Adaptive procedures for boundary FDR control

Sarah Mostow and Daniel Xiang

November 15, 2025

Abstract

A cornerstone of the multiple testing literature is the Benjamini-Hochberg (BH) procedure, which guarantees control of the FDR when p-values are independent or positively dependent. While BH controls the average quality of rejections, the guarantee can be misleading for discoveries made near the rejection threshold, which are typically more likely to be false than the average rejection. For independent p-values with $\mathrm{Uniform}(0,1)$ null distribution, the Support Line procedure (SL, Soloff et al. (2024)) provably controls the error probability for the rejection at the edge of the discovery set (i.e. the one with largest p-value) at level qm_0/m, where m_0 is the number of true null hypotheses and q is a tuning parameter. In this work, we study an adaptive version of the SL procedure that operates in two steps: the first step estimates m_0 from non-significant statistics, and the second step re-runs the SL procedure at an adjusted level $qm/\hat{m}_0$. The adaptive procedures are shown to control the false discovery probability for the "boundary" rejection under an independence assumption. Simulation studies suggest that some but not all of the proposed two-stage procedures maintain error control under positive dependence, and that substantial power is gained relative to the original SL procedure. We illustrate differences between the procedures on publicly available meta-data from the recent literature in behavioral psychology.

1 Introduction

Statisticians have long deliberated how to determine, given a list of p-values arising from a simultaneous multiple testing experiment, which ones to consider significant. Testing each null hypothesis at the traditional per-comparison level of $\alpha = 0.05$ ignores the multiple testing problem and has led to replicability issues in areas of psychological research (Collaboration (2015); Benjamin et al. (2018)). The Bonferroni correction resolves the issue of inflated type 1 error by dividing the threshold by the number of tests, but can be overly conservative in scientific applications (see e.g. Reiner et al. (2003); Risch and Merikangas (1996)). Middle ground is attained by the False Discovery Rate (FDR, Benjamini and Hochberg (1995)), which measures the expected proportion of the discovery set consisting of false positives. By tolerating a fixed proportion of false rejections, the FDR approach allows the analyst to reject more null hypotheses than they would with a family-wise correction, while maintaining a form of type 1 error control more stringent than uncorrected testing.

While the FDR approach corrects for multiplicity, it can also suffer from the "free-rider" problem: a large p-value may be included in the rejection set not on its own merits but simply due to the presence of a few very small p-values in the dataset. Finner and Roters (2001)

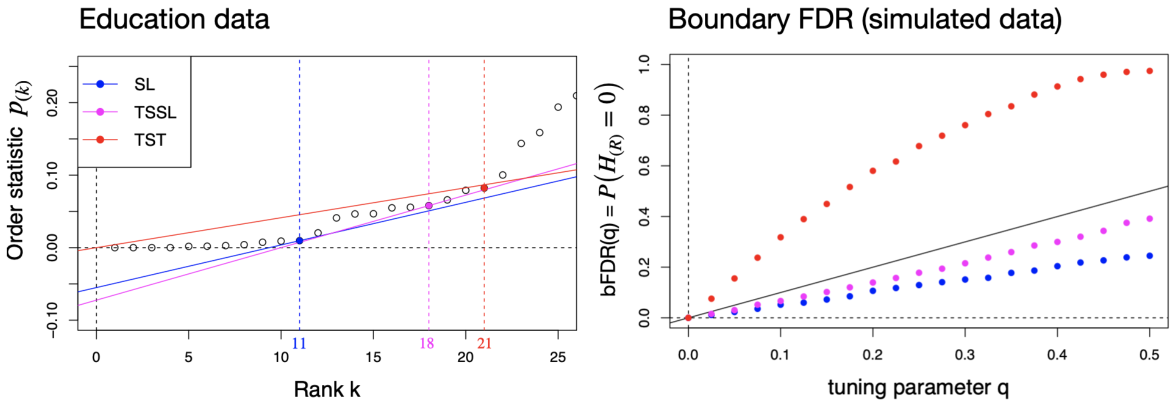


Figure 1: In the left panel, the TST(q'), SL(ℓ) and TSSL(ℓ') procedures are illustrated on an educational dataset from Williams et al. (1999), with $q = 0.1$ and $\ell = 0.2$. In the right panel, we plot the bFDR curve for the procedures in the "1234 configuration" simulation setting of Benjamini et al. (2006), where $\pi_0 = 1/2$ and non-null p-values are computed from a one-sided Gaussian location test with mean μ equal to 1, 2, 3, or 4 with equal weight.

describe the "problematic nature of the FDR concept", where by including some p-values that are known to exhibit strong evidence, the analyst may test their favored hypothesis at a more lenient significance level. Soloff et al. (2024) proposed a correction using the Support Line (SL) procedure and showed that its last rejection (i.e. the one with largest p-value) is a false positive with probability $\pi_0 q$ under independence assumptions, where π_0 denotes the proportion of null hypotheses that are true. The right panel of Figure 1 shows this probability as a function of the tolerance parameter $0 \leq q \leq 1$ for the BH (red) and SL (blue) procedures in a simulation setting from Benjamini et al. (2006). For $q = 0.2$, the last discovery of the FDR procedure is over 50% likely to be false, while the last discovery of SL run at level $q = 0.4$ is false less than 20% of the time. We review this error criterion, called the boundary false discovery rate (bFDR, Soloff et al. (2024); Xiang et al. (2025)), and the SL procedure in Section 2.

Under independence assumptions, the procedures (BH, SL) control their respective error rates (FDR, bFDR) at level $\pi_0 q$. Given π_0, we could in principle run these procedures at level q/π_0 to exhaust the prescribed error tolerance, achieving the desired level q exactly. A substantial body of work has sought to incorporate an estimate of π_0 into FDR-controlling procedures by developing adaptive methods (see e.g. Benjamini and Hochberg (2000); Storey (2002); Benjamini et al. (2006); Gao (2025)), which we review in Section 2. The general strategy is to construct a data-driven estimate $\hat{\pi}_0 > 0$ and then apply BH at level $q/\hat{\pi}_0$. Roughly speaking, if $1/\hat{\pi}_0$ is expected to be smaller than $1/\pi_0$ on average, this strategy increases the power of BH while maintaining FDR $\leq q$.

In this paper, we develop analogous methods for bFDR control by adapting the SL procedure to data-driven estimates of π_0. We provide theoretical results on their finite-sample bFDR control (Section 3) and simulation studies demonstrating the performance gains of these methods relative to the standard SL procedure (Section 4) for independent and positively dependent test statistics. Figure 1 illustrates the two stage SL procedure (TSSL), an adaptive method that estimates the number of true null hypotheses with the number of non-significant p-values according to SL run at level $q' := \frac{q}{1+q}$, and then re-runs the SL procedure at level $q'/\hat{\pi}_0$. This procedure is illustrated along with the standard (unadjusted) SL procedure and an adaptive BH procedure (TST, Benjamini et al. (2006)) on an educational dataset analyzed by Benjamini and Hochberg (2000) in the left panel of Figure 1. TSSL with bFDR tolerance $q = 0.2$ makes more rejections than SL, and fewer than the adaptive BH procedure with FDR tolerance $q = 0.1$. Shown in the right panel are the results of a numerical experiment where the TSSL procedure makes more rejections than the original SL procedure while still controlling bFDR below the tuning parameter. In Section 3, we also consider Storey-type estimators for π_0 where the hyper-parameter λ is either fixed before seeing data (Schweder and Spjøtvoll, 1982; Storey, 2002; Soloff et al., 2024), or selected based on a data-driven stopping rule (Gao, 2025).

