Bernoulli **30**(1), 2024, 743–769
https://doi.org/10.3150/23-BEJ1615

A central limit theorem for the Benjamini-Hochberg false discovery proportion under a factor model

DAN M. KLUGER[a] and ART B. OWEN[b]

Department of Statistics, Stanford University, Stanford, CA, USA, [a]*kluger@stanford.edu,* [b]*owen@stanford.edu*

The Benjamini-Hochberg (BH) procedure remains widely popular despite having limited theoretical guarantees in the commonly encountered scenario of correlated test statistics. Of particular concern is the possibility that the method could exhibit bursty behavior, meaning that it might typically yield no false discoveries while occasionally yielding both a large number of false discoveries and a false discovery proportion (FDP) that far exceeds its own well controlled mean. In this paper, we investigate which test statistic correlation structures lead to bursty behavior and which ones lead to well controlled FDPs. To this end, we develop a central limit theorem for the FDP in a multiple testing setup where the test statistic correlations can be either short-range or long-range as well as either weak or strong. The theorem and our simulations from a data-driven factor model suggest that the BH procedure exhibits severe burstiness when the test statistics have many strong, long-range correlations, but does not otherwise.

Keywords: Empirical cumulative distribution function; functional central limit theorem; functional delta method; multiple hypothesis testing; Simes line

1. Introduction

The Benjamini-Hochberg (BH) procedure is a widely used method for balancing Type I and Type II errors when testing many hypotheses simultaneously. The procedure is designed to control the False Discovery Rate (FDR), which is the *expected value* of the proportion of discoveries that are false (FDP), below a user specified threshold (Benjamini and Hochberg (1995)). The procedure was originally shown to guarantee FDR control when the test statistics are assumed to be independent, an assumption unlikely to hold in most application settings. The BH procedure was later proven in Benjamini and Yekutieli (2001) to control the FDR when there are dependent test statistics satisfying the Positive Regression Dependency (PRDS) property. While PRDS is quite restrictive (for example, it does not hold for two-sided hypothesis tests when the test statistics are correlated or when there are negatively correlated test statistics (Fithian and Lei (2022))), under more general conditions simulation studies have found BH to conservatively control the FDR (Farcomeni (2006), Kim and van de Wiel (2008)).

While FDR control is important, the motivation for this paper is our concern that FDR control alone can give investigators who use BH false confidence in a low prevalence of false discoveries among their rejected hypothesis. This can happen if the distribution of the FDP has both a wide right tail and a mean that is still below the user specified threshold. As an example, it would be worrisome in the plausible scenario that an investigator is led to believe that roughly 10 percent of their discoveries are false, when in fact, a majority of them are false. To address such concerns, a number of multiple testing procedures have been proposed to control the tail probability that the FDP exceeds a user specified threshold (Korn et al. (2004), Romano and Shaikh (2006), Romano and Wolf (2007)). Efron (2007) also raised concerns about high variability of FDP due to correlations of the test statistics and proposed an empirical Bayes approach for estimating a dispersion parameter of the test statistics and controlling

FDR conditionally on the dispersion parameter. Despite the promise of these methods, the BH procedure remains overwhelmingly popular and the default method of choice for investigators with multiple testing problems. It is therefore important to determine under which conditions are we assured that the distribution of the FDP will be well concentrated about its mean, the FDR, and under which conditions there is a risk that the distribution of the FDP has a wide right tail. Throughout this text we will refer to the former scenario with low variability of the FDP about its mean as the non-bursty regime, and the alarming, latter scenario where occasionally the FDP is much larger than expected as the bursty regime.

While previous simulations in the literature can be used to identify some settings where the BH procedure will exhibit burstiness (see for example, Figure 5 of Friguet, Kloareg and Causeur (2009) or Figure 1 of Delattre and Roquain (2015)), the aim of this paper is to gain a theoretical understanding of when burstiness is a concern for BH. We identify dependency structures among test statistics that in conjunction with certain proportions of nonnulls make BH prone to delivering bursts of false discoveries. We find other settings where such bursts must be rare. Our results are asymptotic and hold in a two-group mixture model previously studied by Genovese and Wasserman (2004), Delattre and Roquain (2016) and Izmirlian (2020). In that model, independent Bern(π_1) variables define which hypotheses are nonnull, the null p-values have the Unif$(0,1)$ distribution and the nonnull p-values have some other distribution in common.

The BH procedure and the asymptotic distribution of the FDP is well studied for the setting where the test statistics are independent. Finner and Roters (2001, 2002) study properties of the number of false discoveries under independence both when there are no nonnulls and when the nonnull p-values are always 0. The limiting distribution of the FDP was studied by Genovese and Wasserman (2004) under independence of the test statistics; however, their asymptotic FDP results are derived for the "plug-in" method (Benjamini and Hochberg (2000)) rather than the standard BH procedure. To our knowledge, a CLT for the FDP of the BH procedure itself was first explicitly stated in Neuvial (2008), which uses a functional delta method argument. Izmirlian (2020), using a CLT for a randomly stopped process, provides a simpler proof of a CLT for the FDP and corrects an error in Neuvial's asymptotic variance formula. These works show that in the two group mixture model with Bernoulli parameter π_1 and with m hypotheses to test, when using the BH procedure at FDR control level q, $\sqrt{m}(\mathrm{FDP} - (1-\pi_1)q)$ converges as $m \to \infty$ to a centered Gaussian with variance that depends on q, π_1, and the common nonnull p-value distribution.

There are fewer results on the limiting distribution of the FDP for dependent test statistics. Farcomeni (2007) derives a CLT for the FDP of the plug-in procedure when the p-values are stationary and satisfy some mixing conditions but for brevity omits explicitly stated FDP CLTs for the standard BH procedure. Using the proof methodology of Neuvial (2008), Delattre and Roquain (2011) derive a CLT for the FDP of the BH procedure for one sided testing, when the test statistics follow an equicorrelated Gaussian model, with correlation parameter $\rho \to 0$ as the number of tests $m \to \infty$. Delattre and Roquain (2016) extend this result to settings where the Gaussian test statistics follow arbitrary dependence structures but the average pairwise correlation of the test statistics, and the average 2nd and 4th powers of the pairwise correlations of the test statistics satisfy some constraints.

In Delattre and Roquain (2016), CLTs for the FDP are derived under two distinct regimes. In their first regime, the average pairwise correlation among test statistics is strictly greater than $O(1/m)$ for m tests. Under this regime, the FDP is not $\sqrt{m}$-consistent for the product of the FDR control parameter with the limiting proportion of nulls, and the FDP only converges to a Gaussian with scale factors much smaller than $\sqrt{m}$. Their other regime considered has an average correlation among test statistics that is at most $O(1/m)$. For this regime Delattre and Roquain (2016) derive a CLT for the FDP with $\sqrt{m}$ scaling, but they require a restrictive assumption which they call "vanishing-second order", precluding settings where there are short-range correlations of constant order. Examples of test statistic correlation

matrices to which Delattre and Roquain (2016) will not apply include tridiagonal Toeplitz correlation matrices as well as block correlation matrices of constant block size, both of which are simple models of interest for studying multiple testing under dependence.

With the aim of identifying dependency structures for which the investigator should be concerned about bursty behavior of the BH procedure, in this paper, we introduce a model that allows for a combination of long-range and potentially strong dependence among the test statistics via a factor model, along with additional strongly-mixing noise that has rapidly decaying long-range dependence. The model also allows for the proportion of nonnulls among the m hypothesis tests to vary as a function of m. Under some regularity conditions on the factor model and on the noise with rapidly decaying long-range dependence, we prove a CLT for the FDP under more general conditions than prior CLTs. We also establish a CLT for the False Positive Ratio (FPR), which is the proportion of false discoveries among all tests conducted. The new CLTs hold conditionally on the realized latent variable of the factor model. Applying these new results, we make the following contributions to the literature of asymptotic results for the BH procedure:

1. CLTs of the FDP for simple models, with short-range and constant-order dependency structures, that were not covered by the results of Delattre and Roquain (2016). Examples include block correlation structures with fixed block size and banded correlation structures. These results allow for non-stationary test statistics, so they cannot be inferred from theorems in Farcomeni (2007) either.
2. Conditional CLTs in settings where the long-range dependency is modeled by a factor model which includes scenarios not covered in either Farcomeni (2007) or Delattre and Roquain (2016).
3. CLTs for the FPR rather than just the FDP because the FDP limiting behavior is unilluminating when it converges in probability to 1.
4. CLTs where the expected proportion of nonnulls varies as the number of test statistics grows, allowing for a sparse allocation of nonnulls.
5. A discussion of the dependency regimes under which the investigator should be concerned about BH having bursty behavior, such as the setting where the number of nonnulls is $o_p(\sqrt{m})$, and the dependency structure contains a factor model component.

To qualify point 2 above, we note that Delattre and Roquain (2016) include some CLTs for the FDP under long-range dependency that our theorems do not. Ours all have a $\sqrt{m}$ scaling. They include some with a slower than $\sqrt{m}$ scaling. For instance, they get such a CLT under an equicorrelated Gaussian model with correlation $\rho \to 0$ but $\sqrt{m}\rho \to \infty$. To clarify point 5 above, we characterize what causes alarming burstiness of the BH FDP when the test statistics are dependent, which to our knowledge has not been analyzed theoretically before. A separate issue is the pathologically low power of BH when the FDR control level q is below a critical threshold (Chi, 2007). This issue was studied in greater generality by Zhang, Fan and Yu (2011) using the framework of Storey, Taylor and Siegmund (2004).

The proof of our main theorem builds upon the proof structure seen in Neuvial (2008) and Delattre and Roquain (2016). As is done in those works, we derive a functional CLT (FCLT) for the empirical cumulative distribution functions (ECDFs) of the null and nonnull p-values, compute the Hadamard derivative of the FDP written as a functional of the two ECDFs, and apply the functional delta method to obtain a CLT for the FDP. While the proofs of previous FDP CLTs in the literature require establishing an FCLT for the p-value ECDFs defined on $[0,1]$, in Section 2.3.4, we define a focal interval $[a,b] \subset (0,1)$, and our proof demonstrates that merely an FCLT for the ECDFs restricted to $[a,b]$ is needed. Our use of a focal interval allows us to obtain a CLT for the FDP in new settings where the null and nonnull p-value ECDFs are poorly behaved asymptotically in either $[0,a)$ or $(b,1]$. To obtain an FCLT when restricting our attention to $[a,b]$, we use an FCLT from Andrews and Pollard (1994) for bounded function classes. The regularity conditions in Andrews and Pollard (1994) are conducive to

obtaining an FCLT of the p-value ECDFs in settings where the test statistics follow block or banded correlation structures (see point 1 above). Therefore, in addition to the contributions enumerated above, our use of a focal interval and our use of an FCLT from Andrews and Pollard (1994) are contributions to proof methodology for BH asymptotics.

Our results suggest that approaches which estimate and remove the factor model components from the test statistics prior to applying BH can alleviate burstiness issues. A number of such approaches for estimating and removing factor model components in multiple testing settings have been proposed and have shown promise in simulations (Fan, Han and Gu, 2012, Fan et al., 2019, Friguet, Kloareg and Causeur, 2009, Sun, Zhang and Owen, 2012, Wang et al., 2017). Our CLT for the BH method itself can be a useful step towards deriving CLTs for methods that first estimate and remove factor model components and subsequently apply BH.

