The Benjamini–Hochberg Procedure Can Fail to Control the FDR for Correlated Two-Sided Gaussian Tests

Edgar Dobriban*

July 14, 2026

Abstract

We show that the Benjamini–Hochberg procedure can fail to control the false discovery rate (FDR) at its nominal level for correlated two-sided Gaussian p-values. We construct a factor model for which, at level $\alpha = 0.01$, a rigorous interval-arithmetic certificate proves $\mathrm{FDR} > 0.0104$ for all sufficiently large numbers of hypotheses. This disproves a conjecture widely believed to be true for twenty years. Monte Carlo experiments are consistent with the theoretical result. The proof was obtained by GPT-5.6 Pro and carefully checked by the author.

1 The conjecture and the counterexample

Multiple hypothesis testing is a central topic in modern statistics, with applications in biology, genomics, astronomy, economics, finance, and many other fields. Over the last three decades, controlling the *false discovery rate* (FDR) (Benjamini and Hochberg, 1995) has become a standard target in applications involving many tests.

The Benjamini–Hochberg (BH) procedure is the standard method for FDR control. It controls the FDR when the p-values are independent (Benjamini and Hochberg, 1995), or when they satisfy the weaker positive regression dependence condition of Benjamini and Yekutieli (2001). See also Sarkar (2002); Blanchard and Roquain (2008); Finner et al. (2007) and the references below.

A crucial setting that had remained unresolved is that of correlated two-sided Gaussian tests. This setting matters because many data sets involve correlated tests; for example, neighboring genes or genetic variants can be correlated. Two-sided tests are also common because the direction of an effect is often unknown in advance, requiring simultaneous sensitivity to positive and negative effects.

In this paper we show, contrary to prior conjectures and supporting evidence (e.g., Reiner-Benaim, 2007; Benjamini, 2010), that the Benjamini–Hochberg procedure need not control the FDR at its nominal level for correlated two-sided Gaussian tests.

*Department of Statistics and Data Science, University of Pennsylvania. E-mail address: dobriban@wharton.upenn.edu.

More specifically, we consider the following setting. Let $X \sim \mathcal{N}_m(\mu, \Sigma)$, where Σ is a correlation matrix. For $1 \le i \le m$, form the two-sided Gaussian p-value $P_i = 2\overline{\Phi}(|X_i|)$, where Φ is the standard normal CDF and $\overline{\Phi} = 1 - \Phi$. The Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg, 1995) at level $\alpha \in (0,1)$ uses critical values $t_r = \alpha r/m$. Let $R = \max\Big\{r : \#\{i : P_i \le t_r\} \ge r\Big\}$, with $R = 0$ if the set is empty. The BH procedure rejects those hypotheses $H_i : \mu_i = 0$ with $P_i \le t_R$. If $I_0 = \{i : \mu_i = 0\}$ denotes the true-null index set, define the number of false rejections by $V = \#\{i \in I_0 : P_i \le t_R\}$. The false discovery proportion (FDP), the fraction of rejected hypotheses that are true nulls, is $\mathrm{FDP} = V/\max(R, 1)$, and the false discovery rate is its mean, $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$.

A key question about the BH procedure is the following:

Does the Benjamini–Hochberg procedure control the false discovery rate at level α, i.e., do we have $\mathrm{FDR} \le \alpha$, for every m, every mean vector, every correlation matrix, and every $\alpha \in (0,1)$?

Before this work, a positive answer was widely believed. Farcomeni (2006); Reiner-Benaim (2007) provided empirical and theoretical evidence in support of FDR control, and Kim and van de Wiel (2008) provided additional evidence through extensive simulations. In an influential review, Benjamini (2010) wrote that "*convincing simutheoretical evidence indicates that [FDR control] holds for two-sided z-tests with any correlation structure.*" This motivates referring to the positive answer as a conjecture.

More recently, Sarkar (2023) wrote that "*The answer to this question is generally believed to be yes, and is conjectured so in the literature since results of numerical studies investigating the question and reported in numerous papers strongly support it.*" Sarkar (2023) continued that "*proving this conjecture [. . .] seems an urgent and important undertaking.*" Similar statements were made by Sarkar and Zhang (2025) and Ghosh and Sarkar (2025).

