We introduce a new class of methods for finite-sample false discovery rate (FDR) control in multiple testing problems with dependent test statistics where the dependence is noted as deleted or revised text. Our approach separately calibrates a data-dependent \( p \)-value rejection threshold for each hypothesis, relaxing or tightening the threshold as appropriate to target exact FDR control. In addition to our general framework we propose a concrete algorithm, the dependence-adjusted Benjamini–Hochberg (dBH) procedure, which thresholds the BH-adjusted \( p \)-value for each hypothesis. Under positive regression dependence the dBH procedure uniformly dominates the standard BH procedure, and in general it uniformly dominates the Benjamini–Yekutieli (BY) procedure (also known as BH with log correction), which makes a conservative adjustment for worst-case dependence. Simulations and real data examples show substantial power gains over the BY procedure, and competitive performance with knockoffs in settings where both methods are applicable. When the BH procedure empirically controls FDR (as it typically does in practice) the dBH procedure performs comparably.

Note the second revision is color-coded by added text, [DELETED: deleted text], and revised text, relative to the first revision.

1. Introduction. Despite the immense popularity of the false discovery rate (FDR) paradigm and the Benjamini–Hochberg (BH) method for large-scale multiple testing (Benjamini and Hochberg, 1995), the literature on FDR-controlling methods has long been dogged by their uncertain validity when applied to dependent \( p \)-values. In particular, the BH procedure is only known to control FDR under restrictive positive dependence assumptions, or after a severe correction to the significance level (Benjamini and Yekutieli, 2001). Apart from specific supervised learning settings where knockoff methods (Barber and Candès, 2015; Candès et al., 2018) can be applied, practitioners still commonly default to the uncorrected BH method, choosing to forego theoretical guarantees and rely instead on consistent empirical findings that the BH procedure is robust to the kinds of dependence that occur in practice (e.g. Farcomeni, 2006; Kim and van de Wiel, 2008).

This article introduces new methods for finite-sample FDR control under dependence. Our key technical idea is to decompose the FDR according to the additive contribution of each hypothesis, and use conditional inference to adaptively calibrate a separate rejection rule for each hypothesis to directly control its FDR contribution. Equipped with this tool, we prove finite-sample FDR control for a broad class of multiple testing methods. We also propose a concrete algorithm, the dependence-adjusted Benjamini–Hochberg procedure (dBH), that operates by adaptively calibrating a separate BH-adjusted \( p \)-value cutoff for each hypothesis. Specifically, our method rejects \( H_i \) if the adjusted \( p \)-value is below a calibrated threshold

**MSC2020 subject classifications:** Primary 62F03; secondary 62G10.

**Keywords and phrases:** false discovery rate, multiple testing, dependent test statistics, Benjamini–Hochberg procedure.
\( \hat{c}_i \), which may be larger or smaller than \( \alpha \). The dBH procedure can be applied in a wide variety of discrete and continuous, parametric and nonparametric models, provided that the analyst is able to exploit some knowledge of the dependence structure. The present work emphasizes multivariate Gaussian and linear regression models, but we also discuss Gaussian graphical models, binomial models, and permutation tests. We show empirically that our methods perform similarly to BH, but with provable FDR control.

Because \( \hat{c}_i \) can be larger than \( \alpha \), the dBH procedure can be, and often is, somewhat more powerful than the usual BH procedure. We show that dBH is uniformly more powerful than BH under positive dependence, in the sense that it makes at least as many rejections, almost surely. In addition, versions of the method are uniformly more powerful than the corrected version of BH (known as the Benjamini–Yekutieli (BY) procedure), which assumes worst-case dependence. By instead exploiting knowledge about the dependence structure, our method commonly achieves dramatic improvements in power relative to BY.

1.1. Multiple testing and the false discovery rate. In a multiple testing problem, an analyst observes a data set \( X \sim P \), and rejects a subset of null hypotheses \( H_1, \ldots, H_m \). We assume \( P \in \mathcal{P} \) for some parametric or non-parametric model \( \mathcal{P} \), and each null hypothesis \( H_i \subseteq \mathcal{P} \) represents a submodel; without loss of generality, the \( i \)th alternative hypothesis is \( \mathcal{P} \setminus H_i \). We assume the analyst computes a \( p \)-value \( p_i(X) \) to test each \( H_i \), where \( p_i \) is marginally super-uniform (i.e., stochastically larger than \( \text{Unif}(0,1) \)) under \( H_i \). Let \( \mathcal{H}_0(P) = \{ i : P \in H_i \} \) denote the set of true null hypotheses, and \( m_0 = |\mathcal{H}_0| \). Much of our discussion will treat the parametric setting \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d \} \), often parameterized so that \( H_i \) concerns only \( \theta_i \), for example \( H_i : \theta_i = 0 \) or \( H_i : \theta_i \leq 0 \).

A multiple testing procedure is a decision \( \mathcal{R}(X) \subseteq [m] = \{ 1, \ldots, m \} \) designating the set of rejected hypotheses. An analyst who rejects \( H_i \) for each \( i \in \mathcal{R}(X) \) makes \( V = |\mathcal{R} \cap \mathcal{H}_0| \) false rejections (sometimes called “false discoveries”) out of \( R = |\mathcal{R}| \) total rejections. Benjamini and Hochberg (1995) define the false discovery proportion (FDP) as

\[
\text{FDP}(\mathcal{R}(X); P) = \frac{V}{R \vee 1},
\]

where \( a \vee b = \max\{a, b\} \) and \( a \wedge b = \min\{a, b\} \). The false discovery rate (FDR) is defined as the expected FDP:

\[
\text{FDR}_{P}(\mathcal{R}) = \mathbb{E}_P\left[ \text{FDP}(\mathcal{R}(X); P) \right].
\]

A standard goal in multiple testing is to maximize the expected number of rejections while controlling the FDR at a pre-set significance level \( \alpha \), typically 5\%, 10\%, or 20\%.

The most widely used method for FDR control is the Benjamini–Hochberg (BH) procedure, an example of the more general class of step-up procedures. Let \( p(1) \leq \cdots \leq p(m) \) denote the order statistics of the \( m \) \( p \)-values. Then the step-up procedure for an increasing sequence of thresholds \( 0 \leq \Delta(1) \leq \cdots \leq \Delta(m) \leq 1 \) finds the largest index \( r \) for which \( p(r) \leq \Delta(r) \) and rejects all of the corresponding hypotheses up to that index. That is, we reject the hypotheses with the smallest \( R(X) \) \( p \)-values, where

\[
R(X) = \max \big\{ r : p(r)(X) \leq \Delta(r) \big\}.
\]

The BH(\( \alpha \)) procedure takes \( \Delta_\alpha(r) = \alpha r/m \). For a general family of thresholds \( \Delta_\alpha(r) \) that are non-decreasing in \( \alpha \) and \( r \), we denote the generic step-up procedure as \( \text{SU}_\Delta(\alpha) \). We denote the corresponding testing procedures as \( \mathcal{R}^{\text{BH}(\alpha)} \) and \( \mathcal{R}^{\text{SU}_\Delta(\alpha)} \) respectively.

As \( \alpha \) increases, the BH(\( \alpha \)) procedure becomes more liberal, with nested rejection sets. We define the BH-adjusted \( p \)-value (Yekutieli and Benjamini, 1999) as the level at which \( H_i \) is barely rejected, (as calculated by the \( \text{p.adjust} \) function in R):

\[
q_{i}^{\text{BH}}(X) = \min \left\{ \alpha : i \in \mathcal{R}^{\text{BH}(\alpha)} \right\}.
\]
$q_i^{BH}(X)$ can be regarded as a variant of the $q$-value introduced by Storey (2003), and we refer to it as the BH $q$-value for short. The same definition may be extended to any other multiple testing procedure where the minimum is well-defined, by replacing $R^{BH(\alpha)}$ with that method’s rejection set. For step-up procedures with $\Delta_\alpha(r)$ right-continuous in $\alpha$, the rejection sets are right-continuous too so the minimum is always well-defined.

Benjamini and Hochberg (1995) showed that the BH($\alpha$) procedure controls FDR at exactly $\alpha m_0/m$ if the $p$-values are independent, but results for dependent $p$-values are more complex.

1.2. FDR under dependence. We can begin to understand the role of dependence by first making a standard decomposition of the FDR according to the contribution of each true null hypotheses:

\[ \text{FDR} = \mathbb{E}\left[ \frac{V}{R \vee 1}\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[ \frac{V_i}{R \vee 1}\right], \]

where $V_i = 1\{H_i \text{ rejected}\}$. Under independence, BH controls each term in the sum at $\alpha/m$, attaining FDR control at level $\alpha m_0/m$.

Positive dependence between $V_i$ and $R$ tends to reduce each term in (3), making methods like BH conservative (e.g. Benjamini and Yekutieli, 2001; Sarkar et al., 2002). In particular, Benjamini and Yekutieli (2001) show the BH procedure is known to be conservative under positive regression dependence on a subset (PRDS), a relaxation of the positive regression dependence (PRD) condition introduced in Lehmann (1966). For $a, b \in \mathbb{R}^d$, we say $a \preceq b$ if $a_i \leq b_i$ for all $i = 1, \ldots, d$, and a set $A \subseteq \mathbb{R}^d$ is increasing if $a \in A$ and $a \preceq b$ implies $b \in A$. We say that $p_{-i} = (p_1, p_2, \ldots, p_{i-1}, p_{i+1}, \ldots, p_m)$ is PRD on $p_i$ if $\mathbb{P}(p_{-i} \in A \mid p_i)$ is increasing in $p_i$ for any increasing set $A$. Benjamini and Yekutieli (2001) show that the BH($\alpha$) procedure controls FDR conservatively at $\alpha m_0/m$, provided that $p_{-i}$ is PRD on $p_i$, for every $i \in \mathcal{H}_0$; this condition is called PRDS. Subsequently, many procedures designed to control FDR under independence have also been shown to control FDR under positive dependence. Notable exceptions include the Storey-BH method (Storey, Taylor and Siegmund, 2004), whose estimate of $m_0/m$ can fail badly under dependence\(^1\), and adaptive weighting methods such as AdaPT (Lei and Fithian, 2018) and SABHA (Li and Barber, 2019), whose finite-sample FDR control may be threatened by local random effects that make a cluster of $p$-values smaller together.\(^2\)

Unfortunately, the PRDS condition is quite restrictive. It does hold for one-sided testing with multivariate Gaussian test statistics whose pairwise correlations are all non-negative, or for one- or two-sided testing of uncorrelated multivariate $t$-test statistics. But $p$-values for one-sided testing with any negative pairwise correlations, or for two-sided testing with any correlations at all, no longer satisfy PRDS.

For general, unspecified dependence, Benjamini and Yekutieli (2001) also showed that the much more conservative BH($\alpha/L_m$) procedure controls FDR at level $\alpha$ under arbitrary dependence, where

\[ L_m = \sum_{i=1}^m \frac{1}{i} = \log m + O(1). \]

This method has become known as the Benjamini–Yekutieli (BY) procedure, or sometimes the log-corrected BH procedure. The proof technique was subsequently generalized in the shape

\(^1\)Benjamini, Krieger and Yekutieli (2006) propose another adaptive method that behaves better under positive dependence

\(^2\)Li and Barber (2019) derive an upper bound for the FDR inflation in the multivariate Gaussian case, but the bound depends on unknown aspects of the data distribution.
function approach of Blanchard and Roquain (2008) who show that if \( \nu \) is any probability measure on \( \{1, \ldots, m\} \), then the step-up procedure with

\[
\Delta_\alpha(r) = \frac{\alpha \beta(r)}{m}, \quad \text{where} \quad \beta(r) = \sum_{i=1}^{r} i\nu(\{i\})
\]

also controls FDR under arbitrary dependence between the \( p \)-values. Taking \( \nu(\{i\}) = (iL_m)^{-1} \) recovers the BY\((\alpha)\) procedure, but Blanchard and Roquain (2008) suggest other choices that sometimes improve on the BY procedure’s power.