Figure 1 shows simulations of the FDP under some models that we study in this paper. In each case, there are 25,000 Monte Carlo simulations. The BH procedure is used with $q = 0.1$ on test statistics that are $\mathcal{N}(0,1)$ for null hypotheses and $\mathcal{N}(2,1)$ for alternative hypotheses. Each hypothesis is independently null with probability 0.9 and nonnull otherwise. The models differ in the correlation among test statistics. For the first histogram, $m = 22{,}283$ test statistics were sampled with correlations based on a 3-factor model fit to some Duchenne Muscular Dystrophy data described in Section 6. The second histogram is for the same correlation matrix after dividing the off-diagonal entries by 10. Next are two block correlation models with blocks of size 100 and within-block correlations of 0.5 or 0.05. To keep $m = 22{,}283$ one of the blocks had only 83 test statistics in it. In all four settings the FDR is seen to be controlled below 0.1, as desired. The positive False Discovery Rate (pFDR), defined as the

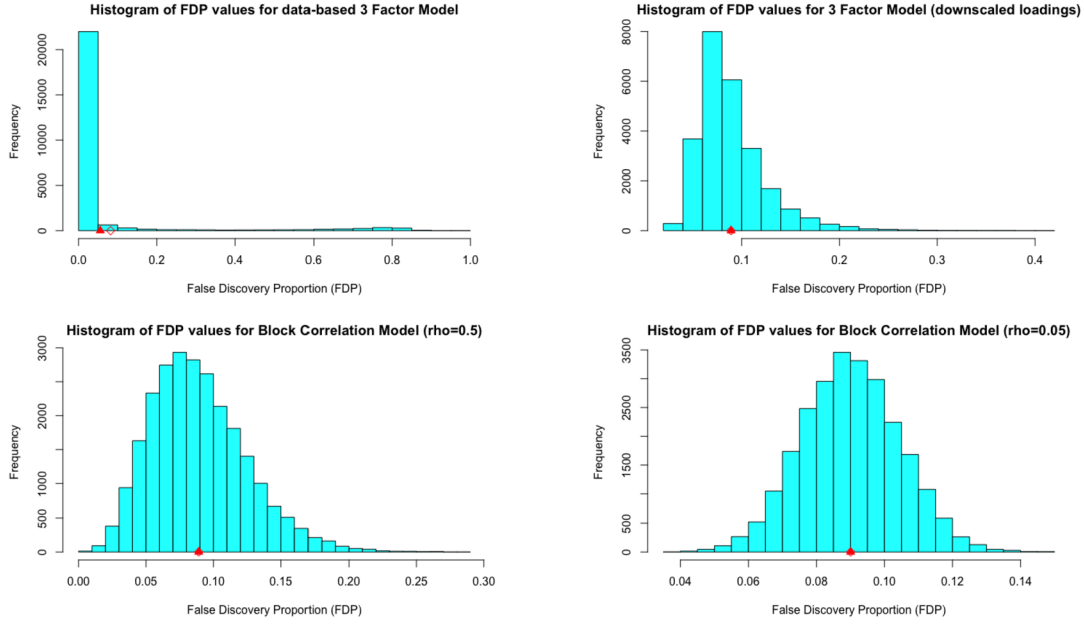

Figure 1. These are histograms of the false discovery proportion in 25,000 simulations. The data come from the two-group mixture model as described in the text. Each histogram's mean is marked with a triangle and each histogram's mean amongst nonzero FDP values is marked with a hollow diamond, which estimate the FDR and pFDR respectively. For three of the histograms, the triangle and diamond are close enough to overlap. The target FDR control is $q = 0.1$.

expected value of the FDP conditional on there being at least one rejection, does not exceed 0.1 in these simulations either. Of the four histograms, the one with data-driven correlations shows a long tailed distribution for FDP that we consider extremely bursty, the two with synthetic block correlations show FDPs that are typically quite close to the target FDR of $q = 0.1$, and the one with downscaled data-driven correlations is intermediate.

The organization of this paper is as follows. In Section 2, we describe our multiple testing setup and our model to account for both long-range correlations and short-range correlations among the test statistics. We also introduce our notation, definitions and the conditions under which our results hold. In Section 3, we state our most general CLTs for the FDP and FPR of the BH procedure. These hold conditionally on the common latent factors in our model. The proofs of these theorems are provided in the supplemental material (Kluger and Owen, 2024). In Section 4, we exploit these theorems to obtain FDP CLTs (that are not conditional on a latent factor) for settings where the long-range dependence is rapidly decaying and no factor model component is needed to account for long-range dependencies. In Section 5, we exploit these theorems to obtain FDP and FPR CLTs conditional on the latent factor that give insight into the burstiness of the BH procedure, and we show that the burstiness of the BH procedure is particularly alarming when the test statistics follow a factor model and the number of nonnulls is sparse (for example, if the number of nonnulls is $o_p(\sqrt{m})$ where m is the number of hypotheses tested). In Section 6, we describe the Duchenne Muscular Dystrophy dataset and the 3-factor model that was fit to it, and we show simulations based on the fitted factor model. In Section 7, we discuss these results and their implications for multiple testing.

2. Setup and definitions

In this section we introduce our notation for the two-group mixture model. Our version relies on a factor analysis model that we also introduce. We also review the BH procedure and state our regularity conditions in this section.

2.1. Two-group mixture model with factors

Our setting has m hypothesis tests indexed by $i = 1, \dots, m$ and our asymptotics let $m \to \infty$. In a two-group mixture model, the marginal distribution of each of the m p-values is a mixture of a common null distribution and a common nonnull distribution, both of which do not depend on i. For $1 \le i \le m < \infty$, let $H_{mi} \in \{0, 1\}$ be an indicator variable with $H_{mi} = 1$ if and only if hypothesis i of m is nonnull. We take $H_{m1}, \dots, H_{mm} \overset{\text{iid}}{\sim} \text{Bern}(\pi_1^{(m)})$ for $\pi_1^{(m)} \in (0, 1)$.

Our two-group mixture model is thus based on a sequence of nonnull probabilities, and letting $\pi_1^{(m)} \to 0$ will let us model sparsity of nonnull hypotheses. For instance, with $\pi_1^{(m)} = \lambda/m$, the number of nonnulls has constant expectation λ and has an asymptotic Poisson distribution.

To focus on dependency among tests it is convenient to assume Gaussian test statistics X_{mi} for $1 \le i \le m < \infty$. In our two-group mixture model, the test statistics where the null holds have mean zero and the ones where the alternative hypothesis holds have common mean $\mu_A > 0$. We assume that $X_{mi} = \mu_A H_{mi} + Z_{mi}$ where $(Z_{m1}, \dots, Z_{mm})$ is multivariate Gaussian, and we induce dependence among our p-values by introducing correlations among the Z_{mi}. We briefly remark that the common nonnull mean assumption is not necessary for our theoretical results to hold, but is made for cleaner exposition, as our primary interest is in investigating how the dependency between the test statistics can drive bursty behavior.

We study two kinds of dependence operating simultaneously. One is an α-mixing dependence that decays rapidly as the distance between hypothesis indices i increases. This model captures some of

the dependence one expects from hypotheses corresponding to a linearly ordered variable such as the position of a single nucleotide polymorphism (SNP) along the genome.

The other form of dependence we include is a factor model. Uses of factor models in multiple hypothesis testing include Friguet, Kloareg and Causeur (2009), Lucas, Kung and Chi (2010), Sun, Zhang and Owen (2012) and Gerard and Stephens (2020). A factor model can capture the dependency structure commonly seen among the test statistics in multiple testing problems involving gene expression data because it can capture important aspects of the correlation matrix of gene expression measurements. For example, Owen (2005) gives conditions where the correlation matrix for m test statistics measuring association of a single phenotype with expression levels of m genes is actually equal to the correlation matrix of the sampled gene measurements.

We construct the k-factor model for the array $\{Z_{mi} : 1 \le i \le m < \infty\}$ as follows. We assume that the number of factors k remains fixed as $m \to \infty$, as is assumed in Fan et al. (2019), among others. We let $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$ be the latent factor, which we suppose is only drawn once and does not change as $m \to \infty$. We let $\{\boldsymbol{L}_{mi} : 1 \le i \le m < \infty\}$ be a triangular array of fixed 'loading' vectors in $\mathbb{R}^k$. The factor model component of Z_{mi} is $\boldsymbol{L}_{mi}^\mathsf{T}\boldsymbol{W}$. For our α-mixing model with possibly strong short-range correlations but rapidly diminishing long-range correlations, we let $\{\Sigma^{(m)}\}_{m=1}^\infty$ be a sequence of covariance matrices and for each m, we let $(\varepsilon_{m1}, \ldots, \varepsilon_{mm}) \sim \mathcal{N}(0, \Sigma^{(m)})$ independently of $\boldsymbol{W}$. To combine both dependency structures, we let $Z_{mi} = \boldsymbol{L}_{mi}^\mathsf{T}\boldsymbol{W} + \varepsilon_{mi}$, giving our correlation structure for the array $\{Z_{mi} : 1 \le i \le m < \infty\}$.

We suppose that all of the test statistics have the same variance and without loss of generality, we take this common variance to be one. We do not assume the factor model to have a perfect fit, and assume instead that $\|\boldsymbol{L}_{mi}\|_2^2 + \Sigma_{ii}^{(m)} = 1$ where $\Sigma_{ii}^{(m)} > 0$ for all i, m. Because $Z_{mi} = \boldsymbol{L}_{mi}^\mathsf{T}\boldsymbol{W} + \varepsilon_{mi}$, these assumptions give $Z_{m1}, \ldots, Z_{mm} \sim \mathcal{N}(0,1)$ along with the two kinds of dependency discussed above.

We let φ and Φ denote the probability density function (PDF) and the cumulative distribution function (CDF), respectively, of $\mathcal{N}(0,1)$ and we let $\bar{\Phi} = 1 - \Phi$ be the complementary CDF. Then our p-values for one-sided hypothesis tests are

$$P_{mi} = \bar{\Phi}(X_{mi}) = \bar{\Phi}\big(\mu_A H_{mi} + \boldsymbol{L}_{mi}^\mathsf{T}\boldsymbol{W} + \varepsilon_{mi}\big) \tag{1}$$

for $1 \le i \le m < \infty$, and so $P_{mi} \sim \mathrm{Unif}(0,1)$ for the true null hypotheses. Fixing $q \in (0,1)$, throughout the text we will let $\tau_{\mathrm{BH},m}$, V_m, FDP_m, and FPR_m denote the rejection threshold, the number of false discoveries, the FDP, and the FPR respectively when applying the Benjamini-Hochberg procedure at level q to the p-values $(P_{m1}, \ldots, P_{mm})$. The formulas for these quantities are given explicitly in the next subsection, where we review the BH procedure. In our main theorems, we state CLTs for the quantities FDP_m and FPR_m conditionally on the value of the latent factor $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$.

2.2. The BH procedure

In this subsection, we describe how the BH procedure is conducted at level q on m tests with p-values $(P_{m1}, \ldots, P_{mm})$. First take the sorted p-values $P_{m(1)} \le \cdots \le P_{m(m)}$ and set $P_{m(0)} = 0$. The number of rejected hypothesis will be given by

$$R_m \equiv \max\Big\{j : P_{m(j)} \le \frac{jq}{m},\ j \in \{0, 1, \ldots, m\}\Big\}. \tag{2}$$

The BH procedure rejects the hypotheses that correspond to the R_m smallest p-values: that is it will reject all hypothesis i for which $P_{mi} \le P_{m(R_m)} \equiv \tau_{\mathrm{BH},m}$. As noted in Neuvial (2008), $\tau_{\mathrm{BH},m}$ can

equivalently be defined as the largest $t \in [0,1]$ at which the empirical CDF (ECDF) of the p-values is at least as large as t/q. We leverage this equivalence in our theorem proofs.

Letting $H_{m1}, \ldots, H_{mm}$ be as defined in Section 2.1, the number of false discoveries is

$$V_m \equiv \sum_{i=1}^{m} I\{P_{mi} \leq \tau_{\mathrm{BH},m}, H_{mi} = 0\}. \tag{3}$$

Then $\mathrm{FPR}_m \equiv V_m/m$ and $\mathrm{FDP}_m \equiv V_m/\max\{R_m, 1\}$.

2.3. Definitions and conditions

Here we present some definitions as well as regularity conditions sufficient for our conditional CLTs to hold. All of the definitions, conditions and formulas in this subsection are conditional on a fixed value of the latent factor $\boldsymbol{W} \in \mathbb{R}^k$.

2.3.1. Variance and mixing conditions on ε

Recall from our setup that $\mathrm{var}(\varepsilon_{mi}) > 0$ for all m, i, allowing us to define $\tilde{\varepsilon}_{mi} \equiv \varepsilon_{mi}/\sqrt{\mathrm{var}(\varepsilon_{mi})}$ for convenience. It is also helpful to introduce the following condition, which forces the variance of all ε_{mi} terms to be bounded away from zero.

Condition 1. $S_L \equiv \sup_{1 \leq i \leq m < \infty} \|\boldsymbol{L}_{mi}\|_2^2 < 1$.

Note about Condition 1. recalling that $\|\boldsymbol{L}_{mi}\|_2^2 + \mathrm{var}(\varepsilon_{mi}) = 1$ in our model, this condition provides a uniform bound $\mathrm{var}(\varepsilon_{mi}) \geq 1 - S_L > 0$ for all $1 \leq i \leq m < \infty$.

To describe the mixing condition on $(\varepsilon_{mi})_{1 \leq i \leq m < \infty}$, for $1 \leq i \leq m < \infty$, define $\xi_{mi} \equiv (\varepsilon_{mi}, H_{mi})$. Now let $\mathcal{A}_1^n(m)$ be the σ-field generated by the variables ξ_{mi} for $1 \leq i \leq n$ and $\mathcal{A}_{n+d}^{\infty}(m)$ be the σ-field generated by the variables ξ_{mi} for $n + d \leq i \leq m$. For integers $d \geq 1$ our α-mixing parameters $\alpha(d) \in [0,1]$ are defined by

$$\alpha(d) \equiv \sup_{n,m \in \mathbb{N}} \sup_{\substack{A_0 \in \mathcal{A}_1^n(m) \\ A_1 \in \mathcal{A}_{n+d}^{\infty}(m)}} \left| P(A_0 \cap A_1) - P(A_0)P(A_1) \right|. \tag{4}$$

Condition 2. There exists an even integer $Q > 2$ and $\gamma > 0$ such that both

$$\text{(i)} \quad \frac{\gamma}{2+\gamma} + \frac{2}{Q} < 1 \quad \text{and} \quad \text{(ii)} \quad \sum_{d=1}^{\infty} d^{Q-2} \alpha(d)^{\frac{\gamma}{Q+\gamma}} < \infty$$

Notes about Condition 2. Throughout the text we will let Q, γ be such numbers. Note that it is possible that this condition can be loosened to allow Q to be rational, but then we need to trust a claim in Andrews and Pollard (1994) that their Theorem 2.2 would still hold for Q not an even integer.