In this paper we show that the conjecture fails, by constructing the following Gaussian factor model. Let Z, $(\varepsilon_i)_{i\ge1}$, $(\eta_j)_{j\ge1}$, $(\xi_k)_{k\ge1}$ be mutually independent standard normal random variables. For a fixed N, define three coordinate blocks by

$$X_i^{(0)} = \frac{3}{10}Z + \frac{\sqrt{91}}{10}\varepsilon_i, \qquad 1 \le i \le 96N, \tag{1}$$

$$X_j^{(1)} = \frac{12}{5} - \frac{3}{10}Z + \frac{\sqrt{91}}{10}\eta_j, \qquad 1 \le j \le N, \tag{2}$$

$$X_k^{(2)} = \frac{22}{5} - \frac{18}{25}Z + \frac{\sqrt{301}}{25}\xi_k, \qquad 1 \le k \le 3N. \tag{3}$$

The block in (1) contains the $96N$ true nulls. The other two blocks contain the $4N$ nonnulls. The null block moves with Z, whereas both signal blocks move against it, at two different strengths.

Theorem 1 (The Benjamini–Hochberg Procedure Can Fail to Control the FDR for Correlated Two-Sided Gaussian Tests)**.** *For each integer $N \ge 1$, let $m_N = 100N$, let the first $96N$ hypotheses be true nulls, and let FDR_N denote the FDR in the model above. Then the Benjamini–Hochberg procedure at level $\alpha = 0.01$ satisfies, for all sufficiently large N,*

$$\mathrm{FDR}_N > 0.0104 > \alpha.$$

Consequently, the conjecture is false.

Proof. The construction and all analytic details occupy the remainder of the paper. The final strict numerical inequality is established by the complete outward-rounded certificate in Appendix B. □

AI usage. The proof was obtained by GPT-5.6 Pro. The model was asked directly to prove or disprove the conjecture and was provided only with the mathematical definition of the Benjamini–Hochberg procedure. After about 90 minutes of reasoning, the model produced a proof, an example, and code for the numerical certificate, which form the basis of this paper.[1] The author carefully checked the entire argument and the associated numerical certificate. Subsequently, the author asked the model to provide additional simulations, related work, and illustrations for a paper draft, and wrote the final version by editing the AI-generated draft.

Thus, this work falls into a line of work where AI models have helped professional mathematical scientists resolve open problems, see e.g., Feldman and Karbasi (2025); Jang and Ryu (2025); Salim (2025); Bubeck et al. (2025); Alexeev et al. (2025); Alexeev and Mixon (2025); Dobriban (2025); Abouzaid et al. (2026); OpenAI (2026b); Wang (2026); OpenAI (2026a), etc.

2 Relation to existing FDR analysis

The conjecture lies between two classical regimes. Under mutual independence of the null p-values and independence from the nonnull p-values, ordinary BH satisfies the sharper bound $\mathrm{FDR} \le \pi_0 \alpha$, where $\pi_0 = |I_0|/m$ (Benjamini and Hochberg, 1995). The same bound holds under positive regression dependence on the subset of true nulls (PRDS) (Benjamini and Yekutieli, 2001; Sarkar, 2002; Blanchard and Roquain, 2008).

By contrast, under completely arbitrary dependence the universal finite-sample guarantee for the unmodified linear step-up rule carries a harmonic inflation factor; replacing α by α/H_m, where $H_m = \sum_{r=1}^{m} r^{-1}$, restores level-$\alpha$ control (Benjamini and Yekutieli, 2001).

For one-sided Gaussian tests, nonnegative correlations provide an important PRDS setting (Benjamini and Yekutieli, 2001; Sarkar, 2002). The two-sided transformation is qualitatively different: $\{P_i \le t\} = \{X_i \ge c_t\} \cup \{X_i \le -c_t\}$ folds together two tails that induce opposite conditional shifts in correlated coordinates. Consequently, the usual monotone-regression and total-positivity arguments used for one-sided statistics do not automatically transfer to the folded Gaussian vector. This obstruction is discussed explicitly by Sarkar and Zhang (2025) and Ghosh and Sarkar (2025).

Earlier work of Reiner-Benaim (2007) combined low-dimensional analysis, upper bounds, and simulation evidence, and Benjamini (2010) described the evidence for arbitrary-correlation control as "simutheoretical" while noting that a complete proof was unavailable. The later literature therefore developed dependence-adjusted or shifted alternatives with provable control rather than a proof for ordinary BH (Fithian and Lei, 2022; Sarkar, 2023; Sarkar and Zhang, 2025; Ghosh and Sarkar, 2025).