These methods control FDR under worst-case dependence assumptions, but their generality typically comes at a price of substantial conservatism and diminished power compared to the BH procedure. As a result the BH procedure is often still used in applications where PRDS does not hold. This “off-label” use of BH owes in part to theoretical and empirical evidence that, under forms of dependence that typically arise in practice, such as the multivariate Gaussian distribution, BH is more often conservative than it is anti-conservative; see Farcomeni (2006) for extensive simulations. Roquain and Villers (2011) find that for one-sided multivariate Gaussian testing in the \( m = 2 \) case, the FDR can exceed the nominal \( \alpha \) when the test statistics are negatively correlated, but it does not approach the BY bound of \( \alpha L_m = 3\alpha/2 \). While Guo and Rao (2008) show that the BY bound can be attained for general \( m \), they acknowledge their construction is very artificial.

A second strategy is to prove asymptotic control in regimes where the limiting problem is simpler. For example, Genovese and Wasserman (2004) and Storey, Taylor and Siegmund (2004) study regimes where the empirical distributions of null and non-null \( p \)-values converge to limiting deterministic functions as \( m \to \infty \). This line of analysis was developed further in Ferreira and Zwinderman (2006) and Farcomeni (2007) and extended in Fan, Han and Gu (2012) and Fan and Han (2017) beyond the weak-dependence regime, by estimating and subtracting off principal component structure in the correlations. While these analyses provide valuable insights, the results hold only in the limit where \( R \to \infty \); but FDR control is often desired in problems where \( R \) may be relatively small, even if \( m \) is large. Troendle (2000) and Romano, Shaikh and Wolf (2008) study resampling-based approaches in a different asymptotic regime where \( m \) is fixed but the non-null \( p \)-values converge in probability to zero; in finite samples this is likely an optimistic assumption. Recent work by Xie et al. (2011) and Heller and Rosset (2021) obtain precise results finding optimal methods for FDR control with dependent test statistics, assuming foreknowledge of empirical Bayes models where the parameter values are known under the alternative hypothesis.

Recently discovered knockoff methods (Barber and Candès, 2015) offer an alternative means of FDR control under dependence for testing coefficients in linear regression models, and have been extended to testing conditional independence in supervised learning settings where a model for the joint distribution of predictor variables is available (Candès et al., 2018). Knockoff methods, which operate by feeding synthetic noise variables to a supervised learning procedure, represent a sharp methodological departure from classical multiple testing procedures like BH. Knockoffs can be more or less powerful than the BH procedure for context-dependent reasons that are not yet fully understood. We discuss their relative strengths and weaknesses compared to classical procedures like BH in Section 7.1.

A related alternative type I error criterion is to control the probability that FDP > \( \gamma \), often called the FDX or \( \gamma \)-FDP, at some prespecified level \( \alpha \) (e.g. Perone Pacifico et al., 2004; Lehmann and Romano, 2005a; Genovese and Wasserman, 2006). One motivation for targeting a quantile of the FDP rather than its expectation is that the FDP may have high variance, especially when the \( p \)-values are strongly dependent on one another (Korn et al., 2004; Owen, 2005). Guo et al. (2014) and Delattre et al. (2015) provide methods for controlling the \( \gamma \)-FDP under dependence, and we discuss the FDP distribution in our simulation studies.
In this work, we propose a new methodological framework for controlling FDR under dependence in a wide variety of discrete and continuous, parametric and nonparametric models. Rather than assume worst-case dependence, we begin with a baseline procedure like BH or BY and calibrate its FDR by exploiting full or partial knowledge of the dependence.

2. FDR control by conditional calibration.

2.1. Conditional calibration: a new strategy. Our method operates by adaptively calibrating a separate rejection threshold for each of the \( m \) \( p \)-values to control each term in (3), which we will call the FDR contribution of \( H_i \). Let \( \tau_i(c; X) \) be some possibly data-dependent rejection threshold for \( p_i \), with calibration parameter \( c \geq 0 \). We assume \( \tau_i \) is non-decreasing in \( c \) for all \( X \), and \( \tau_i(0; X) = 0 \) almost surely. We will be primarily interested in the effective BH threshold, defined as the \( p \)-value rejection threshold “estimated” by the BH(\( c \)) procedure:

\[
\tau_{BH}(c; X) = \frac{c R_{BH}(c)}{m}.
\]

Because the BH \( q \)-value \( q_i(X) \) is below \( \alpha \) if and only if \( p_i(X) \leq \tau_{BH}(\alpha; X) \), we can roughly interpret \( \tau_{BH} \) as an inverse \( q \)-value transformation, and \( c \) as a \( q \)-value cutoff. More generally, we define the effective SU\(_\Delta \) threshold as \( \tau_{SU\Delta}(c; X) = \Delta_c(R_{SU\Delta}(c)) \). We will suppress the dependence of \( p_i, q_i, \tau_i, \tau_{BH} \), and \( \tau_{SU\Delta} \) on \( X \) for convenience when no confusion can arise.

Taking the decomposition in (3) as our starting point, we will aim to calibrate the threshold for \( p_i \), choosing \( \hat{c}_i \) to directly control the \( i \)th term in the sum:

\[
E_{H_i} \left[ \frac{V_i}{R\vee 1} \right] = \sup_{p \in H_i} E_P \left[ \frac{1\{p_i \leq \tau_i(\hat{c}_i)\}}{R\vee 1} \right] \leq \frac{\alpha}{m}.
\]

We will use \( E_{H_i}[\cdot] \) as a shorthand notation for \( \sup_{p \in H_i} E_P[\cdot] \) throughout. More generally, we could control the \( i \)th term at \( \kappa_i \) provided \( \sum_{i \in H_0} \kappa_i \leq \alpha \); see Remark 2.1.

There are two main challenges in solving for \( \hat{c}_i \) in (6). First, the expectation depends in a possibly complicated way on the entire distribution of \( X \), whereas \( H_i \) typically only constrains the distribution of \( p_i \). Our first idea is to achieve (6) by controlling a more tractable conditional expectation, given some conditioning statistic \( S_i \) that blocks most or all of the nuisance parameters from influencing the conditional analysis. Often \( S_i \) is independent of \( p_i \), but we only require that \( p_i \) is conditionally superuniform given \( S_i \):

\[
\sup_{P \in H_i} \mathbb{P}_P(p_i \leq \alpha \mid S_i) \leq \alpha, \quad \text{for all } \alpha \in [0, 1].
\]

This style of conditioning is a well-established device for handling nuisance parameters in inference problems, especially for exponential family models and permutation tests, and has seen recent application in approaching complex decision problems like multiple testing (e.g. Weinstein, Fithian and Benjamini, 2013; Barber and Candès, 2015; Candès et al., 2018) and post-selection inference (e.g. Tibshirani et al., 2016; Lee et al., 2016; Fithian, Sun and Taylor, 2014).

Under independence, (7) is satisfied with \( S_i = p_{-i} \). A standard FDR control proof for the BH procedure, introduced in Benjamini and Yekutieli (2001), conditions on \( p_{-i} \) and applies the following key lemma:

**Lemma 1.** Let \( p^{(i \leftarrow 0)} = (p_1, \ldots, p_{i-1}, 0, p_{i+1}, \ldots, p_m) \). If \( R \) is a step-up procedure with threshold sequence \( \Delta(1), \ldots, \Delta(m) \), then the following are equivalent:

1. \( p_i \leq \Delta(R(p^{(i \leftarrow 0)})) \),
2. \( i \in \mathcal{R}(p) \), and
3. $\mathcal{R}(p) = \mathcal{R}(p^{(i-0)})$.

Lemma 1 establishes in particular that BH($\alpha$) is equivalent to thresholding each $p_i$ at the data-adaptive threshold $\alpha R(p^{(i-0)})/m$, which depends only on the other $p$-values. This useful technical result appears throughout the literature: beyond its use in Benjamini and Yekutieli (2001), it is involved in the proof of Theorem 2.1 in Ferreira and Zwinderman (2006) and Lemma 8.1 of Roquain et al. (2009), and appears in a more general form as Lemma 7.1 in Roquain and Villers (2011). We provide a proof for completeness in Appendix A.

Let $R^0 = R(p^{(i-0)})$, which depends only on $p_{-i}$. Then for the standard BH procedure under independence, applying Lemma 1 gives
\[
\mathbb{E}_{H_i} \left[ \frac{V_i}{R \lor 1} \mid p_{-i} \right] = \mathbb{E}_{H_i} \left[ \frac{\{p_i \leq \alpha R^0/m\}}{R^0} \right] = \frac{1}{R^0} \frac{\alpha R^0}{m} = \frac{\alpha}{m},
\]
with equality if $p_i$ is exactly uniform under $H_i$. Marginalizing over $p_{-i}$ and summing over $i \in H_0$ yields FDR $\leq \alpha m_0/m$.

For dependent $p$-values, conditioning on $p_{-i}$ will usually not satisfy (7). As a simple example, suppose $Z \sim N_m(\mu, \Sigma)$ where $\Sigma$ is a known covariance matrix with diagonal entries $\Sigma_{ii} = 1$, and we wish to test $H_i : \mu_i = 0$ against a one- or two-sided alternative. If $\Sigma_{ij} \neq 0$ then $p_i$ and $p_j$ are not independent and the distribution of $p_i$ given $p_{-i}$ depends on $\mu_j$. However, the conditioning statistic $S_i = Z_{-i} - \Sigma_{-i,i}Z_i$ is independent of $Z_i$, and after conditioning on $S_i$ the data distribution depends only on $\mu_i$, which is fixed at zero under the null.

In this example the data set can be reconstructed from $S_i$ and $Z_i$, and if the $p$-values are one-sided then it can also be reconstructed from $S_i$ and $p_i$. Figure 1 illustrates the conditional FDR contribution of $H_i$ for one-sided testing under three conditions: independence ($\Sigma = I_m$), where the contribution is exactly $\alpha/m$; positive-dependence ($\Sigma_{ij} \geq 0$ for all $i, j$), where the contribution is below $\alpha/m$; and worst-case dependence, where the contribution can be as high as $\alpha L_m/m$. In Figure 1b, the red line $R(p_i; S_i)^{-1}$ is increasing in $p_i$ because, fixing $S_i$, the other $p$-values are increasing functions of $p_i$.

The second main challenge is that the number of rejections $R$ in the denominator depends on all of $\hat{c}_1, \ldots, \hat{c}_m$, so all $m$ calibration problems are coupled to one another. To deal with this, we substitute an “estimator” $\tilde{R}_i(X)$ for the event value of $|\mathcal{R}(X) \cup \{i\}|$, the number of rejections if we also reject $H_i$. As we will see, our goal is for $\tilde{R}_i$ to be an accurate and easily computable lower-bound for $R$ on the event where $i$ is rejected, so that $V_i/\tilde{R}_i$ is a good upper bound for $V_i/(R \lor 1)$.

2.2. Our method. We now present a generic two-step FDR-controlling method, possibly with a third randomization step to handle cases where $\tilde{R}_i$ fails to lower-bound the size of the rejection set.

**Step 1: Calibration.** First, we use $\tilde{R}_i$ to estimate the conditional FDR contribution $\mathbb{E}_{H_i} [V_i/(R \lor 1) \mid S_i]$ as a function of the calibration parameter, and take $\hat{c}_i(S_i)$ as large as possible while controlling the conditional FDR contribution at $\alpha/m$. We say a calibration parameter $\gamma \geq 0$ is valid for the $i$th calibration problem if
\[
g_i^*(\gamma; S_i) = \sup_{p \in H_i} \mathbb{E}_p \left[ \frac{\{p_i \leq \tau_i(\gamma)\}}{R_i} \mid S_i \right] \leq \frac{\alpha}{m},
\]
suppressing the dependence of $\tilde{R}_i$ on $X$ for compactness of notation.