Condition 2 will hold when the correlation between ε_{mi} and ε_{mj} is a rapidly decaying function of $|i - j|$. If this correlation is always zero for each $|i - j| > M$ (making the error sequences $(\varepsilon_{mi})_{1 \leq i \leq m}$ M-dependent for each m), Condition 2 will hold. In the following remark, we argue that Condition 2 will typically hold when $(\tilde{\varepsilon}_{mi})_{1 \leq i \leq m}$ is modelled by a stationary ARMA process or by a stationary GARCH process (a definition of these processes can be found in Paolella (2019), for example).

Remark 1. Suppose that the standardized errors $\tilde{\varepsilon}_{mi}$ just depend on i and not on m. If $(\tilde{\varepsilon}_i)_{i\in\mathbb{Z}}$ can be modelled by a stationary ARMA model with absolutely continuous errors with respect to Lebesgue measure on $\mathbb{R}$, then Condition 2 will hold. To see this, note that by Theorem 1 in Mokkadem (1988), such a stationary ARMA process $(\tilde{\varepsilon}_i)_{i\in\mathbb{Z}}$ will be geometrically completely regular and hence the α-mixing coefficients of $(\tilde{\varepsilon}_i)_{i\in\mathbb{Z}}$ will be $O(\theta^d)$ for some $\theta \in (0,1)$. By independence of the H_{mi} and since $\mathrm{var}(\varepsilon_i) > 0$, this implies that $\alpha(d) = O(\theta^d)$ will hold for that same $\theta \in (0,1)$, further implying that Condition 2 will hold. By similar reasoning, if $(\tilde{\varepsilon}_i)_{i\in\mathbb{Z}}$ is modeled by a stationary GARCH process, Theorem 8 in Lindner (2009) implies that under certain conditions on the GARCH process errors, Condition 2 will hold.

2.3.2. Definitions of some subdistributions of p-values and their condition

For any positive integer m, define $\pi_0^{(m)} \equiv 1 - \pi_1^{(m)}$, and then write

$$H_{mi0} \equiv 1 - H_{mi} \quad \text{and} \quad H_{mi1} \equiv H_{mi}.$$

Our subsequent definitions use $r = 0$ for quantities based on the null hypotheses and $r = 1$ for quantities from the nonnull hypotheses. For $t \in [0,1]$ and $r \in \{0,1\}$ let

$$\hat{F}_{m,r}(t) \equiv \frac{1}{m}\sum_{i=1}^{m} H_{mir} I\{P_{mi} \le t\} = \frac{1}{m}\sum_{i=1}^{m} H_{mir} I\{\bar{\Phi}(\mu_A r + \varepsilon_{mi} + \boldsymbol{L}_{mi}^{\mathsf{T}}\boldsymbol{W}) \le t\}.$$

We call $\hat{F}_{m,0}$ and $\hat{F}_{m,1}$ the empirical subdistribution functions of the null and nonnull p-values respectively. These empirical subdistribution functions sum to the ECDF of the p-values. Let $\gamma_{mir} : [0,1] \to [0,1]$ be the monotone increasing bijection given by

$$\gamma_{mir}(t) \equiv \Pr(P_{mi} \le t \,|\, H_{mi} = r, \boldsymbol{W} = \boldsymbol{w}) = \bar{\Phi}\Big(\frac{\bar{\Phi}^{-1}(t) - \mu_A r - \boldsymbol{L}_{mi}^{\mathsf{T}}\boldsymbol{w}}{\sqrt{1 - \|\boldsymbol{L}_{mi}\|_2^2}}\Big).$$

We aggregate γ_{mir} in the following subdistribution functions

$$F_{m,r}(t) \equiv \mathbb{E}\big(\hat{F}_{m,r}(t) \,|\, \boldsymbol{W} = \boldsymbol{w}\big) = \frac{\pi_r^{(m)}}{m}\sum_{i=1}^{m}\gamma_{mir}(t)$$

and then let

$$F_r(t) \equiv \lim_{m\to\infty} F_{m,r}(t) = \lim_{m\to\infty} \frac{\pi_r^{(m)}}{m}\sum_{i=1}^{m}\gamma_{mir}(t). \tag{5}$$

Condition 3 ensures that these quantities are well defined.

Condition 3. For all $t \in [0,1]$ and $r \in \{0,1\}$, $F_r(t) \equiv \lim_{m\to\infty} F_{m,r}(t)$ exists.

2.3.3. Defining the asymptotic ECDF and the Simes point

Now define $\hat{G}_m, G : [0,1] \to [0,1]$ via

$$\hat{G}_m(t) \equiv \hat{F}_{m,0}(t) + \hat{F}_{m,1}(t) \quad \text{and} \quad G(t) \equiv F_0(t) + F_1(t) \tag{6}$$

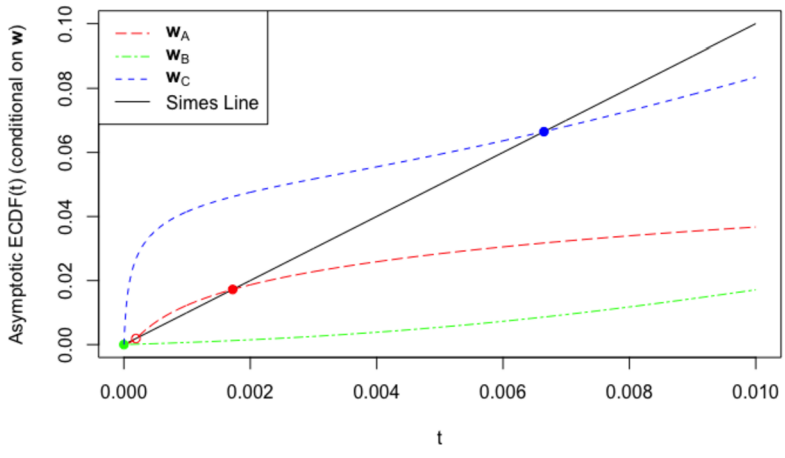

Figure 2. The curves are conditional asymptotic ECDFs of p-values in a 3-factor model based on some Duchenne Muscular Dystrophy data described in Section 6. The three draws satisfy $\Phi(\boldsymbol{w}_A) = (0.8, 0.4, 0.9)$, $\Phi(\boldsymbol{w}_B) = (0.45, 0.56, 0.62)$ and $\Phi(\boldsymbol{w}_C) = (0.02, 0.85, 0.78)$. Filled circles show the Simes points. An open circle for $\boldsymbol{w}_A$ shows a crossing of the Simes line that is not the Simes point because it is not the final crossing.

for $t \in [0,1]$. Note that $\hat{G}_m$ is the ECDF of the p-values and G is the limiting expected ECDF of the p-values. Throughout the text we will refer to G as the *asymptotic ECDF* because under most dependency structures, we expect G to be the point-wise limit in probability of $\hat{G}_m$.

The rejection threshold for the BH procedure at level q is the largest point t such that the ECDF of the p-values evaluated at t lies above the line through the origin of slope $1/q$, called the Simes line. It is reasonable to expect the limiting p-value rejection threshold for the BH procedure at level q to be the largest t at which $(t, G(t))$ intersects the Simes line. We use the term *Simes point* to describe the largest point where the asymptotic ECDF intersects the Simes line. More precisely, the Simes point is

$$\tau_* \equiv \sup\bigl\{t \in (0,1) \,:\, G(t) \geq t/q\bigr\}, \tag{7}$$

interpreting the supremum of the empty set to be zero. The Simes point satisfies $0 \leq \tau_* \leq q$. The upper limit follows from $G(t) \leq 1$. Both G and τ_* depend on the specific realization of latent factor $\boldsymbol{W} \in \mathbb{R}^k$ on which we condition.

Figure 2 illustrates the Simes points. The setting has $\mu_A = 2$, $\pi_0 = 0.9$ and $q = 0.1$. The horizontal axis has putative p-values over the range $t \in [0, 0.01]$. The Simes line is t/q. There are $m = 22{,}283$ hypotheses corresponding to genes in the GDS 3027 Duchenne Muscular Dystrophy data described in Section 6. For three draws $\boldsymbol{W} \sim \mathcal{N}(0, I_3)$ we show the asymptotic ECDF curves. One of them crosses the Simes line twice and the Simes point is the last crossing. One crosses it only once and one has Simes point $\tau_* = 0$ because the Simes line is never crossed.

We will need continuity of $G(\cdot)$ on $(0,1)$ under Conditions 1 and 3. We do not know whether G must be continuous at 0 or 1, but our results do not depend on that.

Proposition 2.1. *Under Conditions 1 and 3, G is continuous on* $(0,1)$.

Proof. It is sufficient to show that G is Lipschitz continuous on $(\epsilon, 1-\epsilon)$ whenever $0 < \epsilon < 1/2$. For any such ϵ, observe that for $r \in \{0,1\}$ and integers $1 \le i \le m < \infty$

$$\gamma'_{mir}(t) = \frac{1}{\varphi(\Phi^{-1}(t))\sqrt{1-\|\boldsymbol{L}_{mi}\|_2^2}}\,\varphi\left(\frac{\bar{\Phi}^{-1}(t) - \boldsymbol{L}_{mi}^{\mathsf{T}}\boldsymbol{w} - \mu_A r}{\sqrt{1-\|\boldsymbol{L}_{mi}\|_2^2}}\right).$$

Now $\varphi(\cdot) \le 1/\sqrt{2\pi}$ and then using Condition 1 it follows that for any $t \in (\epsilon, 1-\epsilon)$,

$$|\gamma'_{mir}(t)| \le \frac{1/\sqrt{2\pi}}{\varphi(\Phi^{-1}(t))\sqrt{1-S_L}} \le \frac{1/\sqrt{2\pi}}{\varphi(\Phi^{-1}(\epsilon))\sqrt{1-S_L}} \equiv C_\epsilon.$$

Since $\sup_{t\in(\epsilon,1-\epsilon)} |\gamma'_{mir}(t)| \le C_\epsilon < \infty$, $|\gamma_{mir}(t) - \gamma_{mir}(s)| \le C_\epsilon |t-s|$ for any $t, s \in (\epsilon, 1-\epsilon)$. This argument holds for any $1 \le i \le m < \infty$ and $r \in \{0,1\}$, and so for any $t, s \in (\epsilon, 1-\epsilon)$ and integer m and $r \in \{0,1\}$,

$$|F_{m,r}(t) - F_{m,r}(s)| \le \frac{\pi_r^{(m)}}{m} \sum_{i=1}^{m} |\gamma_{mir}(t) - \gamma_{mir}(s)| \le C_\epsilon |t-s|.$$

Taking the limit as $m \to \infty$ of the left side of the above inequality, which exists by Condition 3, we get $|F_r(t) - F_r(s)| \le C_\epsilon |t-s|$ for all $t, s \in (\epsilon, 1-\epsilon)$ and for both $r \in \{0,1\}$. Thus, F_r is Lipschitz continuous on $(\epsilon, 1-\epsilon)$ for $r \in \{0,1\}$ which implies $G = F_0 + F_1$ is Lipschitz continuous on $(\epsilon, 1-\epsilon)$. □

2.3.4. Defining a focal interval $[a,b] \subset [0,1]$ for our processes

We are going to work with an interval $[a,b]$ of positive length for which the Simes point τ_* is the unique element $t \in (a,b)$ with $G(t) = t/q$. First we need a technical condition to rule out some pathological behavior. Under this condition there will exist such an interval $[a,b]$.

Condition 4. The Simes point is positive, is the largest point where G actually crosses the Simes line, and is not an accumulation point for points of intersection of G and the Simes line. That is,

(i) $\tau_* > 0$,
(ii) $\tau_* = \sup\{t \in (0,1) : G(t) > t/q\}$, and
(iii) τ_* is not an accumulation point of $\{t \in (0,1) : G(t) = t/q\}$.

Note about Condition 4. For many factor model choices, Condition 4 will hold with some probability in $(0,1)$ depending on the specific realization of $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$. This is due to the dependence of τ_* and G on the specific realization of the latent factor $\boldsymbol{W} \in \mathbb{R}^k$ on which we condition. For example, see Figure 2.