2 Background

2.1 Boundary FDR and SL procedure

Let $H_1, \dots, H_m$ denote the null hypotheses, $\mathcal{H}_0 := \{i : H_i \text{ is true}\}$ the null set, and $p_1, \dots, p_m$ are p-values satisfying $p_i \sim \text{Uniform}(0, 1)$ when H_i is true. The Support Line procedure (SL, Soloff et al. (2024)) is defined

$$R = \text{argmax}_{k=0,\dots,m} \left\{ \frac{qk}{m} - p_{(k)} \right\}, \tag{1}$$

and the rejection set is $\mathcal{R} := \{i : p_i \leq p_{(R)}\}$, where $p_{(0)} := 0$. If $(p_i)_{i \in \mathcal{H}_0}$ are independent of each other and of $(p_i)_{i \notin \mathcal{H}_0}$, then

$$\text{bFDR}(\mathcal{R}) := \mathbb{P}(H_{(R)} \text{ is true}, R > 0) = \pi_0 q, \tag{2}$$

where $\pi_0 := |\mathcal{H}_0|/m$ and the notation $H_{(1)}, \dots, H_{(m)}$ means the null hypotheses ordered according to $p_{(1)} \leq \dots \leq p_{(m)}$, where $H_{(0)} :=$ false, by convention. Note that while the truth statuses of $H_1, \dots, H_m$ may be fixed to begin with, once they are re-ordered according to $p_{(1)} \leq \dots \leq p_{(m)}$, the truth status of $H_{(i)}$ is a non-trivial Bernoulli random variable (Genovese and Wasserman, 2002).

Definition (2) applies to multiple testing procedures that reject for p-values below a threshold $\hat{\tau}$. In principle, one could consider rejection regions other than $[0, \hat{\tau}]$ and extend the definition of "boundary" to the union of boundary points along the edges of the rejection set, but it is not clear in general whether these correspond to the least promising rejections. In this paper, we only consider multiple testing procedures with a rejection region of the form $[0, \hat{\tau}]$.

The SL procedure has a geometric interpretation, which is illustrated in Figure 1. On the plot of p-value order statistics, one can imagine bringing a line of slope q/m starting from height $-\infty$ and up until the plot of order statistics lays tangent to this line. The x-axis value at the tangent point (where the plot is "supported" by the line) determines the number of rejections.

2.2 Adaptive procedures for FDR control

Benjamini and Hochberg (2000) proposed the Lowest Slope (LSL) procedure, which also has a geometric interpretation. First the p-values are sorted from smallest to largest $p_{(1)} \leq \dots \leq$

$p_{(m)}$. Starting from $i = 1$, i.e. the smallest p-value $p_{(1)}$, at each step, the slope of the line passing through $(m+1, 1)$ and $(i, p_{(i)})$ is computed and compared to the slope passing through $(m+1, 1)$ and $(i-1, p_{(i-1)})$, with $p_{(0)} := 0$. If the former slope is smaller, then we consider the next smallest p-value $p_{(i+1)}$ and repeat. The iteration stops when the slope increases relative to the previous step. Intuitively, this is equivalent to the Storey procedure with $\lambda = p_{(i)}$, since

$$\hat{\pi}_0^\lambda := \frac{1 + \#\{i : p_i > \lambda\}}{m(1-\lambda)} \text{ with } \lambda = p_{(i)} \Rightarrow \hat{\pi}_0^\lambda = \frac{1+m-i}{m(1-p_{(i)})}$$

is the reciprocal of the slope at step i. The estimator is $\hat{m}_0 = \min\{\lceil 1/S_j \rceil, m\}$, where j is the first time $\hat{\pi}_0^{p_{(i)}}$ stops decreasing, and the corresponding FDR procedure is to run BH at the adjusted level $qm/\hat{m}_0$. This procedure has been studied numerically by Benjamini et al. (2006), and an approximate justification for the procedure's FDR guarantee holds for independent p-values using the proof method in Gao (2025). Schweder and Spjøtvoll (1982) proposed to use the proportion of p-values above a fixed threshold λ to estimate $\hat{\pi}_0$, a method later popularized by Storey (2002). Gao (2025) refined Storey's estimator by introducing a data-driven stopping rule for choosing λ that is generally more powerful than the original π_0 estimate of Benjamini and Hochberg (2000) and retains finite-sample FDR control for independent statistics.

Benjamini et al. (2006) studied adaptive procedures for FDR control and proposed a two-stage adaptive BH procedure that uses the number of non-rejections by the BH procedure in the first stage to estimate m_0. Our goal in the current paper is similar: to formalize the bFDR guarantee for the two-stage SL procedure under independence, and compare the adaptive SL procedure applied with various estimators of π_0 in simulations and on real data.

3 The adaptive bFDR procedures

Two-stage SL procedure. The two-stage support line (TSSL) procedure is defined as follows.

1. Run SL at level q, i.e. compute

   $$R_1 := \text{argmax}_{k=0,\ldots,m} \left\{ \frac{qk}{m} - p_{(k)} \right\}.$$

   If $R_1 = 0$, stop and reject nothing; if $R_1 = m$ reject all m hypotheses and stop; otherwise, go to Step 2.

2. Re-run SL at level $qm/(m - R_1)$, i.e. compute

   $$R_2 := \text{argmax}_{k=0,\ldots,m} \left\{ \frac{qk}{m - R_1} - p_{(k)} \right\},$$

   and reject $H_{(1)}, \ldots, H_{(R_2)}$.

The procedure is equivalent to estimating π_0 using the proportion of non-rejections at the first stage, and re-running SL with the adjusted tuning parameter $q/\hat{\pi}_0$ in the second stage.

This two-stage SL procedure is the 'local' analogue to the TST procedure in Benjamini et al. (2006) and controls its boundary FDR under independence below an inflated error level, as our next result shows. We remark that this result is analogous to Theorem 1 in Benjamini et al. (2006), which states a guarantee for the FDR of a two-stage BH procedure. A proof is recorded in Section 7.

Theorem 3.1. *Let $H_1, \dots, H_m$ denote m null hypotheses, with independent p-values $p_1, \dots, p_m$. Suppose that $p_i \sim$ Uniform$(0,1)$ if H_i is true. Then*

$$\mathrm{bFDR}(\mathcal{R}_2) \leq \frac{q}{1-q},$$

where $\mathcal{R}_2 := \{i : p_i \leq p_{(R_2)}\}$ is the rejection set for the two stage procedure, and the bFDR *of a rejection set is defined in (2).*

Remark 3.1. *If one wishes to control bFDR below level q with provable guarantee under independence, one could run the TSSL procedure at the reduced level $q' := q/(1+q)$. We note that this is analogous to the TST procedure of Benjamini et al. (2006) which runs BH at the reduced level q' in both stages and provably controls FDR below q for independent test statistics.*

Adaptive Storey adjustment. The Adaptive Storey (AS) null proportion estimator (Gao, 2025) is defined:

$$\hat{\pi}_0 \equiv \hat{\pi}_0(p_1, \dots, p_m) := \frac{1 + \#\{i \leq m : p_i > \hat{\lambda}\}}{m(1-\hat{\lambda})},$$

where $\hat{\lambda}(p_1, \dots, p_m) \geq q$ is any stopping time with respect to the filtration:

$$\mathcal{F}_t = \sigma\Big(\sum_{i=1}^{m} 1\{p_i > s\} : q \leq s \leq t \Big), \qquad t \geq q.$$

Consider the SL threshold with tuning parameter $q/\hat{\pi}_0$ defined in the following way:

$$R_{q/\hat{\pi}_0} = \mathrm{argmin}_{k : p_{(k)} \leq q} \Big\{ p_{(k)} - \frac{qk}{\hat{\pi}_0 m} \Big\}.$$

Our next result shows that this adjustment preserves bFDR control for any valid stopping time $\hat{\lambda}$. The proof is adapted from arguments from Soloff et al. (2024) and Gao (2025), and is recorded in Section 7.

Theorem 3.2. *Let $H_1, \dots, H_m$ be null hypotheses, and suppose that for any i where H_i is true, we have that $p_i \sim$ Uniform$(0,1)$ and that p_i is independent of $p_{-i} := (p_j : j \in [m] \backslash \{i\})$. Then*

$$\mathrm{bFDR}(\mathcal{R}_{q/\hat{\pi}_0}) \leq q.$$

For practical reasons, Gao (2025) recommended searching for λ over a grid, $q, q+\delta, q+2\delta, q+3\delta, \dots$ for some small number $\delta > 0$, e.g. $\delta = 1\%$ or 2%. Then $\hat{\lambda}$ is chosen by successively checking the value of $\hat{\pi}_0^{\lambda}$ for each λ in the grid, starting from q and incrementing λ along the grid until the first time that the associated estimate $\hat{\pi}_0^{\lambda}$ stops decreasing, and taking $\hat{\lambda}$ to be this grid point. This describes a valid stopping rule, so Theorem 3.2 implies that the Storey$(\hat{\lambda})$ adjusted SL procedure achieves finite sample bFDR control. We study these procedures numerically in Section 4.