Let $c_i^*(S_i) = \sup_{\gamma \geq 0} \{g_i^*(\gamma; S_i) \leq \alpha/m\} \in [0, \infty]$, the supremum of all valid calibration parameters. Because $\tau_i(0) = 0$, $g_i^*(\gamma; S_i)$ is non-decreasing with $g_i^*(0; S_i) = 0$, any $\gamma < c_i^*$ is valid, but $c_i^*$ may not itself be valid if $g_i^*$ is discontinuous. To cover this case, we
(a) Independent case: $R$ is constant on the event where $p_i$ is rejected, \( p_i \leq \alpha R/m \), as a result of Lemma 1.

(b) Positive-dependent case: $R$ is decreasing in $p_i$, since the other $p$-values are increasing in $p_i$.

(c) Worst-case: the other $p$-values are adversarially configured so that $R$ is always just large enough to reject $p_i$.

allow for $\hat{c}_i$ to be a sequence $\hat{c}_{i,t}$ converging to $c^*_i$ from below: if $c^*_i$ is valid we can simply take $\hat{c}_{i,t} = c^*_i$ for all $t$. We say $\hat{c}_i$ is maximal if $\bigcup_t \{ \hat{c}_{i,t} \}$ includes every valid $c$, almost surely.

Step 2: Initial rejection. Next, we initialize the rejection set, via:

$$\mathcal{R}_+ = \{ i : p_i \leq \tau_i(\hat{c}_i) \}.$$

If $\hat{c}_i$ is a sequence, the condition $p_i \leq \tau_i(\hat{c}_i)$ is understood to mean $p_i \leq \tau_i(\lim_t \hat{c}_{i,t})$. To limit notational bloat in our prose, we will discuss $\hat{c}_i$ as though it is a single value, but our results all apply to the general case.

In practice, we can usually carry out this step using the maximal $\hat{c}_i$ without explicitly calculating $\hat{c}_i$. Instead we can use the $q$-value $q_i(X) = \min \{ c : p_i \leq \tau_i(c) \}$, which is well-defined and easy to calculate for $\tau^\text{BH}$ or $\tau^\text{SU}$-A. Then, for the maximal $\hat{c}_i$, $i \in \mathcal{R}_+$ if and only if the observed $q_i$ is a valid calibration parameter, so we can alternatively write

$$\mathcal{R}_+ = \{ i : g^*_i(q_i; S_i) \leq \alpha/m \}.$$

Let $R_+ = |\mathcal{R}_+|$. If $R_+ \geq \hat{R}_i$ for all $i \in \mathcal{R}_+$, then we can halt the procedure with $\mathcal{R} = \mathcal{R}_+$. Otherwise, we may need to prune the rejection set further.

Step 3 (if necessary): Randomized pruning. If there is some $i \in \mathcal{R}_+$ for which $\hat{R}_i > R_+$, then we must prune the rejection set via a secondary BH procedure. For user-generated uniform random variables $u_1, \ldots, u_m \sim \text{Unif}(0, 1)$, let

$$R(X; u) = \max \left\{ r : \left\{ i \in \mathcal{R}_+ : u_i \leq r/\hat{R}_i \right\} \geq r \right\},$$

and reject $H_i$ for the $R$ indices with $i \in \mathcal{R}_+$ and $u_i \leq R/\hat{R}_i$. The procedure in (9) is equivalent to the BH(1) procedure on “$p$-values” $\tilde{p}_i = u_i \hat{R}_i/R_+$ for $i \in \mathcal{R}_+$. We can skip this step if $R_+ \geq \hat{R}_i$ for all $i \in \mathcal{R}_+$; in that case all $u_i \hat{R}_i/R_+ \leq 1$, so $R = R_+$. 

Fig 1: Visualizing the conditional FDR contribution for BH($\alpha$) under two simplifying assumptions: (i) $p_i$ is uniform under the null, and (ii) the rejection set is a function of $p_i$ and $S_i$, the conditioning statistic. $R^{-1}$, the reciprocal of the number of rejections, is shown as the broken red line. The BH($\alpha$) procedure rejects $H_i$ when $(p_i, R^{-1})$ is below the hyperbola $p_i = \alpha R/m$ and the conditional FDR contribution is given by $\mathbb{E} \left[ \frac{V_i}{R_{\alpha/m}} \mid S_i \right] = \int_{u \leq \alpha R(u)/m} R^{-1}(u; S_i) \, du$, the area of the light red region.
This pruning step is best avoided, since it puts every rejection at risk, even those with minuscule p-values. If \( \hat{R}_i \) frequently overestimates \( R_+ \) by a large margin, the method may suffer very low power even in the presence of strong signals. Even when the power loss from pruned rejections is small, it is scientifically preferable to use a procedure that rarely or never randomizes. We call a calibrated procedure safe if pruning is never necessary. Theorem 3 discusses conditions for our methods to be safe, and we find in our simulations that even when unsafe our dBH procedure very rarely requires pruning.

We show next that this procedure controls FDR at the desired level.

**Theorem 2 (FDR control).** Assume that (7) holds, and \( \hat{c}_{i,t} \) is chosen to guarantee (8), for all \( i \) and \( t \). Then the three-step procedure defined above controls the FDR at or below level \( \frac{\alpha m_0}{m} \).

**Proof.** It is sufficient to show that \( \mathbb{E}\left[ \frac{V_i}{R \vee 1} \right] \leq \frac{\alpha}{m} \) for every \( i \in \mathcal{H}_0 \). For \( i \in \mathcal{R}_+ \) let \( R_i^* = R(X; u(i^{(i-0)}) \) which is independent of \( u_i \). By Lemma 1 applied to the secondary BH procedure, \( R_i^* = R \) on the event \( \{ V_i = 1 \} \). As a result, we can write

\[
\mathbb{E}\left[ \frac{V_i}{R \vee 1} \right] = \mathbb{E}\left[ \frac{1 \{ i \in \mathcal{R}_+ \} \cdot 1 \left\{ u_i \leq \frac{R}{\hat{R}_i} \right\}}{R \vee 1} \right]
\]

\[
= \mathbb{E}\left[ \frac{1 \{ i \in \mathcal{R}_+ \} \cdot 1 \left\{ u_i \leq \frac{R_i^*}{\hat{R}_i} \right\}}{R_i^*} \right]
\]

\[
= \mathbb{E}\left[ \mathbb{E}\left[ 1 \{ i \in \mathcal{R}_+ \} \cdot 1 \left\{ u_i \leq \frac{R_i^*}{\hat{R}_i} \right\} \left| X, u_{-i} \right. \right] \right]
\]

\[
\leq \mathbb{E}\left[ 1 \{ i \in \mathcal{R}_+ \} \right] \frac{R_i^*}{\hat{R}_i}
\]

\[
= \lim_{t \to \infty} \mathbb{E}\left[ \frac{1 \{ p_i \leq \tau_i(\hat{c}_{i,t}) \}}{\hat{R}_i} \right]
\]

\[
= \lim_{t \to \infty} \mathbb{E}\left[ \frac{1 \{ p_i \leq \tau_i(\hat{c}_{i,t}) \}}{\hat{R}_i} \left| S_i \right. \right].
\]

We can move the limit outside the integral in (14) by monotone convergence, because the integrand is bounded by 1. Under \( H_i \), the last expression is no larger than \( \frac{\alpha}{m} \), completing the proof. \( \square \)

We pause to make several further observations:

**Remark 2.1.** If we modify the calibration step replacing \( \frac{\alpha}{m} \) by a generic, possibly data-dependent upper bound \( \kappa_i(S_i) \) in inequality (8), then the proof of Theorem 2 trivially generalizes to guarantee FDR control at level \( \sum_{i \in \mathcal{H}_0} \mathbb{E}_P \kappa_i(S_i) \). Section 7 discusses two extensions of the procedure where this additional flexibility is useful: adaptive estimation of \( \frac{m_0}{m} \) and adaptive hypothesis weighting.

**Remark 2.2.** From the proof of Theorem 2 we see that there are three possible sources of conservatism in the above procedure. First, we typically have \( m_0 < m \). Second, the inequality in (13) is the price we pay in conservatism when \( \hat{R}_i \) underestimates \( R \). Third, the
conditional expectation in (15) may not attain \( \alpha/m \), either because \( \hat{c}_i \) is not maximal, because \( p_i \) has a discrete distribution, or because \( P \) does not attain the supremum in (8) (for example if \( P \) lies in the interior of \( H_i \)).

To operationalize our method, we must fill in the details of what threshold family \( \tau_i \) and estimator \( \hat{R}_i \) we use, how we identify the conditioning statistic \( S_i \), and how we calibrate the threshold in practice. The next sections address these issues in turn.

2.3. The dependence-adjusted BH and BY procedure. While Theorem 2 proves FDR control for a broad class of procedures, our empirical results focus on special cases of our method that are designed to couple tightly with the BH and BY procedures. If we use the three-step method of the previous section with the effective BH threshold \( \tau_i = \tau^{BH} \) and estimator \( \hat{R}_i = [R^{BH}(\gamma\alpha) \cup \{i\}] \), we call the resulting method the dependence-adjusted BH procedure, which we denote \( \text{dBH}_\gamma(\alpha) \). In the special case where we take \( \gamma = 1/L_m \), we call the resulting method the dependence-adjusted BY procedure, denoted \( \text{dBY}(\alpha) \). More generally, let the \( \text{dSU}_\gamma(\alpha) \) procedure use threshold \( \tau_i = \tau^{SU_\gamma} \) and estimator \( \hat{R}_i = [R^{BH}(\gamma\alpha) \cup \{i\}] \). By definition, we assume the methods use maximal \( \hat{c}_i \) by default; this is possible as long as we can check the inequality (8), since the \( q \)-value is always well-defined for \( \tau^{BH} \), and for \( \tau^{SU_\gamma} \) of the form defined in (4).

We can interpret the \( \text{dBH}_\gamma(\alpha) \) calibration parameters \( \hat{c}_1, \ldots, \hat{c}_m \) as calibrated \( q \)-value rejection thresholds for the respective hypotheses, since \( q_i \leq c \) if and only if \( p_i \leq \tau^{BH}(c) \). As a result, we have

\[
R^{BH}(\min, \hat{c}_i) \subseteq R_+ \subseteq R^{BH}(\max, \hat{c}_i).
\]

Provided that \( \min_i \hat{c}_i \geq \gamma \alpha \), the randomization step is avoided and we can replace \( R_+ \) with \( \mathcal{R} \) in (16). In our simulations we take \( \gamma = 0.9 \) as a conservative choice except in the positive-dependent case, where we take \( \gamma = 1 \). Figure 2 illustrates how conditional calibration operates for the \( \text{dBH}_1 \) and \( \text{dBY} \) procedures.

Under a slight strengthening of the PRDS condition, the \( \text{dBH}_1(\alpha) \) procedure is uniformly more powerful than the BH procedure, meaning \( \mathcal{R}^{dBH_1} \supseteq \mathcal{R}^{BH} \) almost surely (this does not rule out the possibility that the two rejection sets are almost surely identical, as is the case under independence). For a given conditioning statistic \( S_i \) we say the \( p \)-values \( p_{-i} \) are conditionally positive regression dependent (CPRD) if \( \mathbb{P}(p_{-i} \in A \mid p_i, S_i) \) is almost surely increasing in \( p_i \) for any increasing set \( A \). If \( p_{-i} \) is CPRD on \( p_i \) for all \( i \in \mathcal{H}_0 \) we say the \( p \)-values are CPRD on a subset (CPRDS). Note CPRDS implies PRDS after marginalizing over \( S_i \), but PRDS does not necessarily imply CPRDS. As we will see in Section 3, the CPRD condition can be a useful sufficient condition for proving PRD in models like the multivariate Gaussian and multivariate \( t \), where the dependence between \( p \)-values is very simple after conditioning on \( S_i \).

