prostate dataset (Stamey et al., 1989) as a function of λ, with intervals for selected coefficients in black and those for nonselected ones in grey. We see a similar paradoxical effect as in the toy example: conditional intervals of selected variables tend to move towards 0, while confidence intervals for nonselected variables tend to move away from 0; see also the Supplementary Material. Both are equal to the unconditional intervals for very large or small

λ, when the probability of selection is close to 0 or 1, but tend to become longer close to the critical threshold for selection. Kivaranovic & Leeb (2021a,b) provide conditions under which intervals obtained from the polyhedral lemma are either bounded or unbounded. The intervals constructed by the method of Liu et al. (2018) have bounded lengths when they are conditional on selection, whereas the intervals are potentially unbounded when they are conditional on nonselection.

The intervals C_i, defined as $C_{i|i\in S}$ if $i \in S$ and $C_{i|i\notin S}$ if $i \notin S$, are unconditional intervals and, due to the absence of a common conditioning event, have no conditional interpretation as a collection. We may present them all as uncorrected intervals, but if we aim to present only a selection $V \subseteq \{1,\ldots,m\}$ from these intervals we must correct for this using methods for correcting unconditional intervals. We may use level $1-\alpha/m$ to obtain simultaneous coverage over the selected intervals, or we may employ the method of Benjamini & Yekutieli (2005) and use level $1-|V|\alpha/m$ to control FCR. This applies if $V = S$ or for any other V. There is no way in which the conditioning of the intervals on $i \in S$ helped for this correction step; in fact, it merely discarded valuable information, lengthening the intervals and moving them towards zero. Arguably, the superior method is simply to start from regular unconditional intervals. This does not provide a uniform improvement of the method of Liu et al. (2018), but it avoids the paradoxes associated with conditioning and tends to produce more attractive intervals.

12. False coverage rate for hypothesis testing

Confidence intervals can be used to test hypotheses, and the properties of confidence intervals imply error control guarantees on the hypotheses. In this section we look briefly into the error rate (2) implied by FCR control (10), which is used by some authors (e.g. Fithian et al., 2017). Assume that we have a collection H_i ($i \in S$) of hypotheses, one for every confidence interval.

If confidence intervals $C_{i|S}$ ($i \in S$) have conditional FCR control, then $R = \{i \colon H_i \cap C_{i|S} = \emptyset\}$ controls the error rate (10). Observation 1 does not directly apply, since the error rate depends not just on R in S. However, that observation immediately generalizes.

Observation 1 (continued). *Observation 1 also holds for error rates $e_{\mathrm{P}}(R, S)$ that depend on S, if $S \subseteq S'$ implies that $e_{\mathrm{P}}(R, S) \geqslant e_{\mathrm{P}}(R, S')$.*

The extra condition holds for the FCR rate (2). The condition implies that replacing S by U makes the error rate more lenient, so for controlling the error rate it helps to take $S = U$, and the result is still trivial. FCR is a paradoxical error rate from the holistic perspective, since it is decreasing in $|S|$ for the same R. This provides an immediate incentive for an analyst to choose S as large as possible.

FCR is sometimes motivated (Zhao & Cui, 2020) by the property that FCR control reduces to FDR control when $S = R$. For this property to hold, we must have that $S = R$ as random variables; it is not sufficient that the realized values are identical. Regarding conditional control of FCR, or other error rates, when $S = R$ as random variables we have the following observation. We say that a testing problem is trivial on $\mathcal{S}$ if $e_{\mathrm{P}}(V) \leqslant \alpha$ for all $\mathrm{P} \in M$ and all $V \in \mathcal{S}$, i.e., if the error rate is already bounded by α everywhere.

Observation 4. *Suppose that a conditional selective method $U \to S \to R$ has $R = S$ surely. Then the testing problem is trivial on $\mathcal{S}$.*

To see that this observation is true, notice that conditional control requires that $E_{\mathrm{P}}\{e_{\mathrm{P}}(R) \mid S\} \leqslant \alpha$ for all $\mathrm{P} \in M$ and all $S \in \mathcal{S}$. If $R = S$ surely, the inequality reduces to $e_{\mathrm{P}}(S) \leqslant \alpha$.

It follows from Observation 4 that only unconditional FCR-controlling methods can be used as a means to construct FDR-controlling methods; conditional FCR control has no relationship to FDR control.

13. Discussion

The literature on selective inference methods based on conditioning often takes the selected set of hypotheses S as given and presents the analyst's task solely as providing confidence intervals or p-values that are valid despite the random nature of S. In this article, we regard this as only the middle step of a bigger procedure, which first selects S from a universe U, then corrects for this selection, and finally uses the resulting p-values or confidence intervals to control an error rate of choice, leading to a final rejected set R. This holistic perspective is perhaps the most important contribution of the present work. All the results in the article are tied to this perspective.