Storey with fixed threshold. Soloff et al. (2024) proposed to adjust the SL procedure using the Storey estimator with $\lambda = 1/2$. Our simulations show this procedure performs favorably in the independent case, relative to all other procedures we consider here, but worse in settings with strong dependence. In the dependent case, we recommend using the smaller level $\lambda = q$, e.g. $\lambda = 0.2$. This π_0 estimator was noted to work well with BH under dependence settings (Blanchard and Roquain, 2009). We find an analogous phenomenon holds for SL and bFDR.

Lowest slope (LSL) estimator. The lowest slope estimate of m_0, also known as *the original* m_0 *estimate* by Benjamini and Hochberg (2000), was described in Section 2.2. Letting $\hat{\pi}_0^{\mathrm{LSL}} := \hat{m}_0/m$, the SL procedure adapted with this estimate of π_0 can be approximately justified using Theorem 3.2. The reason is that for small q, the adaptive Storey procedure with grid $\{q, p_{(i^*)}, p_{(i^*+1)}, \dots\}$ where $i^* = \min\{j : p_j \geq q\}$, is practically identical to the LSL estimate, as the value of λ used for the Storey estimate is typically chosen larger than q.

4 Numerical illustration

In this section, we study the type I error and power of the two-stage procedures on simulated data generated as in the numerical settings of Benjamini and Hochberg (1995) and Benjamini et al. (2006). Specifically, one-sided p-values are computed from unit variance Gaussian test statistics whose means are set according to one of the following configurations.

Alternating configuration. The non-null means are set to

$$\mu_1 = 5 \times 1/4,\ \mu_2 = 5 \times 2/4,\ \mu_3 = 5 \times 3/4,\ \mu_4 = 5 \times 4/4,$$

repeating cyclically for the first m_1 coordinates, where m_1 is a multiple of 4. The null means are all set to zero.

"All at 5" configuration. In this case, the non-nulls are strongly separated from the nulls: the non-null means are equal to 5, and the null means are all equal to 0, i.e.

$$\mu_1 = \cdots = \mu_{m_1} = 5,\ \mu_{m_1+1} = \cdots = \mu_m = 0.$$

Overview. In Section 4.1, we assess the boundary FDR as a function of the tuning parameter q for each procedure in the setting with independent p-values in both the alternating mean configuration and the all-at-5 configuration, and compare the power of these procedures in the alternating mean setting (as it is more realistic). In Section 4.2, we compare the bFDR of the procedures under positive dependence, and also the variability of the null proportion estimators computed in the first stage.

In principle, the local fdr (Efron et al., 2001) is highly relevant for assessing individual probabilities of type 1 error, but estimates of local fdr may be more noisy than useful in unstructured settings with only a few tests. In Section 4.3, we assess the variability of the lfdr estimates at the oracle threshold, and also the variability of the true local fdr at the estimated thresholds. Unless otherwise specified, the number of tests is set to $m = 64$, the null proportion is $\pi_0 = 0.75$, and the tuning parameter is $q = 0.2$. The procedures we compare are as follows:

1. **TSSL(q):** Two-Stage Support Line using q at both stages, which has guaranteed bFDR $\leq \frac{q}{1-q}$ in the independent case.

2. **TSSL(q'):** Two-Stage Support Line using $q' := \frac{q}{1+q}$ at both stages, which guarantees bFDR $\leq q$ in the independent case.

3. **Storey(1/2):** Storey-adjusted Support Line procedure using fixed $\lambda = 0.5$.

4. **Storey(q)**: Storey-adjusted Support Line procedure using fixed $\lambda = q$.

5. **AS(0.1; q):** Adaptive-Storey Support Line procedure with threshold $\hat{\lambda}$ chosen as described in Section 3, with $\delta = 0.1$.

6. **AS(0.01; q):** Adaptive-Storey Support Line procedure using $\delta = 0.01$.

7. **AS(0.1; 0.5):** Adaptive-Storey Support Line procedure using $\delta = 0.1$ but using a grid that starts from 0.5 instead of q.

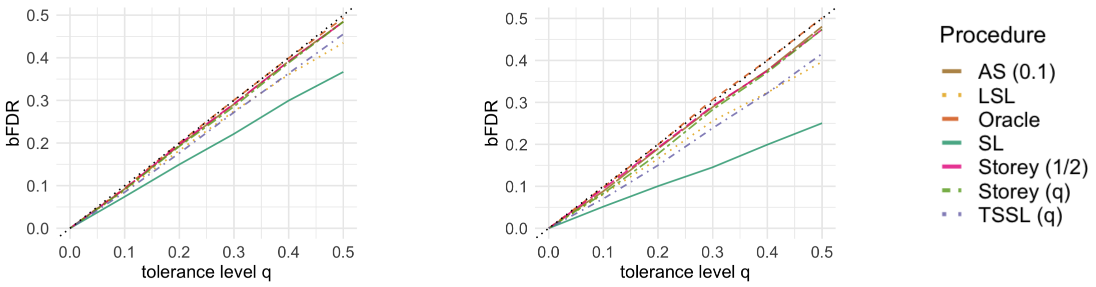
8. **LSL:** Support Line procedure run at level $q/\hat{\pi}_0$ using the Lowest Slope Estimator for π_0.

9. **SL:** Support Line procedure at level q (Benchmark).

10. **Oracle:** Support Line at level q/π_0 (Benchmark).

To estimate the boundary FDR for each procedure, we run $N = 10,000$ independent multiple testing experiments and record for each one the null status of the last rejection. The bFDR is estimated as the proportion of experiments in which the last rejection is a false discovery.

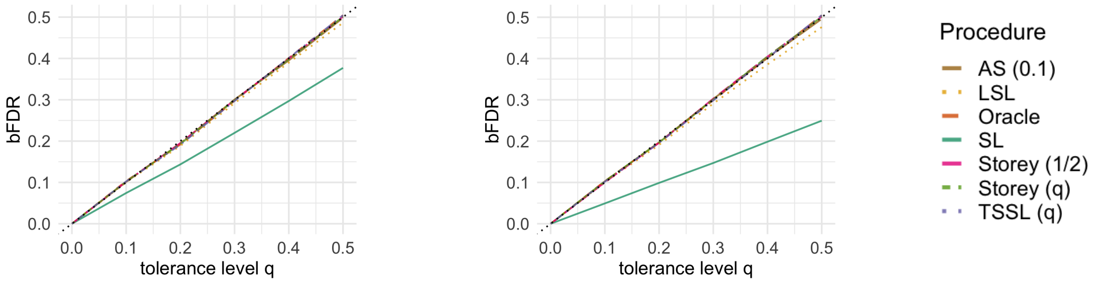
4.1 Independent p-values



(a) $\pi_0 = 0.75$

(b) $\pi_0 = 0.5$

Figure 2: Boundary FDR vs tuning parameter for the alternating configuration independent Gaussian test statistics, $N =$ 10,000 simulations, $m = 64$ p-values, with $\pi_0 = 0.75$ (left) and $\pi_0 = 0.5$ (right).



(a) $\pi_0 = 0.75$

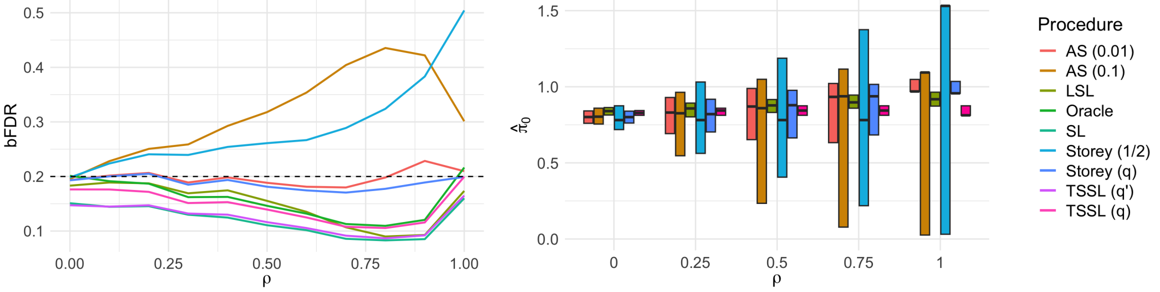
(b) $\pi_0 = 0.5$

Figure 3: Boundary FDR vs tuning parameter for the "all at 5" configuration with $N = 10{,}000$ simulations, $m = 64$ p-values, with $\pi_0 = 0.75$ (left) and $\pi_0 = 0.5$ (right).