**Theorem 3.** Assume \( \hat{c}_1, \ldots, \hat{c}_m \) are maximal. Then

1. If the \( p \)-values are independent with \( p_i \) uniform under \( H_i \), then the \( \text{dBH}_1(\alpha) \) procedure with \( S_i = p_{-i} \) is identical to the BH(\( \alpha \)) procedure.
2. If the \( p \)-values are CPRDS for all \( P \in \mathcal{P} \), then the \( \text{dBH}_1(\alpha) \) procedure is safe, and uniformly more powerful than the BH(\( \alpha \)) procedure.
3. For arbitrary dependence, the \( \text{dBY}(\alpha) \) procedure is safe, and uniformly more powerful than the BY(\( \alpha \)) procedure.
4. Assume the thresholds \( \Delta_\alpha \) are of the form (4). Then for arbitrary dependence, the \( \text{dSU}_\Delta(\alpha) \) procedure is safe, and uniformly more powerful than the SU_\Delta(\( \alpha \)) procedure.
REMARK 2.3. The proof of Theorem 3 only involves values of \( p_i \) in the interval \( p_i \in [0, \alpha] \), and the only increasing sets that appear in the proof are of the form \( \{ R_i \leq r \} \). For the purpose of applying Theorem 3, then, we could relax the definition of CPRD to require only that \( \mathbb{P}(R_i \leq r \mid p_i, S_i) \) is increasing for \( p_i \in [0, \alpha] \). As a result, the second conclusion of Theorem 3 applies to one-sided testing with uncorrelated multivariate \( t \)-statistics even though the \( p \)-values are neither PRDS nor CPRDS, as we show in Section 3.2.

REMARK 2.4. In many of our examples, including one-sided multivariate Gaussian testing, the entire data set can be reconstructed from \( p_i \) and \( S_i \). In that case, the conditional probability is always 0 or 1, and a sufficient condition for CPRD is that, fixing \( S_i \), every other \( p_j \) is an increasing function of \( p_i \).

REMARK 2.5. If the model \( \mathcal{P} \) includes distributions that realize the worst-case dependence as shown in Figure 1c, for the BH(\( \alpha/L_m \)) procedure, then we will have \( g^*_{i}(\alpha/L_m) = \alpha/m \) and the calibration in the dBY(\( \alpha \)) procedure cannot improve on \( \hat{c}_i = \alpha/L_m \). In that case, the dBY(\( \alpha \)) and BY(\( \alpha \)) procedures are equivalent, giving almost surely identical rejection sets. In particular, if \( \mathcal{P} \) is completely generic — if we refuse to assume anything at all about the dependence between the \( p \)-values — then we cannot in general improve on BY, whose FDR control is sharp in such a general model. The dBH method relies on our making some assumption constraining the dependence structure.

2.4. Identifying the conditioning statistic \( S_i \). There are several desiderata for a good conditioning statistic. Most importantly, recall that our method’s validity depends on \( p_i \) being a valid conditional \( p \)-value, so that (7) holds. To facilitate calibration, \( S_i \) should also eliminate or mitigate the influence of nuisance parameters on the conditional distribution of \( X \). Finally, the conditional distribution under \( H_i \) should be analytically and/or computationally tractable. Calibration is conceptually simplified if \( S_i \) is a sufficient statistic for the null submodel \( H_i \), so that the conditional distribution of \( X \) is known under \( H_i \). In that case, we say \( H_i \) is conditionally simple, and \( g^*_{i}(c; S_i) \) is an integral we can directly evaluate. Otherwise, we say \( H_i \) is conditionally composite, and \( P^*_{i} \) is least favorable for calibrating \( \hat{c}_i \) if it almost surely attains the supremum:

\[
(17) \quad g^*_{i}(c; S_i) = \mathbb{E}_P \left[ \frac{1\{p_i \leq \tau_i(c)\}}{\hat{R}_i} \mid S_i \right], \quad \text{for all } c \geq 0.
\]

One class of examples where the choice of conditioning statistic is fairly natural is the case of parametric exponential family models. Appendix B also discusses nonparametric settings with permutation tests.

EXAMPLE 2.1 (Exponential families). In exponential family models, there is a natural choice of \( S_i \) which follows from the classical theory of conditional testing in the style of Lehmann and Scheffé (1955). Suppose our model arises from a full-rank exponential family in canonical form:

\[
(18) \quad X \sim f_\theta(x) = e^{\theta^T(x) - A(\theta)} f_0(x), \quad \theta \in \Theta \subset \mathbb{R}^d,
\]

and for \( i = 1, \ldots, m \leq d \), \( H_i \) takes the form \( H_i : \theta_i = 0 \) or \( H_i : \theta_i \leq 0 \). In this setting, the uniformly most powerful unbiased (UMPU) test rejects \( H_i \) when \( T_i(X) \) is extreme, conditional
α/m

\[ p_i = \frac{\alpha \hat{R}_i}{m} \]

\[ \{ q_i \leq \hat{c}_i \} \]

\[ \alpha/m \]

\[ \{ q_i \leq \hat{c}_i \} \]

\[ \alpha/m \]

\[ \hat{R}_i(p; S)_i^{-1} \]

\[ \hat{R}_i^{-1} \]

\[ \hat{R}_i(p; S_i)^{-1} \]

\[ \hat{R}_i^{-1} \]

\[ \hat{R}_i^{-1} \]

\[ \hat{R}_i(p; S_i)^{-1} \]

(a) The dBH \((\alpha)\) procedure in the positive-dependent case.

(b) The dBY \((\alpha)\) procedure, which is often much more powerful than BY \((\alpha)\).

Fig 2: Illustration of conditional calibration for the dBH \(1\) and dBY procedures. The red line in each plot represents the reciprocal of \(\hat{R}_i = |R^{BH}(\gamma\alpha) \cup \{i\}|\) for \(\gamma = 1\) and \(\gamma = 1/Lm\), respectively. We assume here as in Figure 1 that the full data set \(X\) is a deterministic function of \(p_i\) and \(S_i\). If \(p_i\) is uniform under the null, the conditional FDR contribution is estimated as \(\int_{u: q_i(u; S_i) \leq \hat{c}_i} \hat{R}_i^{-1}(u; S_i) \; du\). The calibration step adds the hatched region to increase the total area, which is the estimated FDR contribution, to \(\alpha/m\). Note that the set \(\{ u : q_i(u; S_i) \leq \hat{c}_i \}\), shown here as a contiguous interval, could have multiple connected components in general.

on the value of \(S_i = T_{-i}(X)\); see e.g. Brown (1986) (Theorem 1.15). As a result, \(T_{-i}(X)\) makes a natural choice of test statistic, because it eliminates nuisance parameters and because \(p_i\) is conditionally valid by construction, so the hypothesis \(H_i\) is conditionally simple and we can evaluate \(g^*_i(c; S_i)\) directly. For one-sided testing, the null hypothesis \(H_i : \theta_i \leq 0\) is conditionally composite, but we will see in Section 2.6 that under mild conditions setting \(\theta_i = 0\) is least favorable.

Section 3 and Appendix B (in the supplementary material) discuss a variety of multiple testing problems in this vein, including multivariate Gaussian test statistics, testing in linear models, multiple comparisons with binary responses, and conditional independence testing in Gaussian graphical models.

2.5. Recursive refinement of \(\hat{R}_i\). The random variable \(\hat{R}_i(X)\) is an unusual estimator in that, by the time the procedure terminates, we will have computed the estimand \(|R(X) \cup \{i\}|\) (which coincides with \(R\) for \(i \in R(X)\)). If they differ, it seems natural to re-run the procedure substituting the “correct value” for the inaccurate estimate. While we can start again at Step 1 with the same threshold family, we will still not have a perfect estimator because changing \((\hat{R}_1, \ldots, \hat{R}_m)\) will affect the entire procedure and change \(R\) in turn. Nevertheless, we may obtain a better procedure if the new \(\hat{R}_i\) is a better estimator for the new rejection set. We call this process recursive refinement of the estimator.

We will denote the original estimator as \(\hat{R}_i^{(1)}\), which leads to original calibration parameter \(\hat{c}_i^{(1)}\) and initial rejection set \(R_+^{(1)}\). We define the recursively refined estimator as

\[
\hat{R}_i^{(2)}(X) = |R_+^{(1)}(X) \cup \{i\}|.
\]
We can then calibrate new thresholds $\hat{c}_i^{(2)}$ solving (8) with respect to $\hat{R}_i^{(2)}$, and proceed as before. In principle, we can repeat this refinement as many times as we want, defining $\hat{R}_i^{(k)}(X) = [R_i^{(k-1)}(X) \cup \{i\}]$ for all $k > 1$, but in most problems of moderate size the computational cost is prohibitive for $k > 2$, for reasons we explain in Section 4.

Recursive refinement is especially useful when we begin with a very conservative estimator, as we do when we use the dBY procedure. If we use the effective BH threshold with the dBH$_c(\alpha)$ estimator, we call the resulting procedure the dBH$_c^2(\alpha)$ procedure, or the dBY$^2(\alpha)$ procedure if $\gamma = 1/L_m$. When the baseline procedure is safe, recursive refinement always yields another safe procedure that is uniformly more powerful:

**Theorem 4.** Assume $c_i^{(1)}, \ldots, c_i^{(k)}$ are maximal for all $i$ and $k$. If $R_i^{(1)}$ is safe, then for every $k \geq 1$, $R_i^{(k+1)}$ is safe and uniformly more powerful than $R_i^{(k)}$.

### 2.6. One-sided testing in exponential family models.

In this section we consider how to test the one-sided hypotheses $H_i : \theta_i \leq 0$ in exponential family models of the form (18). We assume throughout that $0 \in \Theta \cup \mathbb{R}^d$, and that the tests are right-tailed; otherwise we can reparameterize the family (possibly with different reparameterizations for each $i$).

If we were testing a single hypothesis, the UMPU test for $H_i$ would reject for large values of $T_i$, conditional on the value of $S_i = T_i$; these include the z-, t-, and Fisher exact tests we discuss in Section 3 and Appendix B (Lehmann and Romano, 2005b). That is, the conditional test p-value is given by

$$p_i = \lim_{t \to T_i} 1 - F_{i,0}(t \mid S_i),$$

where the conditional distribution $F_{i,\theta_i}(t \mid S_i) = \mathbb{P}_{\theta_i}(T_i \leq t \mid S_i)$ depends only on $\theta_i$.

Because the distribution of $T_i$ is stochastically increasing in $\theta_i$, only the boundary case $\theta_i = 0$ is relevant for calculating the p-value, but we cannot necessarily restrict our attention to the boundary when we calibrate $\hat{c}_i$, unless $\theta_i = 0$ is least favorable in the sense of (17). The next result gives a sufficient condition for least favorability:

**Proposition 5.** Consider testing $H_i : \theta_i \leq 0$ for $i = 1, \ldots, m \leq d$ in an exponential family model of the form (18) with $0 \in \Theta^0 \subseteq \mathbb{R}^d$. Assume for all $i$ that $p_i$ is given by the standard one-sided UMPU test, and that we have for some $\alpha'$ which is not necessarily the target FDR level, almost surely,

(i) $\tau_i(c \mid X) \leq \alpha'$, and

(ii) Under $\theta_i = 0$, the conditional upper-$\alpha'$ quantile of $T_i$ is above its conditional mean:

$$F_{i,0}(t \mid S_i) < 1 - \alpha', \quad \text{for all } t < \mu_{i,0}(S_i),$$

where $\mu_{i,\theta_i}(S_i) = \mathbb{E}_{\theta_i}[T_i \mid S_i]$.

Then, $\theta_i = 0$ is least favorable for $H_i$, for purposes of calibrating $\hat{c}_i$.

**Remark 2.6.** Although there is no universal cap on the $\tau_{\text{BH}}$ threshold, in practice it very rarely exceeds $2\alpha$, and we can choose to modify it by capping it manually at some $\alpha'$. In our implementation of dBH, we cap $c$ at $2\alpha$, effectively capping $\tau_{\text{BH}}$ at $\alpha' = 2\alpha$, as discussed in Appendix C.2.2. In all examples discussed in Section 3, $T_i$ is symmetrically distributed given $S_i$ with $\mu_{i,0}(S_i) = 0$, so the assumptions of Proposition 5 hold for all $\alpha \leq 0.25$. 