Proposition 2.2. *If Conditions 1, 3, and 4 hold, then for any $b \in (q,1)$ there exists a point $a \in (0,q)$ such that*

(i) *$G(a) > a/q$, and*
(ii) *the Simes point τ_* is the unique $t \in (a,b)$ solving $G(t) = t/q$.*

Proof. Pick any $b \in (q,1)$ and suppose that Conditions 1, 3, and 4 hold. By Condition 4,

$$\tau_* \equiv \sup\{t \in (0,1) : G(t) \ge t/q\} = \sup\{t \in (0,1) : G(t) > t/q\} > 0.$$

Now $\tau_* \le q$ because $G(t) \le 1$ for all $t \in (0,1)$. Also, $G(\tau_*) = \tau_*/q$ by continuity of G (see Proposition 2.1). Hence, because $\tau_* = \sup\{t \in (0,1) : G(t) > t/q\}$ but $G(\tau_*) = \tau_*/q$, there exists a sequence $a_n \uparrow \tau_*$ such that for all n, $G(a_n) > a_n/q$ and $a_n < \tau_*$. Since $\{t \in (0,1) : G(t) = t/q\}$ does not have an accumulation point at τ_* and since $G(t) - t/q$ is continuous, there is a sufficiently large N_* with $G(t) > t/q$ for all $t \in [a_{N_*}, \tau_*)$. We choose $a = a_{N_*} \in (0, \tau_*)$ and then property (i) holds by our definition of a_n. Also, $a \in (0,q)$ because $a < \tau_* \le q$. Turning to property (ii), $G(t) > t/q$ for all $t \in [a, \tau_*)$ by the the choice of N_* and a, while for all $t \in (\tau_*, b)$, $G(t) < t/q$ by the definition of τ_*. □

Throughout the text, when conditioning on $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$, if Conditions 1, 3, and 4 hold, we will let $[a,b]$ be an interval satisfying the properties (i) and (ii) with $a \in (0,q)$ and $b \in (q,1)$ that are guaranteed by Proposition 2.2.

2.3.5. Defining our stochastic process and its Gaussian process limit

Our stochastic processes of interest are two jointly distributed random càdlàg functions on $[a,b]$. We will show convergence to a pair of Gaussian processes with continuous sample paths on $[a,b]$. The expressions $C([a,b] \times \{0,1\})$ and $(C[a,b])^2$ are both awkward, while $C[a,b]^2$ denotes functions on a square region. Therefore, we use the symbol $[a,b]_2$ to denote $[a,b] \times \{0,1\}$ and study random elements in $C[a,b]_2$ and $D[a,b]_2$. Explicitly, $C[a,b]_2$ is the collection of all pairs of real valued continuous functions on $[a,b]$ while $D[a,b]_2$ is the collection of all pairs of real valued càdlàg functions on $[a,b]$. We study the following processes in $D[a,b]_2$:

$$\begin{aligned} W_{m,r}(t) &\equiv \sqrt{m}\big(\hat{F}_{m,r}(t) - F_{m,r}(t)\big) \quad \text{and} \\ \hat{W}_{m,r}(t) &\equiv \sqrt{m}\big(\hat{F}_{m,r}(t) - F_r(t)\big). \end{aligned} \tag{8}$$

We are ultimately interested in a functional central limit theorem (FCLT) for the joint process $\big(\hat{W}_{m,0}(\cdot), \hat{W}_{m,1}(\cdot)\big)$, so we must find the limiting joint covariance kernel of this pair of processes. To describe this limiting covariance kernel, we introduce some convenient definitions and notation.

For convenience, throughout the text we will define $\{\Gamma^{(m)}\}_{m=1}^{\infty}$ to be the sequence of correlation matrices corresponding to $\{\Sigma^{(m)}\}_{m=1}^{\infty}$ and, as before, for each m,i define $\tilde{\varepsilon}_{mi} \equiv \varepsilon_{mi}/\sqrt{\mathrm{var}(\epsilon_{mi})} = \varepsilon_{mi}/\sqrt{1 - \|\boldsymbol{L}_{mi}\|_2^2}$. Note that $(\tilde{\varepsilon}_{m1}, \ldots, \tilde{\varepsilon}_{mm}) \sim \mathcal{N}(0, \Gamma^{(m)})$ and that each $\tilde{\varepsilon}_{mi}$ has unit variance. For any $t,s \in [0,1]$ and $|\rho| \le 1$, define

$$\tilde{\rho}(t,s,\rho) \equiv \Pr\left(\tilde{\varepsilon}_1 \ge \bar{\Phi}^{-1}(t), \tilde{\varepsilon}_2 \ge \bar{\Phi}^{-1}(s) \;\middle|\; \begin{pmatrix}\tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2\end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right)\right) - ts. \tag{9}$$

Given a bivariate Gaussian with unit variance and correlation ρ, the above quantity is the covariance between the indicator that the first coordinate of this bivariate Gaussian exceeds its $1-t$ quantile and the indicator that the second coordinate of this bivariate Gaussian exceeds its $1-s$ quantile.

Now for any $s,t \in (a,b)$ and $r_0, r_1 \in \{0,1\}$ and $m \in \mathbb{N}_+$ define

$$c_m^{(r_0,r_1)}(t,s) \equiv \mathrm{cov}\big(W_{m,r_0}(t), W_{m,r_1}(s)\big).$$

It is convenient to break up the expression of $c_m^{(r_0,r_1)}(t,s)$ into two terms. Define

$$c_{m,\mathrm{diag}}^{(r_0,r_1)}(t,s) \equiv \frac{1}{m}\sum_{i=1}^{m}\Big(\pi_{r_0}^{(m)}\gamma_{mir_0}(t \wedge s)I\{r_0 = r_1\} - \pi_{r_0}^{(m)}\pi_{r_1}^{(m)}\gamma_{mir_0}(t)\gamma_{mir_1}(s)\Big),$$

and define

$$c^{(r_0,r_1)}_{m,\text{cross}}(t,s) \equiv \frac{\pi^{(m)}_{r_0}\pi^{(m)}_{r_1}}{m}\sum_{i\neq j}\tilde{\rho}\Big(\gamma_{mir_0}(t),\gamma_{mjr_1}(s),\Gamma^{(m)}_{ij}\Big).$$

In the following proposition we show that $c_m = c_{m,\text{diag}} + c_{m,\text{cross}}$.

Proposition 2.3. *For any $s,t \in [a,b]$ and $r_0, r_1 \in \{0,1\}$ and $m \geq 2$*

$$c^{(r_0,r_1)}_m(t,s) = c^{(r_0,r_1)}_{m,\text{diag}}(t,s) + c^{(r_0,r_1)}_{m,\text{cross}}(t,s).$$

Proof. For $i,j \in [m]$ define

$$\begin{aligned} C^{(r_0,r_1)}_{i,j,m}(t,s) &\equiv \text{cov}\big(H_{mir_0}I\{P_{mi}\leq t\}, H_{mjr_1}I\{P_{mj}\leq s\}\big)\\ &= \text{cov}\big(H_{mir_0}I\{\bar{\Phi}(\tilde{\varepsilon}_{mi})\leq \gamma_{mir_0}(t)\}, H_{mjr_1}I\{\bar{\Phi}(\tilde{\varepsilon}_{mj})\leq \gamma_{mjr_1}(s)\}\big).\end{aligned}$$

For $i \neq j$, H_{mir_0}, H_{mjr_1} and $\tilde{\varepsilon}$ are all independent, so

$$\begin{aligned} C^{(r_0,r_1)}_{i,j,m}(t,s) &= \pi^{(m)}_{r_0}\pi^{(m)}_{r_1}\text{cov}\big(I\{\bar{\Phi}(\tilde{\varepsilon}_{mi})\leq \gamma_{mir_0}(t)\}, I\{\bar{\Phi}(\tilde{\varepsilon}_{mj})\leq \gamma_{mjr_1}(s)\}\big)\\ &= \pi^{(m)}_{r_0}\pi^{(m)}_{r_1}\tilde{\rho}\big(\gamma_{mir_0}(t),\gamma_{mjr_1}(s),\Gamma^{(m)}_{ij}\big).\end{aligned}$$

When $i=j$, $H_{mir_0}H_{mjr_1} = H_{mir_0}I\{r_0=r_1\}$ and $\tilde{\epsilon}_{mi} = \tilde{\epsilon}_{mj}$, so that

$$C^{(r_0,r_1)}_{i,j,m}(t,s) = \pi^{(m)}_{r_0}I\{r_0=r_1\}\gamma_{mir_0}(t\wedge s) - \pi^{(m)}_{r_0}\pi^{(m)}_{r_1}\gamma_{mir_0}(t)\gamma_{mir_1}(s).$$

Since the above expressions hold for any $i,j \in [m]$,

$$\begin{aligned} c^{(r_0,r_1)}_m(t,s) &= \text{cov}\big(W_{m,r_0}(t), W_{m,r_1}(s)\big)\\ &= m\text{cov}\big(\hat{F}_{m,r_0}(t), \hat{F}_{m,r_1}(s)\big)\\ &= \frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m \text{cov}\big(H_{mir_0}I\{P_{mi}\leq t\}, H_{mjr_1}I\{P_{mj}\leq s\}\big)\\ &= \frac{1}{m}\sum_{i=1}^m\sum_{j=1}^m C^{(r_0,r_1)}_{i,j,m}(t,s)\\ &= \frac{1}{m}\sum_{i=1}^m C^{(r_0,r_1)}_{i,i,m}(t,s) + \frac{1}{m}\sum_{i\neq j} C^{(r_0,r_1)}_{i,j,m}(t,s)\\ &= c^{(r_0,r_1)}_{m,\text{diag}}(t,s) + c^{(r_0,r_1)}_{m,\text{cross}}(t,s).\end{aligned}$$

□

Now define

$$c^{(r_0,r_1)}(t,s) \equiv \lim_{m\to\infty} c^{(r_0,r_1)}_m(t,s). \tag{10}$$

By the simplified formula for $c^{(r_0,r_1)}_m(t,s)$, the above limit exists by Condition 3 if we further impose Condition 5 below.

Condition 5. For any $(s,t) \in [a,b]$ and $r_0, r_1 \in \{0,1\}$,

$$\lim_{m\to\infty} \frac{\pi_{r_0}^{(m)}\pi_{r_1}^{(m)}}{m} \sum_{i=1}^{m} \gamma_{mir_0}(t)\gamma_{mir_1}(s) \quad \text{and}$$

$$\lim_{m\to\infty} \frac{\pi_{r_0}^{(m)}\pi_{r_1}^{(m)}}{m} \sum_{i\neq j} \tilde{\rho}(\gamma_{mir_0}(t), \gamma_{mir_1}(s), \Gamma_{ij}^{(m)})$$

both exist.

It is easy to see that the function $c(\cdot,\cdot)$ defined above gives a joint covariance kernel that is symmetric and positive semidefinite because it is the limit of symmetric and positive semidefinite joint covariance kernels c_m.

2.3.6. Regularity conditions on F_0, F_1, $F_{m,0}$ and $F_{m,1}$

Before introducing the main theorems, we introduce another two conditions that will be used in their proof.

Condition 6. Both F_0 and F_1 are differentiable at τ_*.

The final condition is needed to derive a $\big(\hat{W}_{m,0}(\cdot), \hat{W}_{m,1}(\cdot)\big)$ FCLT from a $\big(W_{m,0}(\cdot), W_{m,1}(\cdot)\big)$ FCLT. We would like to hold the subdistribution functions $F_{m,0}$ and $F_{m,1}$ constant as m changes but this does not hold in all cases of interest. Instead we assume that they approach limits F_0 and F_1 at a fast rate.

Condition 7. For $r \in \{0,1\}$, $\lim_{m\to\infty} \sup_{t\in[a,b]} \big|\sqrt{m}\big(F_{m,r}(t) - F_r(t)\big)\big| = 0$.

Notes about Condition 7. As with all of the other conditions in Section 2.3, Condition 7 merely needs to hold for the fixed value of the latent factor $\boldsymbol{W} \in \mathbb{R}^k$ on which we condition. In addition, a version of Theorem 3.1 below will still hold if we loosen Condition 7 to say that there exist Gaussian processes Z_0 and Z_1 on $[a,b]$ that are independent from the noise ε such that for both $r \in \{0,1\}$,

$$\sqrt{m}\big(F_{m,r}(\cdot) - F_r(\cdot)\big) \xrightarrow{\mathcal{D}} Z_r(\cdot).$$

This looser condition can be useful to study asymptotic behavior of the BH procedure in settings where the nonnull effect sizes μ_A are not constant and instead are assumed to come from some prior distribution. However, if we use this looser condition, the resulting theorem statement will be messier.

3. Statement of the theorems

Theorem 3.1. *For the model of Section 2.1, suppose that conditionally on a specific value of the latent factor $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$ that Conditions 1–7 all hold. Then*

$$\sqrt{m}\Big(\mathrm{FDP}_m - \frac{qF_0(\tau_*)}{\tau_*} \mid \boldsymbol{W} = \boldsymbol{w}\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_L^2) \tag{11}$$

as $m \to \infty$ where

$$\sigma_L^2 \equiv \frac{q^2}{\tau_*^2}\big((1+\alpha)^2 c^{(0,0)}(\tau_*,\tau_*) + \alpha^2 c^{(1,1)}(\tau_*,\tau_*) + 2\alpha(1+\alpha)c^{(1,0)}(\tau_*,\tau_*)\big) \tag{12}$$

for the function F_0 given at (5), the asymptotic ECDF G given at (6), the Simes point τ_* from (7), the covariances $c^{(r_0,r_1)}(\cdot,\cdot)$ given by (10) and

$$\alpha \equiv \frac{F_0'(\tau_*) - F_0(\tau_*)/\tau_*}{1/q - G'(\tau_*)}. \tag{13}$$

Proof. See the supplemental material for a proof of this theorem. □

The proof is quite long but to summarize we first derive an FCLT for the joint process $(\hat{F}_{m,0}, \hat{F}_{m,1})$ by proving finite dimensional distribution convergence using a CLT from Neumann (2013) and then extend to an FCLT by using a result from Andrews and Pollard (1994). We then define $\Psi^{(\mathrm{FDP})} : D[a,b]_2 \to \mathbb{R}$ to be a particular function satisfying $\mathrm{FDP}_m = \Psi^{(\mathrm{FDP})}(\hat{F}_{m,0}, \hat{F}_{m,1})$ with probability converging to 1 as $m \to \infty$. Then we argue that $\Psi^{(\mathrm{FDP})}$ is Hadamard differentiable at (F_0, F_1) tangentially to $C[a,b]_2$ and compute the Hadamard derivative by mimicking the approach in Neuvial (2008). To complete the proof of the CLT given in (11), we tie these results together with the functional delta method in Chapter 20.2 of van der Vaart (1998). Using the same proof technique we obtain the following conditional CLT for the ratio V_m/m.