In Figures 2 and 3, we plot the bFDR against the tolerance level q, for independent p-values at two different levels of π_0, in both the alternating and "all at 5" configurations. We can see that the bFDR for the adaptive procedures is much closer to the oracle bFDR and to q compared to the standard Support Line procedure. The improvement is more pronounced when the alternative distribution is more extreme (as in the "all at 5" configuration) and when $\hat{\pi}_0$ is smaller.

Figure 5, which shows the power of each procedure relative to the oracle, also demonstrates that the biggest gains in power are achieved when $\hat{\pi}_0$ is small. We also see that TSSL(q') does not always outperform the Support Line procedure, particularly when q is small and π_0 is close to 1.

4.2 Correlated p-values



(a) bFDR vs correlation ρ

(b) Second and third quartiles of $\hat{\pi}_0$ estimates for $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$

Figure 4: Numerical experiment: equicorrelated Gaussian test statistics with means generated according to the alternating configuration, $N = 10{,}000$ simulations, $m = 64$ p-values, $q = 0.2$, and $\pi_0 = 0.75$.

Figure 4a shows bFDR versus the dependence parameter ρ for the same numerical setting as in Section 4.1 but with positively equi-correlated Gaussian noise, where the equi-correlation parameter ρ varies between 0 and 1. Figure 4b suggests that the adaptive procedures' per-

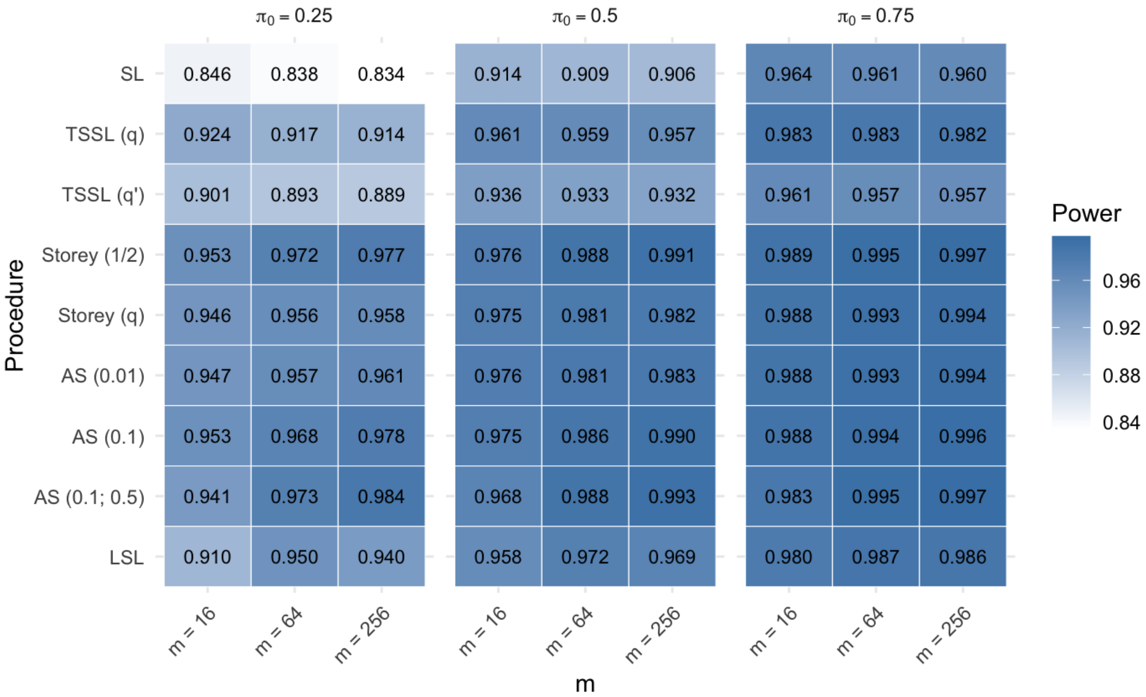


Figure 5: Power for each procedure relative to oracle SL, using the alternating configuration with $\pi_0 = 0.25, 0.5, 0.75$, $m = 16, 64, 256$, and $q = 0.2$. $N = 10,000$ simulations.

formance under strong positive dependence can be somewhat explained by the variability of their $\hat{\pi}_0$ estimates, since the procedures with the greatest bFDR violation also seem to have far higher variance in estimating π_0. Futhermore, while the choice of δ for the Adaptive Storey procedure did not seem to make a considerable difference for bFDR control in the independent case, we see a drastic difference in the correlated case. Hence, if one suspects there is strong positive dependence in their test statistics, it may not be wise to use the standard Storey-type adjustment (with $\lambda = 1/2$) for bFDR control. However, the Storey adjustment with fixed $\lambda = q = 0.2$ does not follow this trend, and performs surprisingly well in this setting.

4.3 Variability of lfdr estimates

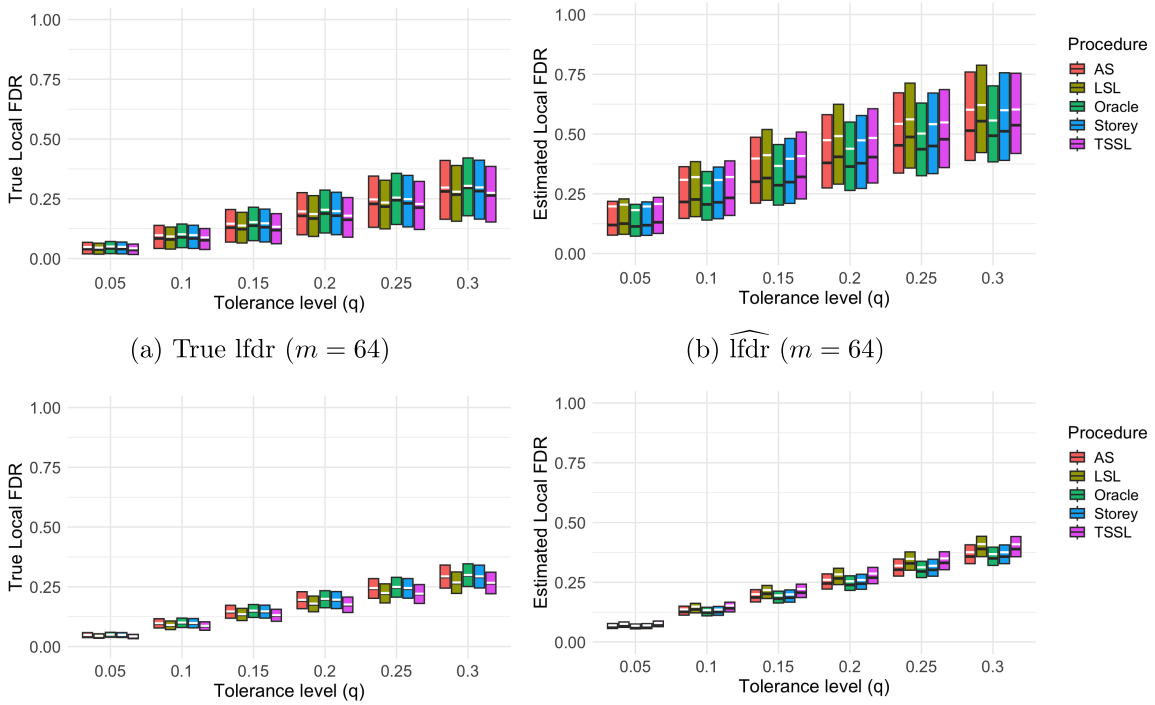
One concern with the Support Line procedure is that lfdr estimates could be quite variable – particularly for smaller values of m – since they rely on nonparametric density estimates. If there are too few p-values near the cutoff, the estimate $\hat{f}$ could have such high variance so as to not be useful in practice. To assess the extent to which this phenomenon occurs in our numerical settings, we examine box plots of $\hat{\pi}_0$ and true and estimated lfdr at the p-value cutoffs at various choices of q. Results are shown for setting with $m = 64$ and $m = 1024$ p-values in Figure 6. The true lfdr for Configuration 1 is given by

$$\mathrm{lfdr}(t) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\frac{\frac{1}{4}\sum_{j=1}^{4}\phi(\Phi^{-1}(1-t)-\frac{5j}{4})}{\phi(\Phi^{-1}(1-t))}},$$

where t is the cutoff p-value for each procedure. The estimated lfdr is given by $\widehat{\text{lfdr}}(p^*) = \hat{\pi}_0/\hat{f}(p^*)$, where $\hat{f}$ is the Grenander estimate used in the R package `fdrtool` (Strimmer, 2008) and p^* is the cutoff for the Oracle Support Line procedure, i.e.