2.7. Two-sided testing and directional error control. We say a hypothesis is two-sided if it can be written as $H_i : \theta_i(P) = 0$, for some parameter $\theta_i$ mapping $P$ to $\mathbb{R}$, where the range includes both positive and negative values. Because a two-sided hypothesis frequently represents a “measure-zero” set in the model $P$, rejecting $H_i$ is more meaningful when we can also draw an inference about the sign of $\theta_i$. If we write $H_i$ as the intersection of the one-sided hypotheses $H_{i,+} : \theta_i(P) \leq 0$ and $H_{i,-} : \theta_i(P) \geq 0$, a directional inference is one that rejects exactly one of $H_{i,+}$ and $H_{i,-}$ along with $H_i$.

For multiple testing of two-sided hypotheses with directional inferences, let $R^+$ denote the set of indices for which we declare $\theta_i > 0$ (reject $H_{i,+}$) and $R^-$ the set for which we declare $\theta_i < 0$ (reject $H_{i,-}$), with $R$ the disjoint union of both sets. Let $p_{i,+}$ and $p_{i,-}$ denote $p$-values for each of the two tests, which we assume are conditionally valid:

$$\sup_{P \in H_{i,+}^\leq} \mathbb{P}_P(p_{i,+} \leq \alpha \mid S_i) \leq \alpha \quad \text{and} \quad \sup_{P \in H_{i,-}^\geq} \mathbb{P}_P(p_{i,-} \leq \alpha \mid S_i) \leq \alpha, \quad \forall \alpha \in [0,1].$$

(20)

We assume the two one-sided tests are based on a common test statistic $T_i$, with $H_{i,+}$ rejected when $T_i$ is large and $H_{i,-}$ rejected when $T_i$ is small, where the critical thresholds possibly depend on $S_i$. For the sake of simplicity we also assume the two-sided test is equal-tailed, in the sense that $p_i = 2 \min \{p_{i,+}, p_{i,-}\}$.

In general, multiple testing procedures that are valid for two-sided hypotheses do not necessarily justify directional conclusions even if the constituent single hypothesis tests do (Shaffer, 1980; Finner, 1999). Testing two-sided hypotheses with directional inferences creates more opportunities to make errors: defining $\mathcal{H}_{i,+} = \{i : \theta_i \leq 0\}$ and likewise $\mathcal{H}_{i,-} = \{i : \theta_i \geq 0\}$, the number of directional errors is $V_{\text{dir}} = V^+ + V^-$, where $V^+ = \mathcal{H}^\leq \cap R^+$, and $V^- = \mathcal{H}^\geq \cap R^-$. The directional FDP is defined as $\text{FDP}_{\text{dir}} = V_{\text{dir}} / (R \lor 1)$, and its expectation $\text{FDR}_{\text{dir}}$ is the directional FDR. The next result gives a natural sufficient condition guaranteeing that our method with directional inferences controls the directional FDR:

**Lemma 6.** Assume $p_{i,+}$ and $p_{i,-}$ are valid in the sense of (20), and that the assumptions of Theorem 2 are satisfied for $S_i$, $\tau_i$, $\zeta_i$, and $\hat{R}_i$, for $i = 1, \ldots, m$. Define

$$g_{i,\theta}(c ; S_i) = \sup_{P ; \theta_i(P) = \theta} \mathbb{E}_P \left[ \frac{1 \{p_i^+ \leq \tau_i(c) / 2\}}{\hat{R}_i} \right] \mid S_i \right], \quad \text{for } a = +, -.$$

If $g_{i,\theta}$ is almost surely non-decreasing in $\theta$ for $\theta \leq 0$, and $g_{i,\theta}$ is almost surely non-increasing in $\theta$ for $\theta \geq 0$, then our three-step method controls the directional FDR.

As an immediate consequence of Lemma 6 and Proposition 5, we see that we can draw directional conclusions for two-sided multiple testing in exponential family models:

**Corollary 7.** Consider testing $H_i : \theta_i = 0$ for $i = 1, \ldots, m \leq d$ in an exponential family model of the form (18) with $0 \in \Theta^o \subset \mathbb{R}^d$. Assume for all $i$ that $p_i = 2 \min \{p_{i,+}, p_{i,-}\}$ where $p_{i,+}$ are given by the standard one-sided UMPU tests, and that we have for some $\alpha'$ which is not necessarily the target FDR level, almost surely,

(i) $\tau_i(c ; X) \leq \alpha'$, and

(ii) Under $\theta_i = 0$, the conditional mean is between the conditional lower- and upper-$\alpha'/2$ quantiles:

$$F_{i,0}(t \mid S_i) < 1 - \alpha' / 2, \quad \text{for all } t < \mu_{i,0}(S_i), \quad \text{and}$$

$$F_{i,0}(t \mid S_i) > \alpha' / 2, \quad \text{for all } t > \mu_{i,0}(S_i),$$

where $\mu_{i,\theta}(S_i) = \mathbb{E}_{\theta_i} [T_i \mid S_i]$. 

Then, our three-step method controls the directional FDR.

For symmetric null distributions, as in all examples discussed in Section 3, \( \alpha' \) can be set to be 1, rendering the first condition unnecessary even without the capping step required for the one-sided tests.

3. Examples. In this section we give additional details about several parametric examples arising from the multivariate Gaussian family. Appendix B (in the supplementary material) discusses three further parametric examples — edge testing in Gaussian graphical models, post-selection \( z \)- and \( t \)-testing, and multiple comparisons to control for a one-way layout with binary outcomes — as well as a nonparametric example, multiple comparisons to control with in a one-way layout with generic responses and \( p \)-values arising from permutation tests. As we will see, in cases where the conditional distribution given \( S_i \) is very simple to describe, the CPRD condition can often be checked easily, giving a potentially useful alternative route to checking PRD.

Let \( Z \sim N_d(\mu, \Sigma) \) with \( \Sigma \succ 0 \), an exponential family model with density

\[
f_{\mu, \Sigma}(z) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (z - \mu)' \Sigma^{-1} (z - \mu) \right\}.
\]

Lemma 8 establishes a useful decomposition of terms in (22).

**Lemma 8.** Assume \( \Sigma \in \mathbb{R}^{d \times d} \) is nonsingular and define

\[
A = (\Sigma^{-1})_{-i,-i}, \quad \zeta = A(\mu_{-i} - \Sigma_{-i,i}(\Sigma_{i,i})^{-1} \mu_i), \quad \text{and} \quad U(z) = z_{-i} - \Sigma_{-i,i}(\Sigma_{i,i})^{-1} z_i.
\]

Then we have

\[
\mu' \Sigma^{-1} z = \frac{1}{\Sigma_{ii}} \mu_i z_i + \zeta' U(z), \quad \text{and} \quad z' \Sigma^{-1} z = \frac{z_i^2}{\Sigma_{ii}} + U(z)'AU(z).
\]

Note that while \( A, \zeta, \) and \( U(z) \) depend on \( i \), we have suppressed that dependence to avoid cluttering our notation. Because the calibration for each hypothesis only requires thinking about reparameterizations and conditional distributions for one hypothesis at a time, we follow this convention in the examples below; the absence of an index \( i \) from the notation for some quantity should not be read as implying that quantity does not depend on \( i \).

3.1. Multivariate \( z \)-statistics. First assume that \( Z \sim N_d(\mu, \Sigma) \) with \( \Sigma \succ 0 \) known and all \( \Sigma_{i,i} = 1 \), and that we wish to test \( H_i : \mu_i = 0 \) or \( H_i : \mu_i \leq 0 \), for \( i = 1, \ldots, m \leq d \). Let \( p_i \) be the standard one- or two-sided \( p \)-value based on \( Z_i \), for \( i \leq m \).

Applying Lemma 8, we can rewrite (22) as a full-rank \( d \)-parameter exponential family:

\[
f_{\mu}(z) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ \mu_i z_i + \zeta' U(z) - \frac{1}{2} (z^2_i + U(z)'AU(z)) - \frac{1}{2} \mu' \Sigma^{-1} \mu \right\}.
\]

As a result we see that conditioning on \( S_i = U(Z) \) eliminates \( \zeta \) from the problem, leaving a one-parameter exponential family model in \( \mu_i \) with \( Z_i \) as its sufficient statistic. Moreover \( Z_i \) and \( S_i \) are independent by standard calculations.
To carry out the dBH$_\gamma(\alpha)$ procedure, we must evaluate whether

\[(24) \, \mathbb{E}_0 \left[ \frac{1 \{q_i \leq c\}}{|R^{BH(\gamma\alpha)} \cup \{i\}|} \mid S_i \right] \leq \frac{\alpha}{m}, \]

plugging in $c = q_i(Z)$, the observed BH $q$-value. Because $Z_{-i} = S_i + \Sigma_{-i,i} Z_i$, it is straightforward to evaluate the expectation in (24) by holding $S_i$ fixed at its observed value and integrating with respect to $Z_i \sim N(0, 1)$, which is least favorable for $H_i : \mu_i \leq 0$ for $\alpha \leq 0.25$, per Remark 2.6.

For one-sided $p$-values $p_i = 1 - \Phi(Z_i)$, $p_{-i}$ is CPRD on $p_i$ if and only if $\Sigma_{ij} \geq 0$ for all $j = 1, \ldots, m$, since in that case every other $p_j$ is conditionally a non-decreasing function of $p_i$. This matches the condition for marginal PRD in Benjamini and Yekutieli (2001), who showed that one-sided $p$-values are PRD in this problem provided all pairwise correlations are non-negative. For two-sided $p$-values, neither CPRD nor PRD holds if any of the $\Sigma_{ij}$ are nonzero.

3.2. Multivariate $t$-statistics. A slightly harder case is to modify the problem in Section 3.1 by assume that $\Sigma = \sigma^2 \Psi$ where $\Psi > 0$ is known but $\sigma^2 > 0$ is unknown. Without loss of generality we assume $\Psi_{i,i} = 1$. To estimate $\sigma^2$, assume we also observe an additional independent vector $W \sim N_{n-d}(0, \sigma^2 I_{n-d})$.

As above, we consider testing $H_i : \mu_i = 0$ or $H_i : \mu_i \leq 0$ for $i = 1, \ldots, m \leq d$. The usual $t$-statistic for $H_i$ is

\[T_i = \frac{Z_i}{\sigma} H_i \sim t_{n-d}, \quad \text{where} \quad (n-d)\sigma^2 = ||W||^2 \sim \sigma^2 \chi^2_{n-d}, \]

from which one- or two-sided $p$-values are computed as usual.

Extending the density in (23) to include $W$ and bringing back the subscript $i$ to $U_i(z)$ to avoid confusion, we obtain the $(d+1)$-parameter exponential family density

\[
\frac{1}{(2\pi \sigma^2)^{n/2} |\Psi|} \exp \left\{ \frac{1}{\sigma^2} \mu_i z_i + \Psi^{\prime} U_i(z) - \frac{1}{2\sigma^2} \left( w^\prime w + z_i^2 \right) - \frac{1}{2\sigma^2} U_i(z)^\prime B U_i(z) - \frac{\mu_i^\prime \Psi^{-1} \mu_i}{2\sigma^2} \right\},
\]

for $B = \sigma^2 A = (\Psi^{-1})_{-i,-i}$, which depends only on the known matrix $\Psi$. Likewise note that

\[(25) \quad U_i(z) = z_{-i} - \Sigma_{-i,i}(\Sigma_{i,i})^{-1} z_i = z_{-i} - \Psi_{-i,i}(\Psi_{i,i})^{-1} z_i = z_{-i} - \Psi_{-i,i} z_i \]

is not a function of unknown parameters. If $V_i(Z, W) = \sqrt{||W||^2 + Z_i^2}$, then conditioning on $S_i = (U_i(Z), V_i(Z, W)^2)$ yields a one-parameter exponential family with parameter $\mu_i/\sigma^2$.

Lemma 9 shows that $T_i$ is independent of $S_i$, and gives an explicit formula for reconstructing the other $t$-statistics from $T_i$ conditional on $S_i$.