If S is simply a step in a procedure that starts with a universe U and ends with a rejected set R, the question arises naturally as to what is the optimal amount of information to invest in choosing S. The simple answer is: none. For both primary roles of S, i.e., automatically accepting hypotheses not in S and discarding all information used to select S, the optimal choice is to choose S as large as possible.

Selection-adjusted p-values of confidence intervals are sometimes presented as the end result of a conditional selective inference procedure, suggesting that selection adjustment is sufficient to address the multiplicity problem. However, the error rate (2) thus controlled is equivalent to the per-comparison error rate, i.e., unadjusted testing, on S. It does not correct for the multiplicity of S itself. The larger S is, therefore, the more and the lower the selection-adjusted p-values will be. From the holistic perspective, there is every incentive for the analyst to choose S as large as possible, eventually reaching unadjusted testing when $S = U$. In our view, it is appropriate to present selection-adjusted p-values or confidence intervals without further multiple testing adjustment only if the choice of S is not under the control of the analyst, and only if unadjusted methods would have been appropriate if S were nonrandom and given a priori.

We have given several examples of uniform improvements of conditional methods by unconditional ones, as well as general conditions under which such improvements are possible. Some of these improvements are useful and substantial; others are small or may appear artificial. We do not have a general recipe for such improvements, and we emphasize that improvements are generally not unique. In several case studies we have constructed improved procedures that are either still selective, i.e., focusing power on a small and promising set S of hypotheses, or still conditional, i.e., valid conditional on the information used to find this same S. Invariably, we found that good selective procedures were not conditional, and good conditional procedures were not selective. Apparently, prioritizing hypotheses in S and conditioning on this prioritization are conflicting goals. A multiple testing procedure that focuses its power on a promising set S should exploit the information that S is a promising set; a conditional procedure discards the same information by conditioning on it.

Choosing $S = U$, as we advocate, essentially means reverting to unconditional, as opposed to more stringent conditional, error rates. In our view this is good enough: common unconditional error rates such as the familywise error are seldom criticized for being too lenient. Some authors (e.g., Kuffner & Young, 2018) have argued that it is better to control conditional error rates because they avoid unwarranted use of ancillary information. We find this difficult to accept as a general argument, since in most procedures S is not ancillary in the usual sense, but rather based on a bona fide summary of the available evidence in part of the data.

Finally, we remark that allowing inspection of the data prior to making inferential decisions is not exclusively the domain of conditional methods. In fact, simultaneous methods allow users to postpone some inferential decisions until after they have seen all the data (Goeman & Solari, 2011), something conditional methods could never allow.

Acknowledgement

Solari is also affiliated with the Department of Economics at the Ca' Foscari University of Venice, Italy.

Supplementary material

The Supplementary Material includes proofs of all the propositions and lemmas, as well as further examples.

References

Al Mohamad, D., Van Zwet, E. W., Cator, E. & Goeman, J. J. (2020). Adaptive critical value for constrained likelihood ratio testing. *Biometrika* **107**, 677–88.

Andrews, I., Bowen, D., Kitagawa, T. & McCloskey, A. (2022). Inference for losers. *Am. Econ. Assoc. Papers Proc.* **112**, 635–42.

Andrews, I., Kitagawa, T. & McCloskey, A. (2019). Inference on winners. Working Paper no. 25456, National Bureau of Economic Research, Cambridge, Massachusetts.

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

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biomet. J.* **52**, 708–21.

Benjamini, Y., Hechtlinger, Y. & Stark, P. B. (2019a). Confidence intervals for selected parameters. *arXiv:* 1906.00505.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Benjamini, Y., Krieger, A. M. & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

Benjamini, Y., Taylor, J. & Irizarry, R. A. (2019b). Selection-corrected statistical inference for region detection with high-throughput assays. *J. Am. Statist. Assoc.* **114**, 1351–65.

Benjamini, Y. & Yekutieli, D. (2005). False discovery rate: Adjusted multiple confidence intervals for selected parameters. *J. Am. Statist. Assoc.* **100**, 71–81.

Berger, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *J. Am. Statist. Assoc.* **84**, 192–9.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.* **41**, 802–37.

Bi, N., Markovic, J., Xia, L. & Taylor, J. (2020). Inferactive data analysis. *Scand. J. Statist.* **47**, 212–49.

Carrington, R. & Fearnhead, P. (2023). Improving power by conditioning on less in post-selection inference for changepoints. *arXiv:* 2301.05636v2.

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

Chen, Y., Jewell, S. & Witten, D. (2023). More powerful selective inference for the graph fused lasso. *J. Comp. Graph. Statist.* **32**, 577–87.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62**, 441–4.