$$p^* = \max\{t : \pi_0/\hat{f}(t) \leq q\}.$$

We chose to evaluate each of the estimates at the same cutoff p^* to control for the variation in different rejection thresholds among the procedures. Since the oracle threshold is the least conservative one among those considered here, it is positioned closest to the bulk of p-values where the estimation error due to the Grenander estimate is expected to be smaller. We see a small difference in the variance between the true and estimated lfdr values, and we notice that the estimated lfdr values are right-skewed and tend to be larger, which suggests the estimates are conservative.



(a) True lfdr ($m = 64$)

(b) $\widehat{\text{lfdr}}$ ($m = 64$)

(c) True lfdr ($m = 1024$)

(d) $\widehat{\text{lfdr}}$ ($m = 1024$)

Figure 6: Shown above are the interquartile ranges of the true lfdr, estimated lfdr, and $\hat{\pi}_0$ estimates for independent test statistics generated according to alternating configuration with $\pi_0 = 0.75$ and $N = 10{,}000$. Black lines show medians and white lines indicate means. Here TSSL is run at level q, Storey uses fixed $\lambda = 0.5$, and Adaptive Storey uses $\delta = 0.1$.

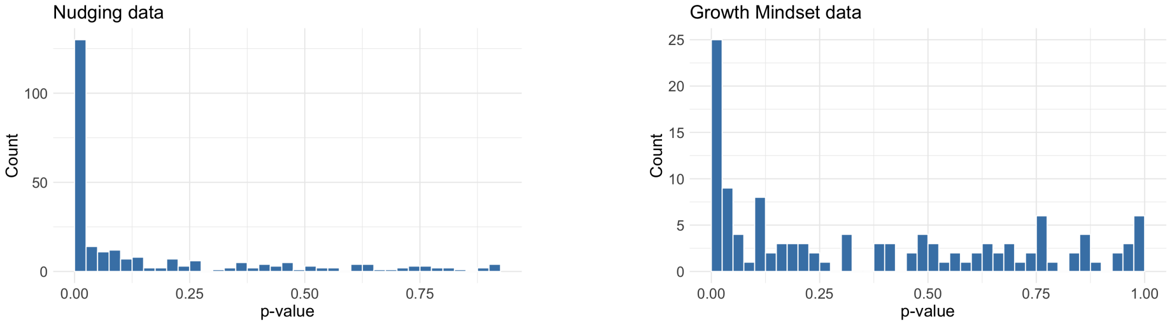
5 Meta-data analysis

5.1 Comparing the procedures on real datasets

As mentioned in the introduction, several areas in the psychology literature have encountered scrutiny due to replicability issues that can be traced back to the 'significance-threshold' standard of evidence (Benjamin et al., 2018). In this section, we use the adaptive procedures to estimate a range of p-value cut-off thresholds that correspond to bFDR tolerance for values q between 0 and 30%, for two meta-datasets of behavioral psychology experiments: one from growth mindset intervention literature (Tipton et al., 2023; Macnamara and Burgoyne, 2023) and another from psychological nudging (Thaler and Sunstein, 2009; Mertens et al., 2022).

Table 2 shows the number of rejections for each procedure for the nudging and growth mindset datasets, respectively. We see considerable gains in power for the adaptive procedures compared to the standard Support Line procedure, with the Storey procedures resulting in the highest number of rejections. We also see little difference between the various Storey and Adaptive Storey procedures. We note that for the growth-mindset data, TSSL(q') and SL have the same number of rejections, which aligns with the observation from Figure 5 that TSSL(q') does not demonstrate gains in power over SL when π_0 is close to 1.

Figure 7 displays histograms of both datasets and Table 2 lists the number of rejections for each procedure for both datasets. The nudging data has a higher proportion of near-zero p-values than the growth mindset data, which is consistent with the results in Tables 1 and 2; the nudging data has smaller $\hat{\pi}_0$ estimates and a higher percentage of rejections.



(a) Histogram of 261 p-values from nudging dataset

(b) Histogram of 122 p-values from growth mindset dataset

Figure 7: Histograms of p-values for nudging (Thaler and Sunstein, 2009; Mertens et al., 2022) and growth mindset (Tipton et al., 2023; Macnamara and Burgoyne, 2023) datasets

Table 1: $\hat{\pi}_0$ estimates for Nudging and Growth Mindset datasets

	Nudging			**Growth mindset**		
Procedure	$q = 0.1$	$q = 0.2$	$q = 0.3$	$q = 0.1$	$q = 0.2$	$q = 0.3$
TSSL(q)	0.62	0.56	0.51	0.93	0.85	0.78
TSSL(q')	0.62	0.56	0.56	0.93	0.85	0.78
Storey(1/2)	0.28	0.28	0.28	0.75	0.75	0.75
Storey(q)	0.40	0.36	0.33	0.77	0.70	0.73
LSL	0.39	0.39	0.39	0.77	0.77	0.77
AS(0.01)	0.38	0.35	0.33	0.72	0.70	0.72
AS(0.1, 0.5)	0.29	0.29	0.29	0.80	0.80	0.80
AS(0.1)	0.29	0.29	0.29	0.73	0.73	0.75

Table 2: Number of rejections for each procedure

	Nudging			**Growth mindset**		
Procedure	$q = 0.1$	$q = 0.2$	$q = 0.3$	$q = 0.1$	$q = 0.2$	$q = 0.3$
SL(q)	99 (38%)	115 (44%)	129 (49%)	9 (7%)	18 (15%)	27 (22%)
TSSL(q)	115 (44%)	129 (49%)	162 (62%)	12 (10%)	27 (22%)	34 (28%)
TSSL(q')	115 (44%)	129 (49%)	131 (50%)	9 (7%)	18 (15%)	27 (22%)
Storey(1/2)	129 (49%)	162 (62%)	182 (70%)	18 (15%)	27 (22%)	34 (28%)
Storey(q)	129 (49%)	162 (62%)	174 (67%)	18 (15%)	27 (22%)	34 (28%)
LSL	129 (49%)	162 (62%)	162 (62%)	18 (15%)	27 (22%)	34 (28%)
AS(0.01)	129 (49%)	162 (62%)	174 (67%)	18 (15%)	27 (22%)	34 (28%)
AS(0.1, 0.5)	129 (49%)	162 (62%)	182 (70%)	18 (15%)	27 (22%)	34 (28%)
AS(0.1)	129 (49%)	162 (62%)	182 (70%)	18 (15%)	27 (22%)	34 (28%)

5.2 Comparison with Sellke et al. (2012) calibration

Sellke et al. (2012) define a calibration function for p-values that can be interpreted as a lower bound on posterior probability of Type I error when the non-null p-values are drawn from a Beta(ξ, 1) distribution. Assuming equal prior odds for truth or falsehood of each null hypothesis, the local fdr for a p-value equal to t would be bounded below by

$$\alpha(t) = \frac{\log(1/t)}{e^{-1} + \log(1/t)}. \tag{3}$$

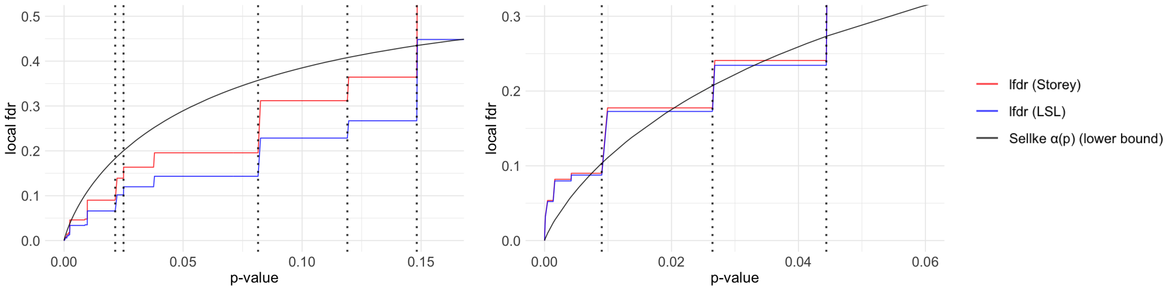
We compare this to the estimate of the local fdr implied by the two-stage procedures studied in this paper, defined in terms of the ratio $\hat{\pi}_0/\hat{f}$ where $\hat{f}$ is the Grenander estimate of a decreasing density (see Grenander (1956), Strimmer (2008)), and $\hat{\pi}_0$ is an estimate for the true null proportion.