Theorem 3.2. *Under the conditions of Theorem 3.1,*

$$\sqrt{m}\Big(\frac{V_m}{m} - F_0(\tau_*) \,\Big|\, \boldsymbol{W} = \boldsymbol{w}\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_R^2) \quad \text{as } m \to \infty \tag{14}$$

where

$$\sigma_R^2 \equiv (1+\beta)^2 c^{(0,0)}(\tau_*,\tau_*) + \beta^2 c^{(1,1)}(\tau_*,\tau_*) + 2\beta(1+\beta)c^{(1,0)}(\tau_*,\tau_*) \tag{15}$$

for

$$\beta \equiv \frac{F_0'(\tau_*)}{1/q - G'(\tau_*)}. \tag{16}$$

Proof. See the supplemental material for a proof of this theorem. □

It is often the case that for a particular factor and noise model, Conditions 1–7 will not hold for all values of the latent factor $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$. Condition 4 will often be violated when drawing $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$. We do not expect a positive probability that τ_* will be an accumulation point of $\{t : G(t) = t/q\}$, nor do we expect a positive probability that $\sup\{t \in (0,1) : G(t) \geq t/q\} \neq \sup\{t \in (0,1) : G(t) > t/q\}$, but we do expect a positive probability that $\tau_* = 0$. Therefore, in the next theorem we describe the asymptotic behavior of the BH procedure, conditional on $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$ when $\tau_* = 0$.

Theorem 3.3. *For the model of Section 2.1, when conditioning on the latent factor $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$, suppose that Conditions 1, 2, and 3 hold, that $\tau_* = 0$, and that Conditions 5, 7 hold when setting $[a,b] = [0,1]$. Then*

$$\tau_{\mathrm{BH},m} \,|\, \boldsymbol{W} = \boldsymbol{w} \xrightarrow{p} 0 \quad \text{and} \quad \frac{V_m}{m} \,|\, \boldsymbol{W} = \boldsymbol{w} \xrightarrow{p} 0. \tag{17}$$

Proof. See the supplemental material for a proof of this theorem. □

Remark 2. Theorem 3.3 will still hold if we loosen the mixing Condition 2 and simply require that the error array $\{\varepsilon_{mi} : 1 \le i \le m < \infty\}$ has summable α-mixing coefficients. This is because the proof of Proposition S.2.1 in the supplemental material (Kluger and Owen, 2024) does not require Condition 2 to hold and merely requires that the error array has summable α-mixing coefficients.

Remark 3. When $\tau_* = 0$ and $\tau_{\mathrm{BH},m} \xrightarrow{p} 0$, it is not guaranteed that $\mathrm{FDP}_m \xrightarrow{p} 0$. For example, in the scenario where $\boldsymbol{L}_{mi} = 0$ for all m,i, the errors $(\varepsilon_{m1},\dots,\varepsilon_{mm})$ are independent and $\pi_1^{(m)} = 1/m$, one can show that $\tau_* = 0$ and $\tau_{\mathrm{BH},m} \xrightarrow{p} 0$, yet $\liminf_{m\to\infty} \Pr(\mathrm{FDP}_m \ge 0.5) > 0$. As another example, Gontscharuk and Finner (2013) provide a scenario where $\tau_{\mathrm{BH},m} \xrightarrow{p} 0$, but asymptotically the false discovery rate exceeds the FDR control parameter q, implying that FDP_m cannot possibly converge to zero in probability in their scenario.

In the above theorems, the limiting of behavior of BH depends on a latent factor which in practice is unobserved. Estimation of the unobserved latent factor is out of scope for this paper but for estimators of the latent factor and properties of these estimators we point the reader to Fan, Han and Gu (2012), Azriel and Schwartzman (2015), Sun, Zhang and Owen (2012), Wang et al. (2017) and Fan et al. (2019). In the next section we focus on results for the case where the correlations are short-range and there is no factor model or latent factor to consider.

4. Corollaries when there is no factor model component

The simplest applications of Theorems 3.1 and 3.2 are to settings where there is no factor model component. That is $k = 0$, or equivalently $\boldsymbol{L}_{mi} = 0$ for $1 \le i \le m$. Then the test statistics are $X_{mi} = \mu_A H_{mi} + \varepsilon_{mi}$ where $(\varepsilon_{m1},\dots,\varepsilon_{mm}) \sim \mathcal{N}(0,\Gamma^{(m)})$ for a correlation matrix $\Gamma^{(m)} \in \mathbb{R}^{m\times m}$. Next suppose, as is usual in the two-group mixture model that $\pi_1^{(m)} = \pi_1 \in (0,1)$ for all $m \ge 1$. Finally, we will assume that the errors $(\varepsilon_{m1},\dots,\varepsilon_{mm})$ satisfy mixing Condition 2, which can hold, for example, if the errors are M-dependent (see Remark 1 for other examples where Condition 2 is met).

Many of the 7 conditions in our theorems hold trivially in this setting. Most trivially, $S_L = 0$ making Condition 1 hold. The mixing Condition 2 holds by assumption. Also in this setting, because $F_0(t) = F_{m,0}(t) = (1-\pi_1)t$ and $F_1(t) = F_{m,1}(t) = \pi_1\bar{\Phi}(\bar{\Phi}^{-1}(t) - \mu_A)$ for each m, Conditions 3 and 7 on the subdistributions can be seen to hold.

To check that Condition 4 ruling out pathologies about τ_* holds note that $G(t) = (1-\pi_1)t + \pi_1\bar{\Phi}(\bar{\Phi}^{-1}(t) - \mu_A)$. A simple calculation shows that $G'(t) = (1-\pi_1) + \exp(\mu_A\bar{\Phi}^{-1}(t) - \mu_A^2/2)$ and

$$G''(t) = -\pi_1 \exp(\mu_A\bar{\Phi}^{-1}(t) - \mu_A^2/2)/\varphi(\bar{\Phi}^{-1}(t)),$$

implying that G is strictly concave on $(0,1)$ and that $G'(t) \to \infty$ as $t \downarrow 0$. By strict concavity of G, and since both G and the Simes line intersect the origin, $\{t > 0 : G(t) = t/q\}$ contains at most one point. It remains to show existence of a point $t > 0$ with $G(t) = t/q$. Since $G'(t) \to \infty$ as $t \downarrow 0$ and $G(0) = 0$, there must be an $\epsilon > 0$ such that $G(\epsilon) > \epsilon/q$. Also $G(1) = 1 < 1/q$, so the continuous function $t \mapsto G(t) - t/q$ must cross 0 at some unique $t_* \in (0,1)$. By continuity of G and uniqueness this unique t_* is the Simes point τ_* defined in (7) and further Condition 4 will be satisfied. Because $\tau_* \in (0,1)$ and because F_0 and F_1 are differentiable on $(0,1)$, Condition 6 also holds.

The only remaining condition to check is Condition 5 on convergence of the covariance kernels. Let $(a,b) \subset (0,1)$ be any open interval containing τ_* and q. Since for $r \in \{0,1\}$, γ_{mir} does not vary

with i or m, the first limit in Condition 5 always holds. Since for each m,i, $\gamma_{mi0}(t) = t$ and $\gamma_{mi1}(t) = \bar{\Phi}(\bar{\Phi}^{-1}(t) - \mu_A)$, Condition 5 holds whenever

$$\varrho(t,s) \equiv \lim_{m\to\infty} \frac{1}{m} \sum_{i\neq j} \tilde{\rho}(t,s,\Gamma_{ij}^{(m)}) \tag{18}$$

exists for all $s,t \in \left[a, \bar{\Phi}(\bar{\Phi}^{-1}(b) - \mu_A)\right]$ with $\tilde{\rho}$ defined at (9). We summarize this along with an application of Theorem 3.1 in Corollary 4.1.

Corollary 4.1. *In the setting of Section 2.1, suppose further that:*

i) *the factor loadings $\boldsymbol{L}_{mi}$ are all zero and*
ii) *the probability $\pi_1 \in (0,1)$ of nonnull hypotheses does not depend on m.*

Let τ_* be the unique $t \in (0,1)$ satisfying $t/q = \pi_0 t + \pi_1 \bar{\Phi}(\bar{\Phi}^{-1}(t) - \mu_A)$ and let $(a,b) \subset (0,1)$ be any open interval containing both τ_* and q. If the ε_{mi} are such that mixing Condition 2 holds and the correlations among the ε_{mi} are such that $\varrho(t,s)$ defined at (18) exists for all $t,s \in \left[a, \bar{\Phi}(\bar{\Phi}^{-1}(b) - \mu_A)\right]$, then

$$\sqrt{m}\big(\mathrm{FDP}_m - \pi_0 q\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\pi_0^2 q^2}{\tau_*^2}\Big(\frac{\tau_*}{\pi_0} - \tau_*^2 + \varrho(\tau_*,\tau_*)\Big)\Big) \tag{19}$$

where $\pi_0 \equiv 1 - \pi_1$.

Proof. As discussed before the statement of the corollary, Conditions 1–7 hold in this setting. Noting that in this setting $\alpha = 0$ and there is no dependency on the latent factor $\boldsymbol{W}$, the result holds by Theorem 3.1. □

We note that this corollary will hold even if $q \notin (a,b)$ and the requirement that $q \in (a,b)$ was included for a cleaner proof of Theorem 3.1.

We also note that if the correlations between the test statistics are known, or even if only the first few moments of the test statistic pairwise correlations are known, the quantity $\varrho(\tau_*,\tau_*)$ can be computed efficiently using the first few terms in a Hermite polynomial expansion, as seen in Theorem 2 of Schwartzman and Lin (2011). Below, we specialize Corollary 4.1 to settings with block diagonal correlations and with Toeplitz correlations. In these settings, Condition 2 will hold and $\varrho(\tau_*,\tau_*)$ will have an easily-expressed formula.

Corollary 4.2 (Block diagonal correlations). *In the setting of Corollary 4.1, suppose that $(\varepsilon_{m1},\ldots,\varepsilon_{mm})$ has a block diagonal correlation matrix with blocks of fixed size s_B in which the off diagonal correlations are ρ_B. Let τ_* be the unique $t \in (0,1)$ satisfying $t/q = \pi_0 t + \pi_1 \bar{\Phi}(\bar{\Phi}^{-1}(t) - \mu_A)$. Then*

$$\sqrt{m}\big(\mathrm{FDP}_m - \pi_0 q\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\pi_0^2 q^2}{\tau_*^2}\Big(\frac{\tau_*}{\pi_0} - \tau_*^2 + (s_B - 1)\tilde{\rho}(\tau_*,\tau_*,\rho_B)\Big)\Big) \tag{20}$$

where $\pi_0 \equiv 1 - \pi_1$ and $\tilde{\rho}$ is defined at (9).

Proof. This follows from a direct application of Corollary 4.1. □

The corollary as written requires m to be a multiple of s_B, but it extends easily to $m \to \infty$ through an arbitrary sequence of m. One can let the "last" block be smaller than the others if necessary.

Another simple correlation structure we can consider has banded Toeplitz correlation matrices for $\varepsilon_{m1},\ldots,\varepsilon_{mm}$.

Corollary 4.3 (Toeplitz correlation). *In the setting of Corollary 4.1, suppose that* $(\varepsilon_{m1},\dots,\varepsilon_{mm})$ *has a Toeplitz correlation matrix*

$$\Gamma_{ij}^{(m)} = I\{i=j\} + \sum_{\ell=1}^{M} \rho_l I\{|i-j|=l\}$$

where $\rho_1,\dots,\rho_M \in (-1,1)$ *are such that* $\Gamma^{(m)}$ *is positive semi-definite for all* $m > M$. *Let* τ_* *be the unique* $t \in (0,1)$ *satisfying* $t/q = \pi_0 t + \pi_1 \bar{\Phi}(\bar{\Phi}^{-1}(t) - \mu_A)$. *Then,*

$$\sqrt{m}\big(\mathrm{FDP}_m - \pi_0 q\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{\pi_0^2 q^2}{\tau_*^2}\left(\frac{\tau_*}{\pi_0} - \tau_*^2 + 2\sum_{\ell=1}^{M} \tilde{\rho}(\tau_*,\tau_*,\rho_\ell)\right)\right) \tag{21}$$

where $\pi_0 \equiv 1-\pi_1$ *and* $\tilde{\rho}$ *is defined at* (9).