**Lemma 9.** $T_i, U_i(Z),$ and $V_i(Z, W)$ as defined above are mutually independent when $\mu_i = 0$, and the $t$-statistic $T_j$ for $j \neq i$ can be reconstructed from them via

\[(26) \quad T_j = \frac{\sqrt{n-d + T_i^2}}{V_i(Z, W)} \cdot U_j(Z) + \Psi_{i,j} \cdot T_i, \]

As in Section 3.1, we can efficiently check the inequality

\[(27) \quad \mathbb{E}_0 \left[ \frac{1 \{q_i \leq c\}}{|R^{BH(\gamma\alpha)} \cup \{i\}|} \mid S_i \right] \leq \frac{\alpha}{m}, \]

for $c = q_i(Z, W)$ by holding $S_i$ fixed at their observed values and integrating with respect to $T_i \sim t_{n-d}$, where the other $t$-statistics are reconstructed via (26).
If $\Psi$ is diagonal then the second term in (26) vanishes. As a result, $T_j^2$ is conditionally an increasing function of $T_j^2$, so that for two-sided testing $p_j$ is increasing in $p_i$ and the test statistics are CPRD.

For right-tailed testing with diagonal $\Psi$, $T_j$ is in general a non-monotone function of $T_i$, so CPRD does not hold, but the relaxed version of the CPRD condition in Remark 2.3 applies when $\alpha < 0.5$, so that the second claim of Theorem 3 still holds. To see why, first note that $U_j(Z) = Z_j$ when $\Psi_{ij} = 0$. As a result, $T_j$ is (conditionally) either always negative, or always positive and increasing in $T_i$ whenever $T_i$ is also positive. Hence, the conditional probability of any increasing event of the form $\{R_{BH}(\alpha) \leq r\}$, which depends only on the positive $t$-statistics, is an increasing function of $p_i$ for $p_i \in (0, 0.5)$. When the test statistics are correlated, even this relaxed version of the CPRDS condition does not hold.

These results mirror the results of Benjamini and Yekutieli (2001), who found in the multivariate $t$ setting with diagonal $\Psi$ that the PRD condition holds for two-sided testing, and that the BH method still controls FDR for one-sided testing despite PRD not holding.

3.3. Testing coefficients in linear models. A third example is the Gaussian linear model wherein we observe fixed covariates $x_i \in \mathbb{R}^d$, with response

$$Y_i \sim x_i^\prime \beta + \epsilon_i,$$

for $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$,

with $\beta \in \mathbb{R}^d$ and $\sigma^2 > 0$ unknown, and test $H_j : \beta_j = 0$ or $H_j : \beta_j \leq 0$, for $j = 1, \ldots, m \leq d$.

If the design matrix $X = (x_1, \ldots, x_n)' \in \mathbb{R}^{d \times n}$ has full column rank then we can make a standard sufficiency reduction to the ordinary least squares coefficients and residual sum of squares (RSS):

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N_d(\beta, \sigma^2(X'X)^{-1}), \quad \text{and} \quad \text{RSS} = \|Y - X\hat{\beta}\|^2 \sim \sigma^2 \chi^2_{n-d}.$$

Let $\psi_j = ((X'X)^{-1})_{j,j}$. Thus the problem is reduced to the multivariate $t$ problem of Section 3.2, with $Z_j = \hat{\beta}_j / \sqrt{\psi_j}$, $\Psi_{i,j} = (X'X)^{-1} / \sqrt{\psi_i \psi_j}$, and $\|W\|^2 = \text{RSS}$.

Applying the logic of Section 3.2, if $X'X$ is diagonal then the two-sided $t$-statistics are CPRD and the one-sided $t$-statistics are close enough that the second claim in Theorem 3 holds.

If instead $\sigma^2$ is known, then the reduction to the multivariate $z$-statistics problem of Section 3.1 is even simpler. In that case $\hat{\beta}$ is sufficient and $Z_j = \hat{\beta}_j / \sqrt{\sigma^2 \psi_j}$. Defining the matrix $D = \text{diag}(\sigma^2 \psi_1, \ldots, \sigma^2 \psi_d)^{-1/2}$, $Z = (Z_1, \ldots, Z_d)$ is multivariate Gaussian with mean $\mu = D\beta$ and correlation matrix $\Sigma = D(X'X)^{-1}D$. Then the one-sided $z$-statistics are CPRD provided that all off-diagonal entries of $(X'X)^{-1}$ are non-negative (note this is very different from requiring positive pairwise correlations between variables).


4.1. An exact homotopy algorithm for dBH$_\gamma$($\alpha$). In this section we discuss an exact homotopy algorithm for dBH$_\gamma$($\alpha$). It can be easily generalized to dSU$_{\gamma, \Delta}$($\alpha$) at the cost of more complex notation; see Appendix C in the supplementary material for details. Recalling the definition of dBH$_\gamma$($\alpha$) in Section 2.3 and that of $\tau^{BH}(c; X)$ in (5), $g^*_i$ can be equivalently formulated as

$$g^*_i(c; S_i) = \sup_{P \in H_i} \mathbb{E} \left[ \frac{1\{p_i \leq cR^{BH}(c)(X)/m\}}{R^{BH}(\gamma_0)(X)} | S_i \right].$$

For all examples discussed in Section 3, $H_i$ is conditionally simple and $p_i = \eta_i(T_i)$ for some univariate transformation $\eta_i$ of test statistic $T_i$, and there exists a bijective mapping $\xi_i$ from $(T_i, S_i)$ to $(T_1, T_2, \ldots, T_m)$. For the multivariate Gaussian case, $T_i = Z_i,
\[ \eta_i(t) = 1 - \Phi(t) \text{ for one-sided testing and } \eta_i(t) = 2(1 - \Phi(|t|)) \text{ for two-sided testing, and} \]
\[ \xi_i(t; s) = (\xi_{i1}(t), \ldots, \xi_{im}(t)) \text{ where } \xi_{ij}(t) = t \text{ and } \xi_{ij}(t) = s_j + \Sigma_{j,i}t. \]
Since \( T_i \) is independent of \( S_i, g_i^*(c; S_i) \) can be equivalently formulated as the following univariate integral:
\[ g_i^*(c; S_i) = \int_{\mathbb{R}} \frac{1{\{\eta_i(t) \leq cR^{i(c)}_{\ell}(t)/m\}}}{{\cal R}^{(c)}_{\ell}(t)} dP(t) \]
where \( P_i(t) \) is the marginal distribution of \( T_i \) under \( H_i \), and \( R^{i(c)}_{\ell}(t) = \rho^{(c)}(\xi_i(t; S_i)) \) denotes the number of rejections by BH if the observed test statistics \( \xi_i(T_i; S_i) \) are replaced by \( \xi_i(t; S_i) \). As remarked at the end of Section 2.3,
\[ {\cal R}_+ = \{ i : g_i^*(c_i(X); S_i) \leq \frac{\alpha}{m} \}. \]
Therefore, it is left to compute \( R^{i(c)}_{\ell}(t) \) with \( c \in \{ c_i, \gamma \alpha \} \).

The test statistics \( (T_1, \ldots, T_m) \) as well as the induced p-values \( (p_1, \ldots, p_m) \) from the new dataset \( \xi_i(t; S_i) \) are both functions of \( t \): \( T_j(t) = \xi_{ij}(t) \) and \( p_j(t) = \eta_j(T_j(t)) = \eta_j(\xi_{ij}(t)) \). Since \( R^{i(c)}_{\ell}(t) \) is an integer-valued function, it must be piecewise constant. We call each point at which \( R^{i(c)}_{\ell}(t) \) changes the value a knot. Then a point is a knot only if \( p_j(t) \) crosses \( cr/m \) for some \( j, r \in [m] \).

For all examples discussed in Section 3, the domain of \( T_i \) is \( \mathbb{R} \), and \( \eta_i, \xi_{ij} \) are differentiable. Then the set of potential knots of \( R^{i(c)}_{\ell}(t) \) is
\[ K_i = \bigcup_{j,r \in [m]} K_{i,j,r}, \quad \text{where } K_{i,j,r} = \left\{ t : \eta_j(\xi_{ij}(t)) = \frac{cr}{m} \right\}, \]
For the one-sided multivariate Gaussian testing problem,
\[ K_{i,i,r} = \left\{ t : 1 - \Phi(t) = \frac{cr}{m} \right\} = \left\{ \Phi^{-1}\left(1 - \frac{cr}{m}\right) \right\}, \]
and for each \( j \neq i, \)
\[ K_{i,j,r} = \left\{ t : 1 - \Phi(S_{ij} + \Sigma_{j,i}t) = \frac{cr}{m} \right\} = \begin{cases} \left\{ (\Phi^{-1}\left(1 - \frac{cr}{m}\right) - S_{ij})/\Sigma_{j,i} \right\} & \text{if } \Sigma_{j,i} \neq 0 \\ \emptyset & \text{if } \Sigma_{j,i} = 0 \text{ and } 1 - \Phi(S_{ij}) \neq \frac{cr}{m} \\ \mathbb{R} & \text{otherwise} \end{cases}. \]
Note that \( S_{ij} \) has an absolutely continuous density, \( \mathbb{P}(K_{i,j,r} = \mathbb{R} \text{ for any } j, r) = 0. \) Thus, with probability 1,
\[ K_i = \bigcup_{r \in [m]} \bigcup_{j: \Sigma_{j,i} \neq 0} K_{i,j,r} \]
where \( K_{i,j,r} \) is a singleton. Similarly, for the two-sided multivariate Gaussian testing problem, it is easy to verify that \( K_i \) has the same form as above except that each \( K_{i,j,r} \) has two elements.

For multivariate t-statistics, \( K_{i,j,r} \) has a more complicated structure though it can still be computed efficiently; see Appendix C.3 in the supplementary material for details.

Let \( t_1 < t_2 < \ldots < t_N \) denote the elements of \( K \) with \( (j_1, r_1), (j_2, r_2), \ldots, (j_N, r_N) \) denoting the indices such that \( t_k \in K_{i,j_k,r_k} \). Let
\[ B^\ell(t) = \left\{ \ell : p_\ell(t) \leq \frac{c\ell}{m} \right\} - \ell, \quad (\ell = 0, \ldots, m). \]
By definition of BH(c),

$$R_i^{(c)}(t) = \max \left\{ \ell : B_i^{(c)}(t) = 0 \right\}. \quad (30)$$

As $t$ moves from $t_k-1$ to $t_k$, $B_i^{(c)}(t)$ remains the same for $\ell \neq r_k$ while $B_i^{(c)}(t_k)$ is incremented by $+1$ or $-1$, depending on whether $p_{jk}(t)$ is increasing or decreasing at $t_k$. For all examples discussed in Section 3, $p_{jk}(t)$ is differentiable, and thus

$$B_i^{(c)}(t_k) - B_i^{(c)}(t_{k-1}) = -\text{sign} \left( p_{jk}'(t_k) \right). \quad (31)$$

For one-sided multivariate Gaussian testing problems, $\eta_{jk}(z) = -\phi(z) < 0$ for any $z \in \mathbb{R}$ and $\xi_{ijk}'(t) = \Sigma_{j,k}$. As a result,

$$B_i^{(c)}(t_k) - B_i^{(c)}(t_{k-1}) = \text{sign}(\Sigma_{j,k}).$$

This motivates a homotopy algorithm to calculate $B_i^{(c)}(t)$ sequentially based on (31) and $R_i^{(c)}(t)$ based on (30). It is not hard to see the computational cost of the homotopy algorithm for a single hypothesis $H_i$ is $O(|K_i|)$. Therefore, the total cost for dBH$_\gamma$($\alpha$) is of order

$$\sum_i |K_i| \leq \sum_i \sum_j \sum_r |K_{i,j,r}|. \quad (32)$$

Naively, it requires $O(m^3)$ computation since there are $m$ summands for $i$ and $j$, corresponding to the hypotheses, and $m$ summands for $r$, corresponding to the thresholds. Nonetheless, we can significantly reduce the size of each sum above by using a step-up method similar to BH, but with sparse increments so that there are only $\log_2 m$ distinct threshold values:

$$\Delta_\alpha(r) = \frac{\alpha \beta(r)}{m}, \quad \text{with } \beta(r) = 2^{[\log_2 r]}. \quad (33)$$

We define the sparse dBH$_\gamma$ (s-dBH$_\gamma$) method as the dSU$_{r,\Delta}$ method with thresholds given in (33). In this case, s-dBH$_\gamma$ only requires $O(m^2 \log m)$ computation even in the worst case.