Cui, X., Dickhaus, T., Ding, Y. & Hsu, J. C. (2021). *Handbook of Multiple Comparisons*. Boca Raton, Florida: CRC Press.

Dahl, F. A., Grotle, M., Šaltytė Benth, J. & Natvig, B. (2008). Data splitting as a countermeasure against hypothesis fishing: With a case study of predictors for low back pain. *Eur. J. Epidemiol.* **23**, 237–42.

Dharamshi, A., Neufeld, A., Motwani, K., Gao, L. L., Witten, D. & Bien, J. (2023). Generalized data thinning using sufficient statistics. *arXiv:* 2303.12931v2.

DiCiccio, C. J., DiCiccio, T. J. & Romano, J. P. (2020). Exact tests via multiple data splitting. *Statist. Prob. Lett.* **166**, 108865.

Dickhaus, T. (2014). *Simultaneous Statistical Inference: With Applications in the Life Sciences*. Heidelberg, Germany: Springer.

Duan, B., Ramdas, A. & Wasserman, L. (2020). Familywise error rate control by interactive unmasking. In *Proc. 37th Int. Conf. Machine Learning (ICML'20)*. JMLR, pp. 2720–9.

Ellis, J. L., Pecanka, J. & Goeman, J. J. (2020). Gaining power in multiple testing of interval hypotheses via conditionalization. *Biostatistics* **21**, e65–79.

Farcomeni, A. & Finos, L. (2013). FDR control with pseudo-gatekeeping based on a possibly data driven order of the hypotheses. *Biometrics* **69**, 606–13.

Fithian, W., Sun, D. & Taylor, J. (2017). Optimal inference after model selection. *arXiv:* 1410.2597v4.

Fuentes, C., Casella, G. & Wells, M. T. (2018). Confidence intervals for the means of the selected populations. *Electron. J. Statist.* **12**, 58–79.

Garcia-Angulo, A. C. & Claeskens, G. (2023). Exact uniformly most powerful postselection confidence distributions. *Scand. J. Statist.* **50**, 358–82.

Goeman, J. J., Hemerik, J. & Solari, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Statist.* **49**, 1218–38.

Goeman, J. J. & Solari, A. (2010). The sequential rejection principle of familywise error control. *Ann. Statist.* **38**, 3782–810.

Goeman, J. J. & Solari, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26**, 584–97.

Heller, R., Meir, A. & Chatterjee, N. (2019). Post-selection estimation and testing following aggregate association tests. *J. R. Statist. Soc.* B **81**, 547–73.

Heller, R. & Solari, A. (2023). Simultaneous directional inference. *arXiv:* 2301.01653v2.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–2.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–6.

Hyun, S., G'Sell, M. & Tibshirani, R. J. (2018). Exact post-selection inference for the generalized lasso path. *Electron. J. Statist.* **12**, 1053–97.

Hyun, S., Lin, K. Z., G'Sell, M. & Tibshirani, R. J. (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics* **77**, 1037–49.

Jacobovic, R. (2022). Simple sufficient condition for inadmissibility of Moran's single-split test. *Electron. J. Statist.* **16**, 3036–59.

Jewell, S., Fearnhead, P. & Witten, D. (2022). Testing for a change in mean after changepoint detection. *J. R. Statist. Soc.* B **84**, 1082–104.

Kivaranovic, D. & Leeb, H. (2021a). A (tight) upper bound for the length of confidence intervals with conditional coverage. *arXiv:* 2007.12448v2.

Kivaranovic, D. & Leeb, H. (2021b). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *J. Am. Statist. Assoc.* **116**, 845–57.

Kuchibhotla, A. K., Kolassa, J. E. & Kuffner, T. A. (2021). Post-selection inference. *Annu. Rev. Statist. Appl.* **9**, 505–27.

Kuffner, T. A. & Young, G. A. (2018). Principled statistical inference in data science. In *Statistical Data Science*. Singapore: World Scientific, pp. 21–36.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.

Lee, J. D. & Taylor, J. E. (2014). Exact post model selection inference for marginal screening. In *Proc. 27th Int. Conf. Neural Information Processing Systems (NIPS'14)*. Cambridge, Massachusetts: MIT Press, pp. 136–44.

Lei, L. & Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *J. R. Statist. Soc.* B **80**, 649–79.

Leiner, J., Duan, B., Wasserman, L. & Ramdas, A. (2023). Data fission: Splitting a single data point. *J. Am. Statist. Assoc.* to appear, DOI: 10.1080/01621459.2023.2270748.

Li, A. & Barber, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Am. Statist. Assoc.* **112**, 837–49.

Liu, K., Markovic, J. & Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv:* 1801.09037.