Proof. This follows from a direct application of Corollary 4.1. □

We check the CLTs provided by Corollaries 4.2 and 4.3 via simulation in Figure 3. We compare the normal approximation given by these CLTs to the normal approximation given by Corollary 4.2 in Delattre and Roquain (2016). We simulate block and banded correlation structures that do not satisfy the sufficient conditions in their Corollary 4.2.

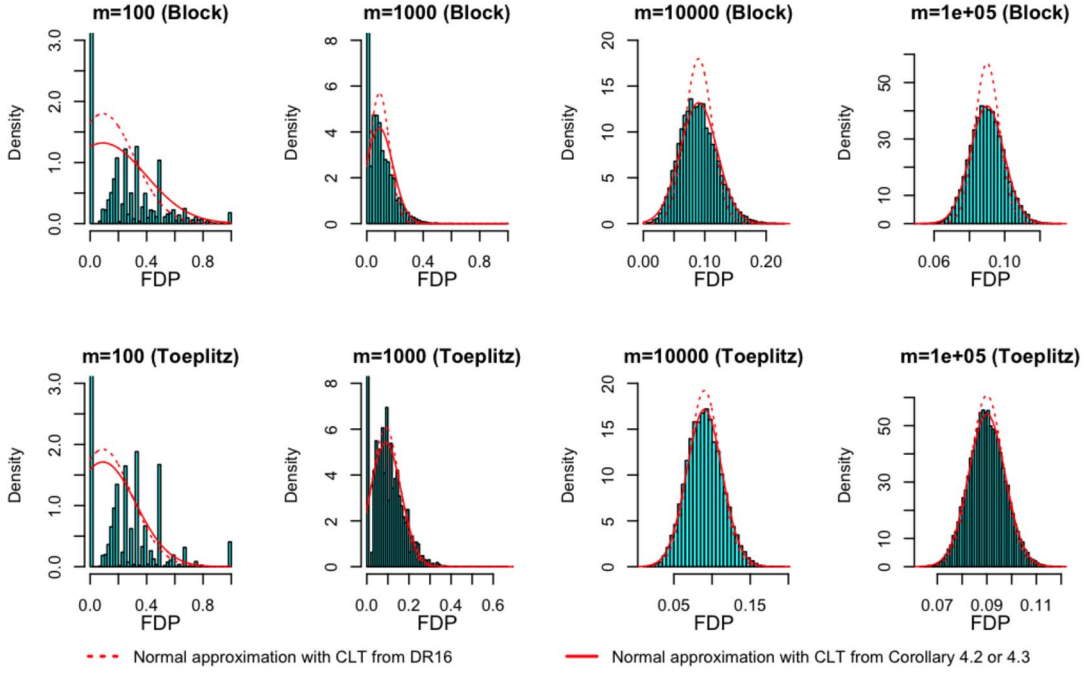

Figure 3. This figure compares the histograms of FDP to our normal approximation and an earlier one by Delattre and Roquain (2016) that does not necessarily cover these cases. Each panel is based on 25,000 Monte Carlo simulations as described in the text.

For each correlation structure and each $m \in \{10^2, 10^3, 10^4, 10^5\}$, we ran 25,000 Monte Carlo simulations with $\mu_A = 2$, $\pi_0 = 0.9$, and $q = 0.1$. The block correlation matrices we considered had block size $s_B = 20$, and within block correlations $\rho_B = 0.5$. The banded Toeplitz correlation matrix had $M = 2$ with $\rho = \rho_1 = 0.65$ above and below the diagonal and $\rho = \rho_2 = 0.3$ two rows above and below the diagonal. Our normal approximation fits very well for the larger values of m. For large m the normal approximation from Delattre and Roquain (2016) appears to accurately estimate the mean but not the variance of the FDP. From this we believe that something in their sufficient conditions must also have been necessary.

5. Burstiness in a factor model

The results in the previous section are CLTs that do not require conditioning on the latent factor as they assumed no long-range correlations modeled via a factor model. The CLTs are messier when factor model components are introduced, so we present two examples for factor model settings where the formulas for the asymptotic distribution of the FDP have some simplifications.

5.1. 1-factor model for long-range equicorrelated Gaussian noise

Suppose that for each m, $H_{m1}, \ldots, H_{mm} \overset{\text{iid}}{\sim} \text{Bern}(\pi_1)$ for a fixed $\pi_1 \in (0,1)$, but we now have a one dimensional latent factor; that is $\boldsymbol{W} \sim \mathcal{N}(0,1)$. For simplicity, we consider the simplest factor model structure: an equicorrelated Gaussian model. In particular, we let $\boldsymbol{L}_{mi} = \sqrt{\rho_1}$ where $\rho_1 \in [0,1)$ for all m,i. We will also allow for errors with shorter range correlations to be added to the model by supposing that $(\tilde{\varepsilon}_{m1}, \ldots, \tilde{\varepsilon}_{mm}) \sim \mathcal{N}(0, \Gamma^{(m)})$ where $\Gamma^{(m)}$ is a correlation matrix with blocks of size s_B and off diagonal within-block correlations of ρ_2. We assume that the blocks are of equal size, except the last one if m does not divide s_B. In this model the test statistics are $X_{mi} = \mu_A H_{mi} + \sqrt{\rho_1}\boldsymbol{W} + \sqrt{1-\rho_1}\tilde{\varepsilon}_{mi}$ and the correlation structure of the errors (not related to the indicators H_{mi} of whether the hypotheses are true) follows a matrix Σ_{B_2} where

$$(\Sigma_{B_2})_{ij} = \begin{cases} 1, & \text{if } i = j, \\ \rho_1, & \text{if } i \text{ and } j \text{ are in different blocks,} \\ \rho_1 + (1-\rho_1)\rho_2, & \text{if } i \neq j \text{ and if } i \text{ and } j \text{ are in the same block.} \end{cases}$$

In such a setting it is easy to show that all of our conditions, except possibly Condition 4 will hold. Condition 4 ruling out pathologies involving τ_* will hold depending on the value of $\boldsymbol{W}$ drawn from $\mathcal{N}(0,1)$. Some $\boldsymbol{w}$ may give $\tau_* = 0$ though we do not expect a positive probability that τ_* will be an accumulation point of $\{t : G(t) = t/q\}$.

Corollary 5.1. *In the multiple hypothesis testing setting of Section 5.1, condition on $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}$ and for $t \in (0,1)$ let $G(t) = (1-\pi_1)\gamma_0(t) + \pi_1\gamma_1(t)$ with $\gamma_r(t) = \bar{\Phi}\big((\bar{\Phi}^{-1}(t) - \mu_A r - \sqrt{\rho_1}\boldsymbol{w})/\sqrt{1-\rho_1}\big)$ for $r \in \{0,1\}$. If $\boldsymbol{w}$ is such that G satisfies Condition 4 that the Simes point is positive with no pathologies, then*

$$\sqrt{m}\Big(\text{FDP}_m - \frac{q\pi_0\gamma_0(\tau_*)}{\tau_*} \,|\, \boldsymbol{W} = \boldsymbol{w}\Big) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{L,2}) \quad \text{as } m \to \infty$$

where $\pi_0 \equiv 1 - \pi_1$ and

$$\sigma^2_{L,2} = \frac{q^2}{\tau_*^2}\Big((1+\alpha)^2 c^{(0,0)}(\tau_*,\tau_*) + \alpha^2 c^{(1,1)}(\tau_*,\tau_*) + 2\alpha(1+\alpha)c^{(1,0)}(\tau_*,\tau_*)\Big)$$

where

$$\alpha = \frac{\pi_0 \gamma_0'(\tau_*) - \pi_0 \gamma_0(\tau_*)/\tau_*}{1/q - G'(\tau_*)}$$

and for $r \in \{0, 1\}$,

$$c^{(r,r)}(\tau_*, \tau_*) = \pi_r \gamma_r(\tau_*) - \pi_r^2 \gamma_r(\tau_*)\gamma_r(\tau_*) + \pi_r^2 (s_B - 1)\tilde{\rho}(\gamma_r(\tau_*), \gamma_r(\tau_*), \rho_2)$$

and

$$c^{(1,0)}(\tau_*, \tau_*) = -\pi_0 \pi_1 \gamma_0(\tau_*)\gamma_1(\tau_*) + \pi_0 \pi_1 (s_B - 1)\tilde{\rho}(\gamma_0(\tau_*), \gamma_1(\tau_*), \rho_2)$$

where $\tilde{\rho}$ *is defined in Equation* (9).

Proof. As mentioned earlier in this section, Conditions 1, 2, 3, 5, 6, and 7 will hold for any value of $\boldsymbol{W}$ drawn. Therefore the above result holds from applying Theorem 3.1 in the setting where Condition 4 also holds. □

Remark 4. In the setting of Corollary 5.1, if m/s_B is small, then perhaps the test statistic correlations can be modeled with a factor model with a bit more than m/s_B factors but asymptotically such an approach would require adding infinitely many factors in the model. In our setup, the equicorrelations ρ_1 are long-range and persist as $m \to \infty$ and hence they are modeled with a factor model whereas the additional noise with correlation blocks of size s_B involves short-range correlations and are therefore not modeled as a factor.

In this subsection, we have demonstrated that Theorems 3.1 and 3.2 can be used to provide further insight into asymptotics of the BH procedure in the setting of Gaussian test statistics with constant pairwise correlation. In such a setting, Finner, Dickhaus and Roters (2007) find the limiting expected values of both the FDP and the FPR as functions of a one-dimensional latent factor. We have extended their results by deriving the limiting distribution of the FDP as a function of a one-dimensional latent factor (the limiting distribution of the FPR can similarly be derived from Theorem 3.2). We also considered a more general setting than the equicorrelated Gaussian model in order to exhibit that Theorems 3.1 and 3.2 can handle settings with both short and long-range correlations simultaneously.

5.2. Setting where number of nonnulls is $o_p(\sqrt{m})$

Here we consider sparse nonnulls with $\pi_1^{(m)} = o(1/\sqrt{m})$. It can be shown with a Chernoff bound for the binomial that in this case the number of nonnulls is $o_p(\sqrt{m})$. Also suppose that under the model for test statistics of Section 2.1, Conditions 1–7 hold. Then it will follow that $F_1 = 0$ and moreover $W_{m,r} = \sqrt{m}(\hat{F}_{m,1} - F_{m,1}) \xrightarrow{p} 0$. If this is the case, then we will have $c^{(1,1)} = 0$, $c^{(1,0)} = 0$, $F_0(\tau_*) = \tau_*/q$ and $\alpha = -1$.

Corollary 5.2. *Suppose that we are in the multiple testing setting of Section 2.1 and that* $\pi_1^{(m)} = o(1/\sqrt{m})$. *If, conditionally on a specific value of the latent factor* $\boldsymbol{W} = \boldsymbol{w} \in \mathbb{R}^k$, *Conditions 1–7 hold, then*

$$\sqrt{m}\Big(\mathrm{FDP}_m - 1 \,|\, \boldsymbol{W} = \boldsymbol{w}\Big) \xrightarrow{p} 0$$

and

$$\sqrt{m}\Big(\frac{V_m}{m} - \frac{\tau_*}{q} \,|\, \boldsymbol{W} = \boldsymbol{w}\Big) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{R,3}) \quad \text{as } m \to \infty$$

where

$$\sigma^2_{R,3} \equiv (1 - qG'(\tau_*))^{-2} c^{(0,0)}(\tau_*, \tau_*).$$

Proof. Apply Theorems 3.1 and 3.2 noting that $F_0(\tau_*) = \tau_*/q$, $\alpha = -1$, $c^{(1,1)} = 0$, $c^{(1,0)} = 0$. □

The corollary indicates that severe bursts can occur; V_m/m can converge to a positive number even while the proportion of hypotheses that are nonnull converges to 0.

We check the result of Corollary 5.2 via simulation in Figure 4. We simulate from the 1-factor model described in Section 5.1, except now the proportion of nonnulls $\pi_1^{(m)}$ is not fixed in m. Instead we set $\pi_1^{(m)} = 5m^{-2/3}$. For each $m \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$, we conditioned on $\boldsymbol{w} = 2.5$ and ran 25,000 Monte Carlo simulations with $\mu_A = 2$, $\pi_0^{(m)} = 1 - 5m^{-2/3}$, $q = 0.1$, $\rho_1 = 0.3$, and $\rho_2 = 0.6$. For these choices of parameters, the conditions of Corollary 5.2 are met.