With all the tricks that are detailed in Appendix C (in the supplementary material), the number of summands in all of the three sums can be further drastically reduced. For multivariate Gaussian testing problems, the number of $i$ has the same order of $R_{BH}(2\alpha)$, the number of $j$ given $i$ has the same order as the range of non-negligible correlation, and the number of $r$ given $i$ and $j$ may be far lower than the total number of thresholds when $\Sigma_{j,i}$ is small. For the case with short-ranged dependence like in the autoregressive (AR) process, and a bounded number of signals, $R_{BH}(2\alpha)$ is bounded with high probability and thus the computation cost of dBH$_\gamma$($\alpha$) is at most $O(m)$. Thus, although the worst-case performance is poor, the cost is highly instance-specific and we find that the algorithm is reasonably fast in many cases. We illustrate in Appendix C.4 that the homotopy algorithm can scale to large-scale problems with millions of hypotheses.

### 4.2. An approximate numerical integration for dBH$_\gamma^2$($\alpha$)

Similar to (28), the conditional expectation $g_i^*(c; S_i)$ in dBH$_\gamma^2$($\alpha$) can be formulated as

$$g_i^*(c; S_i) = \int_{\mathbb{R}} \frac{1{\eta_i(t) \leq cR_i^{(c)}(t)/m}}{\bar{R}_i(t)} dP_i(t), \quad (34)$$

where $\bar{R}_i(t) = R_{dBH}^\gamma(\alpha)(\xi_i(t; S_i))$ denotes the number of rejections by dBH$_\gamma$($\alpha$) if the test statistics shift from $X = (T_i, S_i)$ to $\xi_i(t; S_i)$. Unlike dBH$_\gamma$($\alpha$), the denominator $\bar{R}_i(t)$ has a much more complicated structure and we do not have an efficient homotopy algorithm to
calculate the whole path. In principle, Monte-Carlo integration can guarantee almost sure convergence as the number of random samples grows to infinity because the integrand is bounded. However, it introduces extra randomness to the procedure which is undesirable. For this reason, we approximate (34) via a heuristic numerical integration method that has no guarantee in theory but works well in practice. For illustration, we focus on the one-sided multivariate Gaussian testing problem.

The first step is to reduce (34) to a finite-range integral. Since \( R_i^{(c)}(t) \leq m \), the integrand is 0 whenever \( \eta_i(t) > c \), or equivalently \( t < t_{lo} \equiv \Phi^{-1}(1 - c) \). On the other hand, let \( t_{hi} \equiv \Phi^{-1}(1 - \alpha \epsilon/m) \) for some \( \epsilon < 1 \), then the integral (34) from \( t_{hi} \) to \( \infty \) is upper bounded by \( \alpha \epsilon/m \) because the integrand is bounded by 1 and \( P_i \) is the standard Gaussian distribution. As a consequence, \( \int_{t_{lo}}^{t_{hi}} r_i(t) dP_i(t) \leq g_t^*(c; S_i) \leq \int_{t_{lo}}^{t_{hi}} r_i(t) dP_i(t) + \frac{\alpha \epsilon}{m} \), where \( r_i(t) \) denotes the integrand. If we take \( \epsilon \) to be small, e.g. \( \epsilon = 0.01 \), then the approximation error of \( \int_{t_{lo}}^{t_{hi}} r_i(t) dP_i(t) \) is negligible.

To compute \( \int_{t_{lo}}^{t_{hi}} r_i(t) dP_i(t) \), a naive method is to approximate \( r_i(t) \) by a piecewise constant function evaluated on an equi-spaced grid of \([t_{lo}, t_{hi}]\). However, it may be inefficient since \( r_i(t) = 0 \) whenever \( \eta_i(t) > cR_i^{(c)}(t)/m \). A simple improved version is to find the region of \( t \) in which \( \eta_i(t) \leq cR_i^{(c)}(t)/m \) using the exact homotopy algorithm for \( R_i^{(c)}(t) \), and then discretize the resulting region to approximate \( \int_{t_{lo}}^{t_{hi}} r_i(t) dP_i(t) \).

Naively, the computational cost is the product of the number of hypotheses \( m \), the grid size and cost of the homotopy algorithm to calculate a single \( R_i(t) \). However, as with the homotopy algorithm, we discussed a few tricks in Appendix C that can drastically reduce the number of hypotheses for which the integral \( g_t^*(q_i; S_i) \) needs to be computed. For instance, for a safe procedure, Theorem 4 guarantees that the hypotheses rejected by \( dBH_\gamma(\alpha) \) are also rejected by \( dBH^2_\gamma(\alpha) \), for which the computation of \( g_t^*(q_i; S_i) \) can be avoided. With all tricks discussed in Appendix C, it is even possible that no integral needs to be evaluated, in which case the computational cost of \( dBH^2_\gamma(\alpha) \) reduces to that of \( dBH_\gamma(\alpha) \). In a nutshell, the computational cost of the above algorithm is highly instance-specific.

5. Selected simulations. In Appendix D (in the supplementary material) we provide extensive simulations to compare the power of our approach with the power of several competing procedures including the \( BH(\alpha) \) and \( BH(\alpha/L_m) \) procedures as well as the fixed-X knockoffs (Barber and Candès, 2015), where appropriate. This section includes some highlights from our simulation results. In particular, we focus on the cases with scattered signals where the non-nulls p-values are highly correlated with some null p-values. In Appendix D.1, we consider another structure where the signals are clustered.

We start from a multivariate Gaussian case with \( m = 1000 \) and \( z \sim N_m(\mu, \Sigma) \) where

\[
\mathcal{H}_0^c = \{1, 101, 201, \ldots, 901\}, \quad \text{and} \quad \mu_i = \mu^*, \quad i \in \mathcal{H}_0^c.
\]

We consider two types of covariance structures: (1) an autoregressive structure with \( \Sigma_{ij} = (0.8)^{|i-j|} \); and (2) a block dependence structure with \( \Sigma_{ii} = 1 \) and \( \Sigma_{ij} = 0.5 \cdot 1(|i/20| = |j/20|) \). We perform both one- and two-sided testing using \( BH(\alpha) \), \( dBH_\gamma(\alpha) \), \( dBH^2_\gamma(\alpha) \), \( BY(\alpha) \), \( dBY(\alpha) \) and \( dBY^2(\alpha) \). All these methods are implemented in the \( R \) package \( dbb \). For one-sided testing, we choose \( \gamma = 1 \) because the p-values are CPRD, as shown in Section 3.1. For two-sided testing, we choose \( \gamma = 0.9 \). We set the level \( \alpha = 0.05 \) and tune the signal strength \( \mu^* \) such that \( BH(\alpha) \) has approximately 30% power in a separate Monte-Carlo simulation. We run each of the above 12 methods on 1000 independent samples of z-values and estimate the FDR and power, presented in Figure 3.
We observe that dBH and dBH\(^2\) slightly improve the power of BH, while dBY and dBY\(^2\) significantly improve the power of BY, for one- and two-sided testing with both covariance structures. For one-sided testing, the \(p\)-values are CPRD, so Theorem 4 guarantees that BH, dBH\(_1\), and dBH\(_2\) are all safe procedures with nested rejection sets. For two-sided testing, BH does not provably control the FDR, unlike the other five methods.

For two-sided testing, BH(\(\alpha\)) does not provably control FDR, but the other five methods do. Although the dBY\(_{0.9}^2(\alpha)\) procedure is not safe, the randomized pruning is never invoked over 1000 realizations of each simulation scenario. In all four scenarios, the power of dBY\(^2\) is comparable to that of BH.

Figure 4 shows results for uncorrelated multivariate t-statistics with either \(m = d = 100, n - d = 5\) (Figure 4a) or \(m = d = 1000, n - d = 50\) (Figure 4b). In the first case, the marginal null distribution of each test statistic is heavy-tailed, and very large values tend to
be observed together due to the common variance estimate. In both cases we set the first 10 hypotheses as alternatives with an equal signal strength, tuned so that BH(0.05) has approximately 30% power. We evaluate the same six methods as in the multivariate Gaussian case, except that \( \gamma \) is taken as 1 for both one- and two-sided testing because both are CPRD. The results are qualitatively similar to the multivariate Gaussian results. Notably, the power gains of dBH_1 and dBH_2 over BH are more pronounced for heavier-tailed t-statistics.

Finally, we consider two linear modeling scenarios, for which we evaluate the fixed-X knockoff method as an extra competitor. To apply the knockoff method, we always consider two-sided testing problems with \( n > 2d \). In this section we simulate the fixed design matrix \( X \) as one realization of a random matrix with i.i.d. Gaussian entries with \( n = 3000 \) and \( d = 1000 \), and simulate 1000 independent copies of homoscedastic Gaussian error vectors with \( \sigma^2 = 1 \), each generating an outcome vector \( Y = X\beta + \epsilon \) with \( \beta_1 = \cdots = \beta_{50} = \beta^* \) and \( \beta_{51} = \cdots = \beta_{1000} = 0 \). Again, \( \beta^* \) is tuned so that BH(0.05) has approximately 30% power. For all dBH_\gamma procedures, we choose \( \gamma = 0.9 \) and find that the randomized pruning step is never invoked for any method in the 1000 simulations. For the knockoff method, we generate the knockoff matrix via the default semidefinite programming procedure and choose the knockoff statistic as the maximum penalty level at which the variable is selected (Barber and Candès, 2015). We use the knockoff+ method in all cases to ensure FDR control at the advertised level. The estimated FDR and power with \( \alpha = 0.05 \) and \( \alpha = 0.2 \) are shown in Figure 5a. The comparison between the dBH (dBY) procedures and BH (BY) procedure is qualitatively similar to the previous examples. The fixed-X knockoff has much higher power than all other methods when \( \alpha = 0.2 \), but has near-zero power when \( \alpha = 0.05 \). The former may result from the knockoff method’s use of the lasso for variable selection (Tibshirani, 1996), while the latter is due to the small-sample issue discussed in Section 7.1. Appendix B in the supplementary material gives several more linear modeling examples showing the same qualitative pattern for random design matrices with different correlation structures.

![Fig 5: Estimated FDR and power for two types of linear models.](image)

While the knockoff method often outperform the others when \( \alpha \) is large enough, the reverse can also occur, as we illustrate in a second linear modeling example: the problem of multiple comparisons to control (MCC) in a one-way layout. For each of 100 treatment groups, we sample 30 independent replicates from \( N(\mu_i, \sigma^2) \), and a control group with 30
independent replicates sampled from $N(\mu_0, \sigma^2)$. We then test $H_i : \mu_i = \mu_0$ based on the two-sample t-statistics. In this case, the test statistics are positively equi-correlated. By coding dummy variables this MCC problem is equivalent to a homoscedastic Gaussian linear model with a design matrix $X \in \{0, 1\}^{30 \times 101}$ and a coefficient vector $(\mu_1 - \mu_0, \ldots, \mu_{100} - \mu_0, \mu_0)$.

To ameliorate the small-sample issue of the knockoff method, we set the first $30$ hypotheses to be non-nulls with an equal $\mu_i$ that is tuned so that BH($\alpha$) has approximately 30% power. All of the coefficients are inferential targets except $\mu_0$, which is effectively an intercept term. For fixed-X knockoffs, we follow Barber and Candès (2015) to generate a knockoff matrix that is orthogonal to the column corresponding to $\mu_0$, which is a vector with all entries 1 in this case. The results are presented in Figure 5b. In this case, the fixed-X knockoff is almost powerless for either $\alpha = 0.05$ or $\alpha = 0.2$. In contrast to the previous case, the lack of power is caused by the huge amount of noise generated by knockoffs to handle the equi-correlated covariance structure. The other six methods are less sensitive to this correlation.