Lynch, G., Guo, W., Sarkar, S. K. & Finner, H. (2017). The control of the false discovery rate in fixed sequence multiple testing. *Electron. J. Statist.* **11**, 4649–73.

Meinshausen, N., Meier, L. & Bühlmann, P. (2009). p-Values for high-dimensional regression. *J. Am. Statist. Assoc.* **104**, 1671–81.

Moran, P. A. (1973). Dividing a sample into two parts: A statistical dilemma. *Sankhyā* A **35**, 329–33.

Neufeld, A. C., Gao, L. L. & Witten, D. M. (2022). Tree-values: Selective inference for regression trees. *J. Mach. Learn. Res.* **23**, 1–43.

Panigrahi, S. (2023). Carving model-free inference. *arXiv:* 1811.03142v5.

Panigrahi, S., Fry, K. & Taylor, J. (2023). Exact selective inference with randomization. *arXiv:* 2212.12940v4.

Panigrahi, S. & Taylor, J. (2023). Approximate selective inference via maximum likelihood. *J. Am. Statist. Assoc.* **118**, 2810–20.

Panigrahi, S., Zhu, J. & Sabatti, C. (2021). Selection-adjusted inference: An application to confidence intervals for cis-eQTL effect sizes. *Biostatistics* **22**, 181–97.

Perlman, M. D. & Wu, L. (1999). The emperor's new tests. *Statist. Sci.* **14**, 355–69.

Rasines, D. G. & Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika* **110**, 597–614.

Reid, S., Taylor, J. & Tibshirani, R. (2017). Post-selection point and interval estimation of signal sizes in Gaussian samples. *Can. J. Statist.* **45**, 128–48.

Rinaldo, A., Wasserman, L. & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47**, 3438–69.

Rubin, D., Dudoit, S. & Van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statist. Appl. Genet. Molec. Biol.* **5**, 19.

Schultheiss, C., Renaux, C. & Bühlmann, P. (2021). Multicarving for high-dimensional post-selection inference. *Electron. J. Statist.* **15**, 1695–742.

Solari, A. & Goeman, J. J. (2017). Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. *Biomet. J.* **59**, 776–80.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *J. Urology* **141**, 1076–83.

Taylor, J. & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proc. Nat. Acad. Sci.* **1122**, 7629–34.

Taylor, J. & Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Can. J. Statist.* **46**, 41–61.

Taylor, J. E. (2018). A selective survey of selective inference. In *Proc. Int. Congr. Mathematicians: Rio de Janeiro 2018*. Singapore: World Scientific, pp. 3019–38.

Tian, X. & Taylor, J. (2017). Asymptotics of selective inference. *Scand. J. Statist.* **44**, 480–99.

Tian, X. & Taylor, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46**, 679–710.

Tibshirani, R. J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Assoc.* **111**, 600–20.

Wasserman, L. & Roeder, K. (2009). High dimensional variable selection. *Ann. Statist.* **37**, 2178–201.

Weinstein, A., Fithian, W. & Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *J. Am. Statist. Assoc.* **108**, 165–76.

Weinstein, A. & Ramdas, A. (2020). Online control of the false coverage rate and false sign rate. In *Proc. 37th Int. Conf. Machine Learning (ICML'20)*. JMLR, pp. 10193–202.

Wu, S. S., Wang, W. & Yang, M. C. K. (2010). Interval estimation for drop-the-losers designs. *Biometrika* **97**, 405–18.

Yang, F., Foygel Barber, R., Jain, P. & Lafferty, J. (2016). Selective inference for group-sparse linear models. In *Proc. 30th Int. Conf. Neural Information Processing Systems (NIPS 2016)*. Red Hook, New York: Curran Associates.

Zhang, D., Khalili, A. & Asgharian, M. (2022). Post-model-selection inference in linear regression models: An integrated review. *Statist. Surv.* **16**, 86–136.

Zhao, H. & Cui, X. (2020). Constructing confidence intervals for selected parameters. *Biometrics* **76**, 1098–108.

Zhao, Q., Small, D. S. & Ertefaie, A. (2022). Selective inference for effect modification via the lasso. *J. R. Statist. Soc.* B **84**, 382–413.

Zhao, Q., Small, D. S. & Su, W. (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J. Am. Statist. Assoc.* **114**, 1291–304.

Zhong, H. & Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–34.

Zrnic, T. & Fithian, W. (2023). Locally simultaneous inference. *arXiv:* 2212.09009v4.

Zrnic, T. & Jordan, M. I. (2023). Post-selection inference via algorithmic stability. *Ann. Statist.* **51**, 1666–91.

[*Received on* 27 *July* 2022. *Editorial decision on* 5 *December* 2023]