[1]The conversation is available as a shared ChatGPT conversation.

Our argument uses a classical empirical CDF crossing representation of BH. In independent mixture models, and in several dependent asymptotic regimes, the random BH threshold is compared with the crossing of a limiting p-value CDF and the line t/α (Genovese and Wasserman, 2002, 2004; Storey et al., 2004; Finner et al., 2007).

An innovation in our setting is the construction of a specific Gaussian factor model. Conditioning on a single latent factor produces three independent within-block empirical processes, but leaves a *random* limiting CDF G_Z. The loadings and nonnull means are tuned so that, on a set of latent-factor values of positive Gaussian probability, the nonnull blocks enlarge the BH rejection count at precisely the same time that the conditional null distribution has heavier two-sided tails. This creates a self-consistent threshold with an FDP slightly above α.

A second innovation is that the argument avoids requiring a unique limiting crossing, differentiability at a crossing, or convergence of the BH threshold to an explicitly solved root. Two strict sign conditions merely bracket the threshold. Monotonicity of Gaussian tail probabilities then turns the continuum of (z, c) values into finitely many rational rectangles, and outward-rounded ball arithmetic verifies the required signs. This combination of a one-sided threshold bracketing and a finite interval certificate is the distinctive mechanism of the proof.

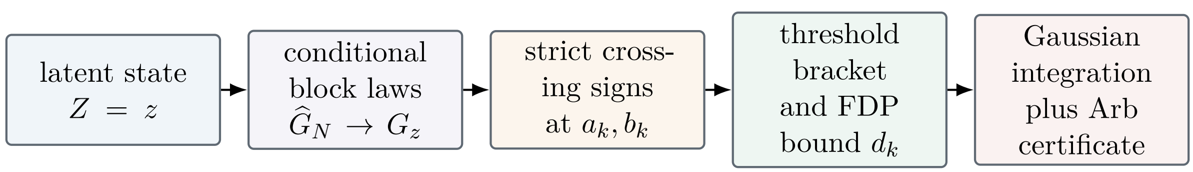

Figure 1: Proof architecture. The latent factor is not averaged out at the start; it indexes a family of deterministic limiting BH problems whose pointwise lower bounds are integrated only at the end.

3 Lemmas for the asymptotic BH threshold

In the remaining sections we present the argument of the proof. The proof is transparent in terms of empirical p-value distribution functions. Interpreting BH as the rightmost crossing of an empirical CDF and the line t/α is standard in the asymptotic FDR literature (Genovese and Wasserman, 2002, 2004; Storey et al., 2004). The following lemma isolates the weaker one-sided bracketing fact that will be needed here.

For a sample of size M_N, let

$$\widehat{G}_N(t) = \frac{1}{M_N} \sum_{i=1}^{M_N} \mathbf{1}\{P_{i,N} \le t\}, \qquad 0 \le t \le 1.$$

Let the BH grid be $\mathcal{T}_N = \left\{ \frac{\alpha r}{M_N} : 1 \le r \le M_N \right\}$, and define the BH p-value threshold

$$\tau_N = \max \left(\left\{ t \in \mathcal{T}_N : \widehat{G}_N(t) \ge \frac{t}{\alpha} \right\} \cup \{0\} \right).$$

This is exactly $\tau_N = \alpha R_N/M_N$: at the grid point $t = \alpha r/M_N$, the inequality $\widehat{G}_N(t) \geq t/\alpha$ is equivalent to having at least r p-values below the rth BH critical value. The argument will leverage the two lemmas below, whose proofs are presented in the appendix.

Lemma 2 (Two strict sign conditions bracket the BH threshold)**.** *Fix $\alpha \in (0,1)$. Suppose $M_N \to \infty$ and, for a continuous distribution function G on $[0,1]$, $\|\widehat{G}_N - G\|_\infty \longrightarrow 0$. Let $0 < v \leq w \leq \alpha$. If*

$$G(t) < \frac{t}{\alpha} \quad \textit{for every } t \in [w, \alpha], \tag{4}$$

and $G(v) > \frac{v}{\alpha}$, then $\limsup_{N\to\infty} \tau_N \leq w$ and $\liminf_{N\to\infty} \tau_N \geq v$.