6. HIV drug resistance data. This section compares our method’s performance against the BH, BY, and knockoff procedures on the Human Immunodeficiency Virus (HIV) drug resistance data of Rhee et al. (2006), reproducing and extending the analysis of Barber and Candès (2015). In each of three separate data sets, we test for associations between mutations present in different HIV samples and resistance to each of 16 different drugs. The data come from three experiments, each for a different drug category: protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs), and nonnucleoside reverse transcriptase inhibitors (NNRTIs).

Following Barber and Candès (2015), we encode mutations as binary with $x_{ij} = 1$ if the $j$th mutation is present in the $i$th sample, discard mutations that occur fewer than three times, and remove duplicated columns in the resulting design matrix $X \in \{0, 1\}^{n \times d}$. For each drug there is a different response vector $Y \in \mathbb{R}^n$ representing a measure of drug resistance. As in Barber and Candès (2015) we do not include an intercept in the model. We also evaluate replicability in the same way as Barber and Candès (2015), by comparing the rejection set to the set of mutations identified in the treatment-selected mutation (TSM) panel of Rhee et al. (2005). We refer to Section 4 of Barber and Candès (2015) for further details.

Figure 6 shows results comparing results for the fixed-X knockoffs, BH, $d\text{BH}_{0.9}^2$, and $d\text{BY}^2$, at significance level $\alpha = 0.2$, as used in Barber and Candès (2015). For the knockoff method, we generate equi-correlated and SDP-based knockoff copies and use as the knockoff statistic the maximum penalty level at which the variable is selected, following the vignette of the knockoff package (Patterson and Sesia, 2018). The performance of two knockoff methods are similar, so we only present the results for SDP-based knockoffs. The latter three have similar power for all seven responses, with the behavior of $d\text{BH}_{0.9}^2$ nearly identical to BH and $d\text{BY}^2$ very slightly less powerful. By contrast, the knockoff method makes somewhat fewer rejections overall than the other methods, but the differences are modest for most drugs. The knockoff method appears to have a higher replicability rate for the TSM panel, possibly because the method is achieving a better tradeoff between Type I and Type II error by using the lasso algorithm to select variables. Alternatively, it may be that the other methods are better able to detect weak signals which are less likely to be replicated in an independent experiment. The $d\text{BH}_{0.9}^2$ method does not require randomization for any of the 16 drugs. 

Figure 7 shows the same results at the more conservative significance level $\alpha = 0.05$, where knockoffs suffers from the small-sample issues discussed in Section 7.1. The relationships between the other three methods are qualitatively the same. Again, $d\text{BH}_{0.9}^2$ does not require randomization for any of the drugs.
7. Discussion. We have presented a new approach for controlling FDR in dependent settings, and proposed new dependence-adjusted step-up methods including the $\text{dBH}^1$, $\text{dBY}$, and $\text{dSU}_{\Delta}$ procedures. The $\text{dBH}^1$ procedure uniformly improves on the BH procedure under (conditional) positive dependence, while the $\text{dBY}$ procedure uniformly improves on the BY procedure. Likewise, our $\text{dSU}$ method can uniformly improve on any shape function method in the style of Blanchard and Roquain (2008).

Practically speaking, our methods offer an alternative to the BH and BY procedures in applications where theoretical FDR control guarantees are attractive. In particular, $\text{dBY}^2$ improves dramatically on the BY procedure and is often competitive even with BH. The $\text{dBH}^2_{0.9}$ procedure offers a balanced approach that is commonly more powerful than BH and
Fig 7: Results on the HIV drug resistance data with $\alpha = 0.05$. The blue segments represent the number of discoveries that were replicated in the TSM panel, while the orange segments represent the number that were not. Results are shown for the fixed-X knockoffs, BH, dBH$_{0.9}$, and dBY.

requires randomization only very rarely, and the dBH$_{0.9}$ procedure offers a scalable version of the same.

More generally, conditional calibration as proposed here is a general-purpose technical device that may prove useful for supplying FDR control proofs in other contexts like grouped, hierarchical, multilayer, or partial conjunction hypothesis testing (e.g. Benjamini and Bogomolov, 2014; Barber and Ramdas, 2016; Lynch and Guo, 2016; Benjamini and Heller, 2008). Numerous challenges remain for future work. One major limitation of our method is the fact that the pruning step does not take into account the size of the original $p$-values, which is hard to justify on scientific grounds. The pruning would be less problematic if the strongest rejections could be preferentially preserved. In addition, further development is needed to
produce algorithms and software for some of the models we did not implement in this work. Finally, the next sections suggest directions of further methodological innovation.

7.1. Comparison with knockoffs. Both the dBH method and the knockoff filter offer finite-sample FDR control for linear models, but with very different statistical tools and methods of proof. In our simulation experiments neither method is a clear overall winner, but some qualitative trends emerge. First, as expected, the dBH procedure performs similarly to BH in power comparisons, so any comparison between knockoffs and dBH is also a comparison between knockoffs and BH. Second, the dBH and BH procedures consistently enjoy better power in experiments where the total number of rejections is relatively small, either because there are very few non-null coefficients to find, the signals are weak, or the FDR significance level is small. This pattern has a clear theoretical explanation: to make rejections, the knockoff+ method requires \((1 + A_t) / R_t \leq \alpha\), where \(A_t\) is the count of \(W_j\)-values smaller than \(-t\) and \(R_t\) is the count above \(t\). As a result the method must either make at least \(1 / \alpha\) rejections or make none at all, and it can be unstable if the number of rejections is on the order of \(1 / \alpha\).

Apart from small-sample issues, it remains unclear in which contexts we should expect one method to outperform the other, and this is an interesting question for future research. Because the knockoffs framework is very general and allows the analyst to bring a great deal of prior knowledge to bear, we expect it can enjoy substantial advantages over BH and dBH in problems where the \(j\)th \(t\)-statistic carries only a small fraction of the total evidence against \(H_j: \beta_j = 0\). In particular, \(U = \hat{\beta}_{-j} - \hat{\beta}_j b_j^2\) is independent of the \(t\)-statistic and \(p\)-value for \(H_j\), but it may hold a wealth of information about \(\beta_j\), especially if we reasonably expect that \(\beta\) is approximately sparse. In our view, it is an important open problem to develop methods that can likewise exploit this kind of information, for example by using adaptive weights as proposed below, while avoiding the randomization and binarization inherent to knockoff methods. In recent work Sarkar and Tang (2021) devise a hybrid method that interprets the knockoff filter as producing two independent estimators for \(\beta\), the latter having diagonal covariance and therefore satisfying CPRDS. This work offers a potentially interesting way forward.

Our method is also extensible to many settings where no knockoff method has been proposed, for example the discrete and nonparametric examples discussed in Appendix B. Because our method operates directly on \(p\)-values it is easily extensible to testing composite hypotheses about parameters, for example to test \(H_j: |\beta_j| \leq \delta\) for a fixed \(\delta > 0\). Finally, in regression problems, there is no requirement that \(n \geq 2d\); we require only that \(n \geq d + 1\), the same dimension required to test individual regression coefficients. Conversely, we have not extended our framework to conditional randomization tests as proposed in Candès et al. (2018), and this may be very challenging in general. Computationally, our method is more scalable for some problems because it avoids solving a semidefinite program or eigendecomposing a large matrix, but the recursively refined variants of our method pose substantial computational challenges of their own. Nevertheless, computational efficiency depends on problem specifics.

7.2. Extension: adapting to the non-null proportion. One arguable weakness of the present work is its conservative control of the FDR at level \(\alpha m_0 / m\). In some problems it would be very useful to correct for this conservatism; for example, in post-screening or other post-selection inference, we may expect \(\pi_0 = m_0 / m\) to be substantially smaller than 1. For independent \(p\)-values, various plug-in methods apply a standard method such as BH at an adjusted level \(\alpha \hat{\pi}_0^{-1}\), for some estimator \(\hat{\pi}_0(X)\). (e.g. Genovese and Wasserman, 2002; Storey, 2002; Storey, Taylor and Siegmund, 2004; Benjamini, Krieger and Yekutieli, 2006; Blanchard and Roquain, 2009).
Inspired by this approach, we can modify our calibration procedure as discussed in Remark 2.1 to use the calibration constraint \( \kappa_i(S_i) = \alpha \hat{\pi}_{0,i}(S_i)^{-1}/m \) in place of \( \alpha/m \) in (8). The resulting method would control FDR at level \( \alpha \) provided that

\[
\sum_{i \in H_0} \mathbb{E}\hat{\pi}_{0,i}(S_i)^{-1} \leq \pi_{0,i}(S_i) - 1/m = m,
\]

for which a sufficient condition is that \( \mathbb{E}_{H_i} \hat{\pi}_0(S_i)^{-1} \leq \pi_0^{-1} \) for each \( i = 1, \ldots, m \). Given any pre-existing estimator \( \tilde{\pi}_0(X) \) for which \( \mathbb{E}\tilde{\pi}_0(X)^{-1} \leq \pi_0^{-1} \), we can construct such an estimator by Rao-Blackwellization:

\[
\hat{\pi}_{0,i}(S_i) = \left( \mathbb{E}_{H_i} \left( \tilde{\pi}_0(X)^{-1} \mid S_i \right) \right)^{-1},
\]

which amounts to a simple calculation if \( H_i \) is conditionally simple.

7.3. Extension: adaptive weights. Another promising extension, the full exploration of which is outside the scope of this work, is to use adaptive weights that exploit side information about the hypotheses. There are a variety of setting where \( p \)-value weights \( w_i \) can substantially improve the power of multiple-testing methods (Benjamini and Hochberg, 1997; Genovese, Roeder and Wasserman, 2006; Dobriban et al., 2015). For fixed weights \( w_1, \ldots, w_m \) that sum to one, it is straightforward to generalize our framework by replacing \( \alpha/m \) by \( \alpha w_i \) in the right-hand side of (8).

More interestingly, however, we might wish to learn the weights from the data, adaptively allowing for some hypotheses to contribute more to the FDR than others. Sun and Cai (2009) discuss methods that exploit dependence between the underlying parameters for “nearby” hypotheses, while assuming the test statistics are independent conditional on covariates, relying on the compound decision theoretic framework of Sun and Cai (2007). This work has been extended to a robust literature on adaptive \( p \)-value weighting for hypotheses with independent test statistics conditional on covariates (e.g. Ignatiadis et al., 2016; Boca and Leek, 2017; Li and Barber, 2019; Lei and Fithian, 2018; Ignatiadis and Huber, 2017; Xia et al., 2017; Tansey et al., 2018). However, there is very little work on adaptive weighting methods when the \( p \)-values are dependent conditional on covariates. This is a major gap in the literature, since true independence between \( p \)-values is rare in applied problems.

Similarly to the strategy described above for estimating \( \pi_0 \), we can accommodate data-adaptive weights by using \( \kappa_i(S_i) = \alpha w_i(S_i) \), provided that \( \sum_{i \in H_0} \mathbb{E}w_i(S_i) \leq 1 \). As above, we can Rao-Blackwellize initial weights \( \tilde{w} \) by setting \( w_i(S_i) = \mathbb{E}_{H_i} \left[ \tilde{w}_i \mid S_i \right] \). If \( \sum_i \tilde{w}_i(X) \leq 1 \) almost surely, then

\[
\sum_{i \in H_0} \mathbb{E}\kappa_i(S_i) = \alpha \sum_{i \in H_0} \mathbb{E}\tilde{w}_i \leq \alpha.
\]

We defer exploration of this idea to future work.

Reproducibility. Our \texttt{R} package \texttt{dbh} is available to download at https://github.com/lihualei71/dbh. A public github repo accompanying the paper with code to reproduce the figures herein can be found at https://github.com/lihualei71/dbhPaper.

Acknowledgments. William Fithian is supported in part by the NSF DMS-1916220 and a Hellman Fellowship from Berkeley. We are grateful to Patrick Chao and Jonathan Taylor for helpful feedback on a draft of this paper.
REFERENCES