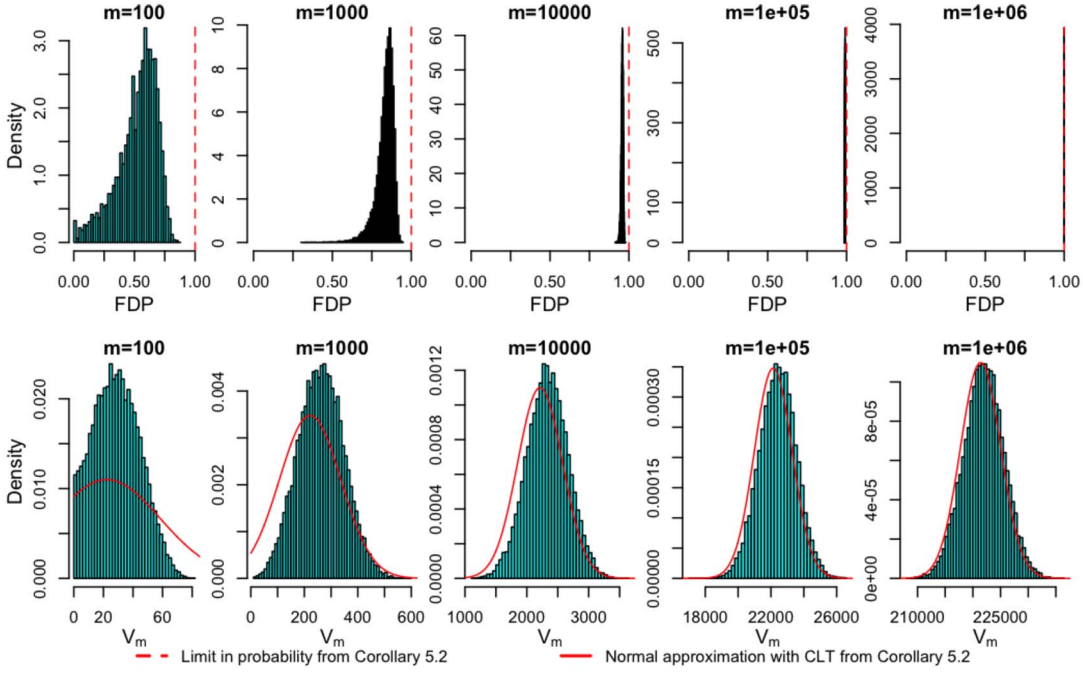

Figure 4. This figure compares the histograms of the FDP and the number of discoveries V_m to the asymptotic estimates of their distributions given by Corollary 5.2. Each panel is based on 25,000 Monte Carlo simulations as described in the text.

6. Data driven factor model example

We fit a 3-factor model to the GDS 3027 Duchenne Muscular Dystrophy (DMD) data, which can be found on the Gene Expression Omnibus. This data set was analyzed in Kotelnikova et al. (2012) and Wang et al. (2022) and had $m = 22{,}283$ genes and $n = 37$ subjects. Of these subjects 23 had DMD and 14 did not. We centered the data for each gene and stored it in a matrix $Y \in \mathbb{R}^{37\times 22{,}283}$. To fit a homoskedastic factor model, we looked at the plot of the singular values of Y and chose to work with the largest three of them for illustrative purposes. We then computed Y_3, the singular value decomposition-based rank 3 approximation to Y, and estimated the homoskedastic noise, σ_E as the standard deviation of the entries in $Y - Y_3$. We let $\tilde{L} \in \mathbb{R}^{m\times 3}$ be the matrix whose columns consist of the first 3 right singular vectors of Y scaled by their corresponding singular values. We subsequently treat $\tilde{L}$ and σ_E as fixed quantities and then assume the following factor model under the global null: $Y^\mathsf{T} = \tilde{L}F + \sigma_E E$, where the entries of $F \in \mathbb{R}^{3\times 37}$ and $E \in \mathbb{R}^{m\times 37}$ are IID standard Gaussians. Under the alternative we suppose that for each nonnull gene, the values of Y for that gene are shifted by a fixed constant for DMD subjects and a different fixed constant, maintaining centering of the columns of Y, for the control subjects.

Since the dataset is from a case-control study, to compute the test statistics we condition on DMD status and assume that the stochasticity in our observations Y comes from the random matrices E and F. The unstandardized test statistics $X_{\text{ustd}} \in \mathbb{R}^m$ are simply the difference-in-means between the DMD group and the control group for each gene and this unstandardized test statistics vector has covariance matrix proportional to $\sigma_E^2 I_m + \tilde{L}\tilde{L}^\mathsf{T}$. The standardized test statistics $X \in \mathbb{R}^m$ are given by dividing each entry of X_{ustd} by the squareroot of the corresponding diagonal entry of $\sigma_E^2 I_m + \tilde{L}\tilde{L}^\mathsf{T}$. The vector of test statistics then satisfy $X = \vec{\mu} + \boldsymbol{L}\boldsymbol{W} + \varepsilon$ where $\vec{\mu}$ is a vector of constant means (which are zero for the null genes), $\boldsymbol{L}$ is a matrix of factor loadings similar to $\tilde{L}$ but with appropriately rescaled rows, $\boldsymbol{W} \sim \mathcal{N}(0, I_3)$ and ε is heteroskedastic, independent, and centered Gaussian noise. Assuming that each standardized nonnull test statistic has the same mean $\mu_A > 0$, that we conduct one-sided testing, and that the nonnulls are determined by IID Bern(π_1) draws, using the test statistics X we are in the multiple hypothesis testing setting of Section 2.1.

Figure 2 in Section 2 shows the asymptotic ECDF of the p-values for three specific realizations of the latent factor $\boldsymbol{w}$ in the data-driven 3-factor model and multiple testing setting described above, with $\mu_A = 2$, $\pi_1 = 0.1$, and $q = 0.1$. Figure 5 shows histograms of the FDP based on 25,000 Monte Carlo simulations for the same data-driven 3-factor model, multiple testing setup, factor outcomes, and parameters as Figure 2.

In the top panel of Figure 5, $m = 22{,}283$ as is the case in the original 3-factor model fit to the GDS 3027 dataset. In the bottom panel, to increase the number of tests and check asymptotic behavior, we copy each row of factor loadings in the original factor model 25 times to get a distribution of the FDP when $m = 22{,}283 \times 25$ tests are conducted. That is much larger than we would need for gene expression and approaches the range we would encounter for SNPs. In case A, the CLT is reasonable for the larger but not the smaller sample size. The CLT fits well for both sample sizes for case C. In case B, the sufficient conditions for the conditional CLT do not hold and Theorem 3.3 holds instead.

This simulation shows some bursty behavior for BH as follows. Cases A and C are both covered by the conditional CLT and there we see that even in cases covered by the conditional CLT, the FDP can vary greatly, being nearly Gaussian with means varying by nearly 100-fold. When cases like case B arise there is no conditional CLT, and by Theorem 3.3, the BH rejection threshold converges to 0 in probability. In case B, we observe a very heavy tail to the FDP distribution, although fewer than 1 out of every 5,000 Monte Carlo simulations yields a nonzero FDP, and no simulation yields more than 1 false discovery. In conclusion, the simulations are consistent with the results of Theorems 3.1 and 3.3.

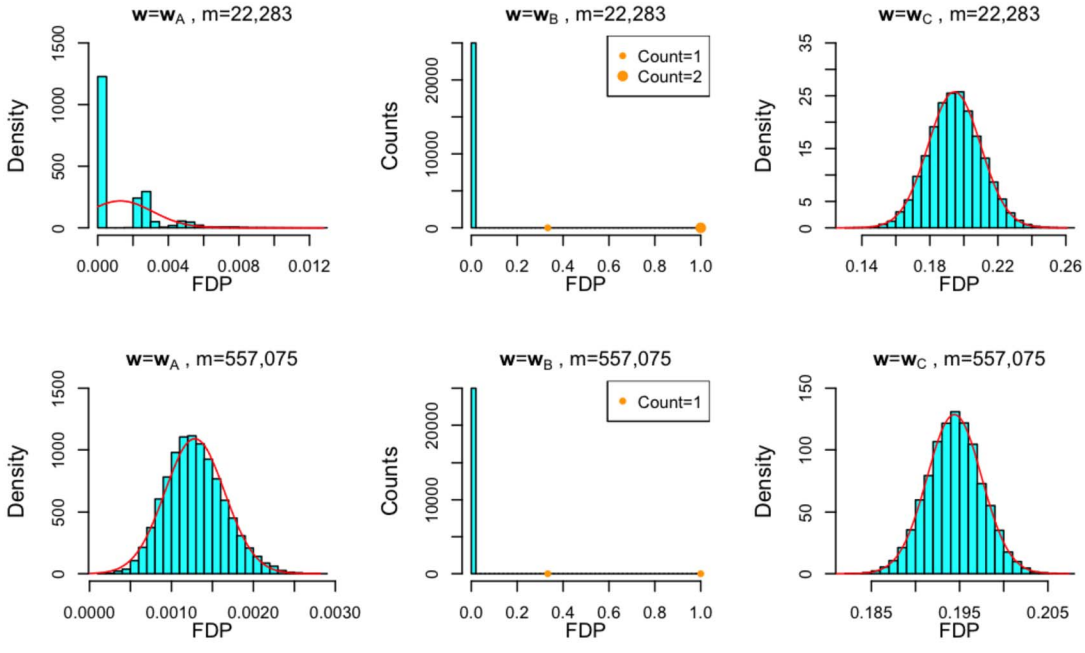

Figure 5. The histograms show 25,000 samples of the number of false discoveries in the two group model with a data-driven three factor model for dependence. The 3-factor model has draws A, B and C as described in the text. The top row has $m = 22{,}283$ hypotheses and the bottom has 25 times as many hypotheses. The nominal FDR control threshold is $q = 0.1$. Draw C yields a higher FDP while draw A yields a lower one. In draw B, nearly all of the 25,000 Monte Carlo simulations yielded an FDP = 0, and the small number of simulations for which FDP $\neq 0$ are denoted by orange circles for visibility. The red curves for A and C are asymptotic Gaussians from Theorem 3.1. Case B does not satisfy the sufficient conditions for the conditional CLT, but satisfies the conditions for Theorem 3.3.

A large FDP tail is not necessarily indicative of alarming bursty behavior for BH, as the FDP can be equal to 1 in scenarios where there is only one false discovery. Looking at the FDP simultaneously with the number of false discoveries V_m gives a clearer sense of whether the bursty behavior is alarming. In Figure 6, we run 25,000 Monte Carlo simulations using the previously described data-driven 3-factor model. In contrast to Figure 5, we do not condition on specific realizations of the latent factor $\boldsymbol{w}$ and we also plot the joint distribution of the FDP and the number of false discoveries rather than the marginal distribution of the FDP. In the simulations, we set the FDR control parameter $q = 0.1$ and repeat the simulations for the nonnull effect size $\mu_A \in \{2,4,6\}$ and for Bernoulli mixture null parameter $\pi_0 \in \{0.9, 0.99, 0.999\}$.

Remark 5. Controlling pFDR using Storey's q-value is another popular multiple testing approach that is heralded for avoiding floods of false positives (Storey and Tibshirani (2003)). Our simulations show that when the nonnull effect sizes are small and the number of nonnulls is sparse the pFDR will be high, implying that controlling the pFDR would mitigate the issue of burstiness in such cases. When there are many nonnulls or when the nonnull effect sizes are large, the pFDR is nearly equal to the FDR (due to few simulations with no discoveries), implying that controlling the pFDR would not mitigate burstiness in such cases.

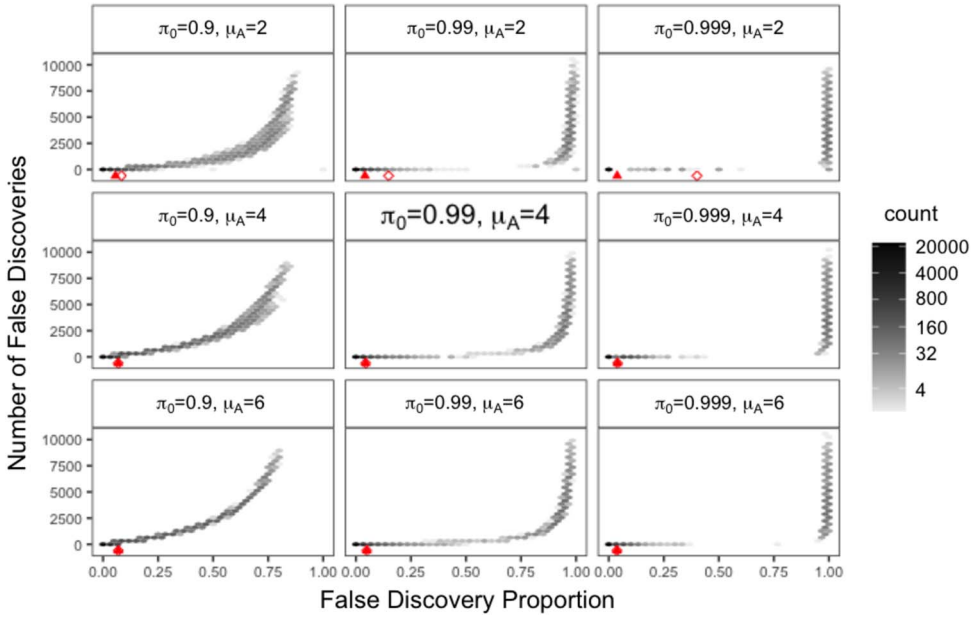

Figure 6. The hexagonal-binned heatmaps each show the results of 25,000 Monte Carlo simulations for the data-driven 3-factor model described in the text, using various nonnull effect sizes μ_A and proportions of nulls π_0. In each simulation, the BH method is applied at level $q = 0.1$ to $m = 22{,}283$ test statistics. For each plot, the mean of the FDP is marked with a solid triangle and the mean amongst nonzero FDP values is marked with a hollow diamond, which estimate the FDR and pFDR respectively. For the plots with $\mu_A \in \{4, 6\}$, the triangle and diamond are close enough to overlap.