Figure 8 shows plots of $\alpha(t)$ and $\widehat{\text{lfdr}}_{\hat{\pi}_0}(t)$ using Storey ($\lambda = 1/2$) and LSL estimates for $\hat{\pi}_0$. Vertical lines are drawn at the common cutoff values $q = 0.1, 0.15, 0.2, 0.25, 0.3$. For the growth mindset p-values, the local fdr estimates roughly agree with the calibration method of Sellke et al. (2012). For the nudge data, $\alpha(t)$ appears to be more conservative than $\widehat{\text{lfdr}}(t)$. This is likely because the $\hat{\pi}_0$ estimates given by the Storey and LSL procedures are smaller than 0.5 (see Table 1), which is the implicit value used in the derivation of (3).

6 Discussion

The adaptive procedures achieved substantial gains in power over the standard SL procedure in both simulations and on real data sets. In the next two paragraphs, we outline two directions of interest for future work.

Dependence. More work is needed to establish bFDR guarantees under dependence and to identify what form of positive dependence tends to preserve the finite-sample bFDR control for Support Line procedures. Figure 4a illustrates the importance of judging whether p-values are correlated when using the adaptive SL procedures on real datasets, since some procedures that have the highest power under independence also exhibited the poorest bFDR control



Psychological nudging p-values (selection-adjusted) (Mertens et al., 2022)

Growth mindset p-values (Macnamara and Burgoyne, 2023)

Figure 8: $\alpha_{\hat{\pi}_0}(p)$ and $\widehat{\text{lfdr}}_{\hat{\pi}_0}(p)$ with vertical lines at p-value cutoffs for $q \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.

under correlation. Future research should explore performance of adaptive SL procedures under a broader class of dependence structures beyond the equicorrelated setting considered in this paper. Additionally, since Adaptive Storey with the stopping rule described in 3 had drastically different bFDR under equicorrelation for different choices of δ, it would be useful to develop guidelines for choosing δ under both independence and dependence assumptions. Our experiments suggest the two-stage SL procedure and Storey ($\lambda = 0.2$) work well under positive dependence.

Tightness of bounds. Numerically, we have found that using *some* estimator for π_0 is typically better than using none, in that the power gain can be substantial while provably maintaining boundary FDR control under independence. However, the theoretical guarantees on TSSL and Storey SL in 3.1 and 3.2 are both inequalities, which allows for the possibility of matching or even under-performing relative to the standard SL procedure, particularly if $\hat{\pi}_0$ is close to 1. We see examples of this in Figure 5 and Table 2. An avenue for future research is to investigate conditions under which the bFDR bounds become tight to clarify when adaptive procedures would provide a definitive theoretical advantage over the unadjusted SL procedure.

Acknowledgements

We thank Yoav Benjamini, William Fithian, Zijun Gao, Etienne Roquain and Jake Soloff for helpful discussions that inspired this work.

7 Proofs

Theorem 3.1. Let $H_1, \dots, H_m$ denote m null hypotheses, with independent p-values $p_1, \dots, p_m$. Suppose that $p_i \sim \text{Uniform}(0,1)$ if H_i is true. Then

$$\text{bFDR}(\mathcal{R}_2) \leq \frac{q}{1-q},$$

where $\mathcal{R}_2 \coloneqq \{i : p_i \leq p_{(R_2)}\}$ is the rejection set for the two stage procedure.

Proof. We decompose the bFDR into two pieces:

$$\mathbb{P}(H_{(R_2)} \text{ is true}, R_2 > 0) = \mathbb{P}(H_{(R_2)} \text{ is true}, R_1 = R_2) + \mathbb{P}(H_{(R_2)} \text{ is true}, R_1 < R_2), \tag{4}$$

with the convention that $H_{(0)} \coloneqq \text{false}$. For the first piece in (4), note by exchangeability

$$\mathbb{P}(H_{(R_2)} \text{ is true}, R_1 = R_2) = m_0 \mathbb{P}(p_{(R_1)} = p_m, R_1 = R_2),$$

where we assumed wlog that H_m is true. Let $q_{(1)} \leq \cdots \leq q_{(m-1)}$ denote the order statistics of $p_1, \dots, p_{m-1}$, and define

$$\Delta_k \coloneqq \frac{qk}{m} - q_{(k)}, \qquad k = 0, \dots, m,$$

where $q_{(m)} \coloneqq 1$ and $q_{(0)} \coloneqq 0$. On the event $\{p_{(R_1)} = p_m\}$, we have for some $k \in \{0, \dots, m-1\}$ that $q_{(k)} < p_m < q_{(k+1)}$ and

$$\frac{q(k+1)}{m} - p_m \geq \big(\max_{j \leq k} \Delta_j\big) \vee \Big(\max_{j > k} \Delta_j + \frac{q}{m}\Big).$$

When non-empty, the intersection of these two constraints is an interval I_k, of the form:

$$I_k \coloneqq \Big(q_{(k)},\ qk/m - \big(\max_{j > k} \Delta_j\big) \vee \big(\max_{j \leq k} \Delta_j - q/m\big)\Big), \tag{5}$$

where we define $I_k \coloneqq \varnothing$ if the right endpoint of (5) is smaller than the left endpoint. It follows from disjointness of $I_0, \dots, I_{m-1}$ that

$$\mathbb{P}(p_{(R_2)} = p_m, R_1 = R_2) \leq \mathbb{P}(\cup_{k=0}^{m-1} \{p_m \in I_k\}) = \sum_{k=0}^{m-1} \mathbb{P}(p_m \in I_k). \tag{6}$$

Next, we consider the case $R_1 < R_2$, which in particular implies $R_1 < m$, and define:

$$\Delta'_k \coloneqq \frac{qk}{m - R_1(1)} - q_{(k)}, \qquad k = 0, \dots, m,$$

where $R_1(p_m)$ denotes the value of R_1 for a given value of p_m, holding $p_1, \dots, p_{m-1}$ fixed. Note that on the event $\{R_1 < R_2\} \cap \{p_{(R_2)} = p_m\}$ we have by Lemma 7.1 that $R_1(p_m) = R_1(1)$.

For the second piece in (4), it follows that on the event $\{R_1 < R_2\} \cap \{p_{(R_2)} = p_m\}$, we have for some k that $q_{(k)} < p_m < q_{(k+1)}$ and

$$\begin{aligned}
\frac{q(k+1)}{m - R_1(1)} - p_m &\geq \big(\max_{j \leq k} \Delta'_j\big) \vee \Big(\max_{j>k} \Delta'_j + \frac{q}{m - R_1(1)}\Big) \\
\frac{q(k+1)}{m} - p_m &< \big(\max_{j \leq k} \Delta_j\big) \vee \Big(\max_{j>k} \Delta_j + \frac{q}{m}\Big),
\end{aligned}$$

which rearrange to yield the following constraints: for some k,

$$\begin{aligned}
q_{(k)} < p_m < q_{(k+1)} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (7)\\
p_m \leq \frac{qk}{m - R_1(1)} - \big(\max_{j>k} \Delta'_j\big) \vee \Big(\max_{j \leq k} \Delta'_j - \frac{q}{m - R_1(1)}\Big) \\
p_m > \frac{qk}{m} - \big(\max_{j>k} \Delta_j\big) \vee \big(\max_{j \leq k} \Delta_j - q/m\big). \qquad\qquad\qquad\qquad\qquad\qquad\qquad (8)
\end{aligned}$$

For each k, we define the analogous interval to I_k as

$$I'_k := \Big(q_{(k)},\ \frac{qk}{m - R_1(1)} - \big(\max_{j>k} \Delta'_j\big) \vee \Big(\max_{j \leq k} \Delta'_j - \frac{q}{m - R_1(1)}\Big)\Big),$$

where $I'_k := \varnothing$ if the right endpoint of the above is smaller than the left endpoint. The two lower constraints (7) and (8) for the value of p_m prompt us to consider two cases.