Remark 6. For the data-driven 3-factor model with $m = 22{,}283$, the distribution of FDP_m is largely driven by the realization of the latent factor $\boldsymbol{w}$. This can be seen in the top panel of Figure 5, and we further quantified this by running 1,000 Monte Carlo simulations from the model for each of 1,000 different randomly generated latent factor vectors. The variance of the FDP between groups with different latent factors was approximately 325 times larger than the average variance within each latent factor vector group.

7. Discussion

Here we discuss the conclusions that can be drawn from our theorems and simulations about when BH exhibits alarming burstiness and when BH is safe from burstiness concerns. We end with a discussion of the relevance and feasibility of factor model-based corrections for addressing burstiness concerns.

Burstiness occurs when there are many strong, long-range correlations between the test statistics. When we model the long-range correlations via a factor model, this phenomenon can be explained by Theorem 3.1. By Theorem 3.1, the asymptotic limit of $\mathrm{FDP}_m \,|\, \boldsymbol{W}$ is $qF_0(\tau_*)/\tau_*$, a quantity that can vary drastically for different realizations of $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$. The variation in $qF_0(\tau_*)/\tau_*$ is greater when the long-range correlations are stronger (or equivalently, when the factor model loading vectors $\boldsymbol{L}_{mi}$ have larger magnitude). Therefore, the FDP has high variability when there are many strong, long-range correlations between the test statistics. Meanwhile, by Theorem 3.2, there could be a flood of false discoveries, making the bursts severe. Our simulations from the 3-factor model fit to the DMD dataset

indicate a wide right tail of the FDP distribution as well as severe bursts (see the top-left panel of Figure 1 and Figure 6). Notably, we find that sparsity of the number of nonnulls exacerbates burstiness issues. This can be explained by Corollary 5.2 and is observed in Figures 4 and 6.

Conversely, our theorems and simulations indicate that there are many settings where the test statistics are correlated, but the BH procedure is free of burstiness concerns. When there are no long-range correlations, no factor model is needed to model the correlations, so the variance of the FDP will decrease rapidly as the number of tests increases, even when the short-range correlations are strong. For example, in the setting of Corollary 4.1, FDP_m converges to a quantity less than the desired FDR control q and has variance of order $1/m$, even with strong short-range correlations. The simulations in the bottom of Figure 1 and all the panels of Figure 3 involve strong short-range correlations and still demonstrate this desirable behavior (the desirable behavior is not seen in the panels of Figure 3 where m is small because, in that case, the "short-range" correlations are actually long-range relative to the number of tests). Even when there are long-range correlations but the long-range correlations are weak, modeled by a factor model with loading vectors $\boldsymbol{L}_{mi}$ of small magnitude, the BH procedure will not exhibit worrisome bursts. With loading vectors $\boldsymbol{L}_{mi}$ with small magnitude, the γ_{mir} terms will not be sensitive to the realized value of $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$, and in turn F_0, F_1, τ_*, and $qF_0(\tau_*)/\tau_*$ will also not be so sensitive to the realized value of $\boldsymbol{W}$. Therefore, when the loading vectors $\boldsymbol{L}_{mi}$ have small magnitude, the asymptotic limit of $\mathrm{FDP}_m \,|\, \boldsymbol{W}$ given in Theorem 3.1 will not oscillate much as $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$ varies. Indeed, in Figure 1, when the long-range correlations are all reduced by a factor of 10 (as we move from the top-left panel to the top-right panel), alarming burstiness is no longer observed.

These results suggest that estimating the correlation structure with a factor model can be useful for identifying whether or not burstiness is a concern in particular applications. The results further suggest that methods which estimate and remove the factor model components from the test statistics prior to applying BH (e.g. methods that estimate $\boldsymbol{W}$ and $(\boldsymbol{L}_{mi})_{i=1}^m$, subtract $\boldsymbol{L}_{mi}^\mathsf{T}\boldsymbol{W}$ from each test statistic, and subsequently apply BH) can alleviate burstiness issues. A number of such approaches for estimating and removing factor model components in multiple testing settings have been proposed and have shown promise in simulations (Fan, Han and Gu, 2012, Fan et al., 2019, Friguet, Kloareg and Causeur, 2009, Sun, Zhang and Owen, 2012, Wang et al., 2017).

While a full discussion of these recent methods is out of scope for this paper, we briefly note that there are two major challenges with estimation and removal of factor model components in a multiple testing setting. First, it is possible that removing the factor model components from the test statistics might remove some of the important signal that one is trying to detect with hypothesis testing. For example, this can happen if a large collection of genes is associated with one of the leading factors, yet at the same time, that collection of genes is also associated with the outcome variable. To avoid such issues, methods which remove the factor model components often rely upon an assumption that the number of nonnulls is sparse. Second, estimating the underlying factor model for a dataset is statistically challenging. It is difficult to estimate the latent factors $\boldsymbol{W}$ and the factor loadings $(\boldsymbol{L}_{mi})_{i=1}^m$ well without a large number of samples, and even choosing the number of latent factors k is a difficult task. For a comparison of methods for estimating the number of latent factors, see Owen and Wang (2016).

Acknowledgements

The authors wish to thank Will Fithian, Kevin Guo, Grant Izmirlian, Lihua Lei, Kenneth Tay, Marius Tirlea, Jingshu Wang, and anonymous reviewers for helpful comments and discussions. The authors also thank Kevin Guo for a proof sketch on how to remove a condition from an earlier version of this paper and Jingshu Wang for sharing the DMD data. Finally, the authors would like to thank two anonymous referees, an Associate Editor and the Editor for their constructive comments that helped improve the quality of this paper.

Funding

DMK was supported by a Stanford Graduate Fellowship and a Stanford Graduate Interdisciplinary Fellowship. ABO was supported by the National Science Foundation under grants IIS-1837931 and DMS-2152780.

Supplementary Material

Proofs of Theorems 3.1, 3.2, and 3.3 (DOI: 10.3150/23-BEJ1615SUPP; .pdf). A supplement with the proofs of Theorems 3.1, 3.2, and 3.3 can be found in the online supplemental material (Kluger and Owen, 2024). The first section of the supplement restates definitions and theorems from the probability literature that we use in our proofs. The second section of the supplement provides proofs of Theorems 3.1, 3.2, and 3.3. The third section of the supplement exhibits Hadamard derivative calculations that are used in our proofs of Theorems 3.1 and 3.2.

References

Andrews, D.W.K. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *Int. Stat. Rev.* **62** 119–132.

Azriel, D. and Schwartzman, A. (2015). The empirical distribution of a large number of correlated normal variables. *J. Amer. Statist. Assoc.* **110** 1217–1228. MR3420696 https://doi.org/10.1080/01621459.2014.958156

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25** 60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

Chi, Z. (2007). On the performance of FDR control: Constraints and a partial solution. *Ann. Statist.* **35** 1409–1431. MR2351091 https://doi.org/10.1214/009053607000000037

Delattre, S. and Roquain, E. (2011). On the false discovery proportion convergence under Gaussian equi-correlation. *Statist. Probab. Lett.* **81** 111–115. MR2740072 https://doi.org/10.1016/j.spl.2010.09.025

Delattre, S. and Roquain, E. (2015). New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *Ann. Statist.* **43** 1141–1177. MR3346700 https://doi.org/10.1214/14-AOS1302

Delattre, S. and Roquain, E. (2016). On empirical distribution function of high-dimensional Gaussian vector components with an application to multiple testing. *Bernoulli* **22** 302–324. MR3449784 https://doi.org/10.3150/14-BEJ659

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302 https://doi.org/10.1198/016214506000001211

Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478

Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *J. Amer. Statist. Assoc.* **114** 1880–1893. MR4047307 https://doi.org/10.1080/01621459.2018.1527700

Farcomeni, A. (2006). More powerful control of the false discovery rate under dependence. *Stat. Methods Appl.* **15** 43–73. MR2281214 https://doi.org/10.1007/s10260-006-0002-z

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scand. J. Stat.* **34** 275–297. MR2346640 https://doi.org/10.1111/j.1467-9469.2006.00530.x

Finner, H., Dickhaus, T. and Roters, M. (2007). Dependency and false discovery rate: Asymptotics. *Ann. Statist.* **35** 1432–1455. MR2351092 https://doi.org/10.1214/009053607000000046

Finner, H. and Roters, M. (2001). On the false discovery rate and expected type I errors. *Biom. J.* **43** 985–1005. MR1878272 https://doi.org/10.1002/1521-4036(200112)43:8<985::AID-BIMJ985>3.0.CO;2-4

Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.* **30** 220–238. MR1892662 https://doi.org/10.1214/aos/1015362191

Fithian, W. and Lei, L. (2022). Conditional calibration for false discovery rate control under dependence. *Ann. Statist.* **50** 3091–3118. MR4524490 https://doi.org/10.1214/21-aos2137

Friguet, C., Kloareg, M. and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571 https://doi.org/10.1198/jasa.2009.tm08332

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197 https://doi.org/10.1214/009053604000000283

Gerard, D. and Stephens, M. (2020). Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics* **21** 15–32. MR4043843 https://doi.org/10.1093/biostatistics/kxy029

Gontscharuk, V. and Finner, H. (2013). Asymptotic FDR control under weak dependence: A counterexample. *Statist. Probab. Lett.* **83** 1888–1893. MR3069893 https://doi.org/10.1016/j.spl.2013.04.025

Izmirlian, G. (2020). Strong consistency and asymptotic normality for quantities related to the Benjamini-Hochberg false discovery rate procedure. *Statist. Probab. Lett.* **160** 108713, 10. MR4061847 https://doi.org/10.1016/j.spl.2020.108713

Kim, K.I. and van de Wiel, M.A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinform.* **9** 114. https://doi.org/10.1186/1471-2105-9-114

Kluger, D.M. and Owen, A.B. (2024). Supplement to "A central limit theorem for the Benjamini-Hochberg false discovery proportion under a factor model." https://doi.org/10.3150/23-BEJ1615SUPP

Korn, E.L., Troendle, J.F., McShane, L.M. and Simon, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124** 379–398. MR2080371 https://doi.org/10.1016/S0378-3758(03)00211-8

Kotelnikova, E., Shkrob, M.A., Pyatnitskiy, M.A., Ferlini, A. and Daraselia, N. (2012). Novel approach to meta-analysis of microarray datasets reveals muscle remodeling-related drug targets and biomarkers in Duchenne muscular dystrophy. *PLoS Comput. Biol.* **8** 1–10. https://doi.org/10.1371/journal.pcbi.1002365

Lindner, A.M. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p, q)–processes. In *Handbook of Financial Time Series* 43–69. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71297-8\protect \T1\textunderscore 2

Lucas, J.E., Kung, H.N. and Chi, J.T.A. (2010). Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers. *PLoS Comput. Biol.* **6** e100920:1–15.

Mokkadem, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl.* **29** 309–315. MR0958507 https://doi.org/10.1016/0304-4149(88)90045-2

Neumann, M.H. (2013). A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics. *ESAIM Probab. Stat.* **17** 120–134. MR3021312 https://doi.org/10.1051/ps/2011144

Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Stat.* **2** 1065–1110. MR2460858 https://doi.org/10.1214/08-EJS207

Owen, A.B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 411–426. MR2155346 https://doi.org/10.1111/j.1467-9868.2005.00509.x

Owen, A.B. and Wang, J. (2016). Bi-cross-validation for factor analysis. *Statist. Sci.* **31** 119–139. MR3458596 https://doi.org/10.1214/15-STS539

Paolella, M.S. (2019). *Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH*. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR3839332

Romano, J.P. and Shaikh, A.M. (2006). On stepdown control of the false discovery proportion. In *Optimality*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **49** 33–50. Beachwood, OH: IMS. MR2337829 https://doi.org/10.1214/074921706000000383

Romano, J.P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35** 1378–1408. MR2351090 https://doi.org/10.1214/009053606000001622

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika* **98** 199–214. MR2804220 https://doi.org/10.1093/biomet/asq075

Storey, J.D., Taylor, J.E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766 https://doi.org/10.1111/j.1467-9868.2004.00439.x

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856 https://doi.org/10.1073/pnas.1530509100

Sun, Y., Zhang, N.R. and Owen, A.B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. MR3058679 https://doi.org/10.1214/12-AOAS561

van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

Wang, J., Zhao, Q., Hastie, T. and Owen, A.B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* **45** 1863–1894. MR3718155 https://doi.org/10.1214/16-AOS1511

Wang, J., Gui, L., Su, W.J., Sabatti, C. and Owen, A.B. (2022). Detecting multiple replicating signals using adaptive filtering procedures. *Ann. Statist.* **50** 1890–1909. MR4474476 https://doi.org/10.1214/21-aos2139

Zhang, C., Fan, J. and Yu, T. (2011). Multiple testing via FDR_L for large-scale imaging data. *Ann. Statist.* **39** 613–642. MR2797858 https://doi.org/10.1214/10-AOS848

Received April 2022 and revised December 2022