Case 1: $q_{(k)} > \frac{qk}{m} - \big(\max_{j>k} \Delta_j\big) \vee \big(\max_{j \leq k} \Delta_j - q/m\big)$.

In this case, I_k as defined in (5) is empty, and I'_k is the set over which p_m achieves the maximum as the $(k+1)^{\text{th}}$ order statistic, which has length:

$$\Delta'_k - \big(\max_{j>k} \Delta'_j\big) \vee \Big(\max_{j \leq k} \Delta'_j - \frac{q}{m - R_1(1)}\Big).$$

Case 2: $q_{(k)} \leq \frac{qk}{m} - \big(\max_{j>k} \Delta_j\big) \vee \big(\max_{j \leq k} \Delta_j - q/m\big)$.

In this case, I_k is non-empty and the length of the set over which p_m achieves the maximum as the $(k+1)^{\text{th}}$ order statistic is:

$$\frac{qk}{m - R_1(1)} - \big(\max_{j>k} \Delta'_j\big) \vee \Big(\max_{j \leq k} \Delta'_j - \frac{q}{m - R_1(1)}\Big) - \Big[\frac{qk}{m} - \big(\max_{j>k} \Delta_j\big) \vee \big(\max_{j \leq k} \Delta_j - q/m\big)\Big].$$

Adding and subtracting $q_{(k)}$, we re-express the above length as:

$$\underbrace{\Delta'_k - \big(\max_{j>k} \Delta'_j\big) \vee \Big(\max_{j \leq k} \Delta'_j - \frac{q}{m - R_1(1)}\Big)}_{\mathbb{P}(p_m \in I'_k | p_{-m})} - \underbrace{\big[\Delta_k - \big(\max_{j>k} \Delta_j\big) \vee \big(\max_{j \leq k} \Delta_j - q/m\big)\big]}_{\mathbb{P}(p_m \in I_k | p_{-m})},$$

since $p_m \mid p_{-m} \sim \text{Uniform}(0,1)$, where $p_{-m} \coloneqq (p_1, \dots, p_{m-1})$. In summary, we have shown

$$\mathbb{P}(p_{(R_2)} = p_m, R_1 = R_2 \mid p_{-m}) \leq \sum_{k=0}^{m-1} \mathbb{P}(p_m \in I_k \mid p_{-m})$$
$$\mathbb{P}(p_{(R_2)} = p_m, R_1 < R_2 \mid p_{-m}) \leq \sum_{k=0}^{m-1} \mathbb{P}(p_m \in I'_k \mid p_{-m}) - \mathbb{P}(p_m \in I_k \mid p_{-m}).$$

The sum is therefore bounded by

$$\begin{aligned}\mathbb{P}(p_{(R_2)} = p_m \mid p_{-m}) &= \mathbb{P}(p_{(R_2)} = p_m, R_1 = R_2 \mid p_{-m}) + \mathbb{P}(p_{(R_2)} = p_m, R_1 < R_2 \mid p_{-m}) \\ &\leq \sum_{k=0}^{m-1} (\mathbb{P}(p_m \in I'_k \mid p_{-m}) - \mathbb{P}(p_m \in I_k \mid p_{-m})) + \mathbb{P}(p_m \in I_k \mid p_{-m}) \\ &= \sum_{k=0}^{m-1} \Delta'_k - \big(\max_{j>k} \Delta'_j\big) \vee \Big(\max_{j \leq k} \Delta'_j - \frac{q}{m - R_1(1)}\Big),\end{aligned}$$

which is telescoping and equal to $\frac{q}{m-R_1(1)}$. Taking expectation on both sides, we have shown

$$\mathbb{P}(H_{(R_2)} \text{ is true}, R_2 > 0) \leq \mathbb{E}\Big[\frac{m_0 q}{m - R_1(1)}\Big].$$

We show in Lemma 7.2 that the right hand side of the above is $\leq \frac{q}{1-q}$, completing the proof. □

Theorem 3.2. Let $H_1, \dots, H_m$ be null hypotheses, and suppose that for any i where H_i is true, we have that $p_i \sim \text{Uniform}(0,1)$ and that p_i is independent of $p_{-i} \coloneqq (p_j : j \in [m]\backslash\{i\})$. Then

$$\text{bFDR}(\mathcal{R}_{q/\hat{\pi}_0}) \leq q.$$

Proof. Suppose wlog H_m is true. By exchangeability of the null p-values, and the tower property, we have

$$\text{bFDR} = m_0 \mathbb{P}(p_{(R_{q/\hat{\pi}_0})} = p_m) = m_0 \mathbb{E}\big[\mathbb{P}(p_{(R_{q/\hat{\pi}_0})} = p_m \mid p_{-m}, 1\{p_m \leq q\})\big]. \tag{9}$$

Define $p'_i \coloneqq p_i/q$, $m' \coloneqq \#\{i : p_i \leq q\}$, and $q' \coloneqq \frac{m'}{m\hat{\pi}_0}$. Then

$$\begin{aligned}R_{q/\hat{\pi}_0} &\coloneqq \text{argmin}_{k:p_{(k)} \leq q}\Big\{p_{(k)} - \frac{qk}{\hat{\pi}_0 m}\Big\} \\ &= \text{argmin}_{k:p_{(k)} \leq q}\Big\{\frac{p_{(k)}}{q} - \frac{k}{\hat{\pi}_0 m}\Big\} = \text{argmin}_{k=0,1,\dots,m'}\Big\{p'_{(k)} - \frac{q'k}{m'}\Big\}.\end{aligned}$$

Notice $\{p_{(R_{q/\hat{\pi}_0})} = p_m\}$ implies $\{p_m \leq q\}$. It follows from Lemma 2 of Soloff et al. (2024) that

$$\mathbb{P}(p_{(R_{q/\hat{\pi}_0})} = p_m \mid p_{-m}, 1\{p_m \leq q\}) \leq \frac{q'\, 1\{p_m \leq q\}}{m'} = \frac{1\{p_m \leq q\}}{m\hat{\pi}_0} = \frac{1\{p_m \leq q\}}{(m-1)\hat{\pi}_0(p_{-m})}$$

since $\hat\lambda \geq q$ implies $\hat\pi_0(p_1, \dots, p_m) = \frac{m-1}{m}\, \hat\pi_0(p_1, \dots, p_{m-1})$ on the event $\{p_m \leq q\}$. Taking expectation on both sides while retaining the conditioning on p_{-m}, and noting that $p_m \sim \text{Uniform}(0, 1)$ independently of p_{-m}, we have

$$\mathbb{P}(p_{(R_{q/\hat\pi_0})} = p_m \mid p_{-m}) \leq \frac{1}{(m-1)\hat\pi_0(p_{-m})}\, \mathbb{E}\big[1\{p_m \leq q\} \mid p_{-m}\big] = \frac{q}{(m-1)\hat\pi_0(p_{-m})}.$$

We were able to pull out $1/\hat\pi_0(p_{-m})$ from the expectation because $\hat\lambda$ is a function of p_{-m} on the event $\{p_m \leq q\}$, and thus $\hat\pi_0$ is determined by p_{-m}. Next, notice that

$$\begin{aligned}\frac{q}{(m-1)}\, \mathbb{E}\Big[\frac{1}{\hat\pi_0(p_{-m})}\Big] &= \frac{q}{(m-1)}\, \mathbb{E}\Big[\frac{(m-1)(1-\hat\lambda)}{1 + \#\{i < m : p_i > \hat\lambda\}}\Big] \\ &\leq q\, \mathbb{E}\Big[\frac{1-\hat\lambda}{1 + \#\{i < m : H_i \text{ is true},\ p_i > \hat\lambda\}}\Big].\end{aligned}$$

Since $\hat\lambda \in [q, 1]$ is a stopping time, and the expression inside the expectation is a supermartingale (see proof of Theorem 1 in Gao (2025)) the OST implies that the above is bounded by

$$\leq q\, \mathbb{E}\Big[\frac{1-q}{1 + \#\{i < m : H_i \text{ is true},\ p_i > q\}}\Big] \leq q\, \frac{1-q}{(1 + (m_0 - 1))(1-q)} \tag{10}$$

following from standard Binomial calculations (see, e.g. the proof of Theorem 1 in Gao (2025)). Together, (9) and (10) imply

$$\text{bFDR} = m_0 \mathbb{P}(p_{(R_{q/\hat\pi_0})} = p_m) \leq m_0\, \frac{q(1-q)}{m_0(1-q)} = q,$$

completing the proof. □

7.1 Technical Lemmas

Lemma 7.1. *Let $P_{(-m)}(t) \in [0, 1]^m$ denote the vector of p-values obtained by replacing p_m with t, and let $R_1(t) : [0, 1]^m \to \{0, ..., m\}$ denote the number of rejections made in stage one of the TSSL procedure run on the vector $P_{(-m)}(t)$. If $p_m > p_{(R_1)}$, then $R_1(p_m) = R_1(1)$.*

Proof. Suppose p_m is the $(k+1)^{\text{th}}$ order statistic and consider what happens to the number of rejections in Stage 1 if we replace p_m with 1. By assumption $R_1(p_m) \leq k$, and since moving p_m to 1 shifts the ranks of $p_{(k+2)}, \dots, p_{(m)}$ down by 1, we have

$$\Big[\max_{j \in \{0,...,k\} \setminus \{R_1(p_m)\}} \Delta_j\Big] \vee \Big[\max_{j > k+1} \Delta_j - \frac{q}{m}\Big] \leq \max_{j \in \{0,...,m\} \setminus \{R_1(p_m)\}} \Delta_j < \Delta_{R_1(p_m)},$$

which implies $R_1(p_m) = R_1(1)$. □

Lemma 7.2. *Suppose $(p_i)_{i \in \mathcal{H}_0}$ are independent and $p_i \sim$ Uniform$(0, 1)$ when H_i is true. Then*

$$\mathbb{E}\left[\frac{qm_0}{m - R_1(1)}\right] \leq \frac{q}{1-q}$$

Proof. Suppose we run TSSL on $P_{(-m)}(1)$. Define the sets V and U as follows:

$$V := \#\{i \le m-1 : H_i \text{ is true, } p_i \le p_{(R_1)}\}$$
$$U := \#\{i \le m-1 : H_i \text{ is true, } p_i > p_{(R_1)}\}.$$

We have by definition

$$(m_0 - 1) + (R_1(1) - V) \le m$$

Because $m_0 = U + V + 1$, we can write

$$U + 1 \le m - R_1(1)$$

By definition, $p_{(R_1)} \le q$ which implies

$$U \ge \#\{i \le m-1 : H_i \text{ is true, } p_i > q\}$$

Let $Y \sim \text{Binomial}(m_0 - 1, 1-q)$. We can see that Y must be stochastically smaller than U, implying that $m - R_1(1)$ is stochastically larger than $Y + 1$. Thus, by Lemma 1 from Benjamini et al. (2006) we have $\mathbb{E}\left[\frac{1}{Y+1}\right] \le \frac{1}{m_0(1-q)}$. Thus, we have

$$\mathbb{E}\left[\frac{qm_0}{m - R_1(1)}\right] = qm_0\mathbb{E}\left[\frac{1}{m - R_1(1)}\right] \le \frac{q}{1-q}.$$

□

7.2 Asymptotic formula for bFDR

Theorem 7.1. *Let $F^{(1)}, F^{(2)}, \ldots$ be a sequence of continuous cdfs, where $p_i \sim F^{(i)}$, and suppose that $\|F_m - \bar{F}\|_\infty \to 0$ in probability as $m \to \infty$, where $F_m(t) := \frac{1}{m}\sum_{i=1}^m 1\{p_i \le t\}$ is the empirical cdf and $\bar{F}_m := \frac{1}{m}\sum_{i=1}^m F^{(i)}$ is the average among the first m cdfs. Further suppose that $\bar{F}_m$ is concave for each m and that there exist unique values $t_1^*, t_2^* \in (0,1)$ such that*

$$t_1^* = \text{argmin}_{t\in[0,1]}\{t - q\bar{F}_m(t)\}, \qquad \bar{F}_m(t_1^*) < 1,$$
$$t_2^* = \text{argmin}_{t\in[0,1]}\left\{t - \frac{q\bar{F}_m(t)}{1 - \bar{F}_m(t_1^*)}\right\}.$$

Let $\tau_1 := \text{argmin}_{t\in[0,1]}\{t - qF_m(t)\}$ and $\tau_2 := \text{argmin}_{t\in[0,1]}\left\{t - \frac{qF_m(t)}{1-F_m(\tau_1)}\right\}$ be the empirical counterparts. Then

$$\lim_{m\to\infty}\left|\text{lfdr}_m(\tau_2) - \frac{q\bar{\pi}_{0,m}}{1 - \bar{F}_m(t_1^*)}\right| = 0,$$

where $\text{lfdr}_m(t) := \bar{\pi}_{0,m}/\bar{f}_m(t)$, $\bar{f}_m(t) := \frac{1}{m}\sum_{i=1}^m f^{(i)}(t)$, and $\bar{\pi}_{0,m} := \#\{i : H_i = 0\}/m$.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al. (2018). Redefine statistical significance, *Nature human behaviour* **2**(1): 6–10.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* **57**(1): 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of educational and Behavioral Statistics* **25**(1): 60–83.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate, *Biometrika* **93**(3): 491–507.

Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence., *Journal of Machine Learning Research* **10**(12).

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science, *Science* **349**(6251): aac4716.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment, *Journal of the American statistical association* **96**(456): 1151–1160.

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type i errors, *Biometrical Journal* **43**(8): 985–1005.

Gao, Z. (2025). An adaptive null proportion estimator for false discovery rate control, *Biometrika* **112**(1): asae051.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**(3): 499–517.

Grenander, U. (1956). On the theory of mortality measurement: part ii, *Scandinavian Actuarial Journal* **1956**(2): 125–153.

Macnamara, B. N. and Burgoyne, A. P. (2023). Do growth mindset interventions impact students' academic achievement? a systematic review and meta-analysis with recommendations for best practices., *Psychological bulletin* **149**(3-4): 133.

Mertens, S., Herberz, M., Hahnel, U. J. and Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains, *Proceedings of the National Academy of Sciences* **119**(1): e2107346118.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics* **19**(3): 368–375.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science* **273**(5281): 1516–1517.

Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously, *Biometrika* **69**(3): 493–502.

Sellke, T., M.J., B. and O Berger, J. (2012). Calibration of p values for testing precise null hypotheses, *The American Statistician* **55**(1): 62–71.

Soloff, J. A., Xiang, D. and Fithian, W. (2024). The edge of discovery: Controlling the local false discovery rate at the margin, *The Annals of Statistics* **52**(2): 580–601.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**(3): 479–498.

Strimmer, K. (2008). fdrtool: a versatile r package for estimating local and tail area-based false discovery rates, *Bioinformatics* **24**(12): 1461–1462.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*, Penguin.

Tipton, E., Bryan, C., Murray, J., McDaniel, M. A., Schneider, B. and Yeager, D. S. (2023). Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity: Commentary on macnamara and burgoyne (2023) and burnette et al.(2023).

Williams, V. S., Jones, L. V. and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement, *Journal of educational and behavioral statistics* **24**(1): 42–69.

Xiang, D., Soloff, J. A. and Fithian, W. (2025). A frequentist local false discovery rate, *arXiv preprint arXiv:2502.16005* .