We also need the corresponding lower bound for the false discovery proportion. Suppose that $|I_{0,N}| = \pi_0 M_N$ for every N, where $\pi_0 \in (0,1]$, and define the true-null empirical CDF

$$\widehat{F}_{0,N}(t) = \frac{1}{\pi_0 M_N} \sum_{i \in I_{0,N}} \mathbf{1}\{P_{i,N} \leq t\}.$$

Lemma 3 (Threshold bracketing implies an FDP lower bound)**.** *In addition to the assumptions of Lemma 2, suppose $\|\widehat{F}_{0,N} - F_0\|_\infty \longrightarrow 0$ for a continuous CDF F_0. Then*

$$\liminf_{N\to\infty} \mathrm{FDP}_N \geq \frac{\alpha \pi_0 F_0(v)}{w}.$$

Figure 2 summarizes the geometry of Lemmas 2–3. Only a feasible point at v and a strictly infeasible terminal interval beginning at w are required.

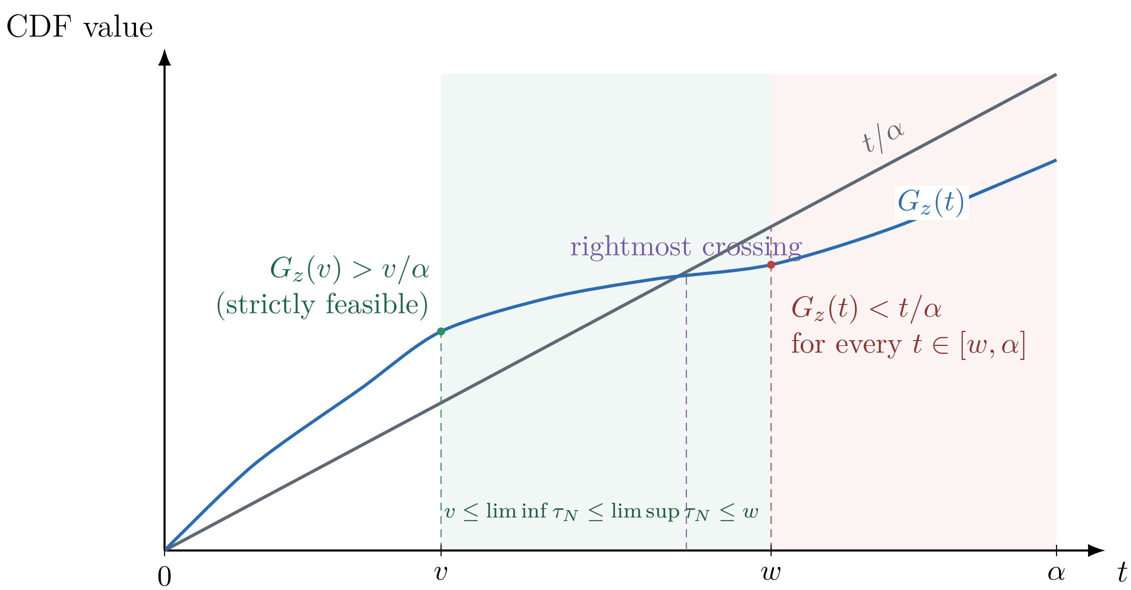

Figure 2: Conceptual BH crossing geometry. The displayed curve is schematic; the proof uses only the two strict sign conditions, not uniqueness or transversality of the crossing.

4 Conditional limiting p-value distributions

We now recall the factor model introduced earlier. Let $a_N \in \mathbb{R}^{100N}$ be the vector whose entries are 3/10 on the first block, $-3/10$ on the second block, and $-18/25$ on the third block. Let D_N be diagonal, with diagonal entries 91/100, 91/100, and 301/625 on the respective blocks. Then $\Sigma_N = a_N a_N^\mathsf{T} + D_N$. Since every diagonal entry of D_N is strictly positive, D_N and therefore Σ_N are positive definite. Moreover,

$$\left(\frac{3}{10}\right)^2 + \frac{91}{100} = 1, \qquad \left(\frac{18}{25}\right)^2 + \frac{301}{625} = 1.$$

Thus every diagonal entry of Σ_N is one, so Σ_N is a correlation matrix.

The mean vector is zero on the first block, 12/5 on the second block, and 22/5 on the third block. In particular, all nonnull means are positive. Every true-null coordinate is marginally $\mathcal{N}(0,1)$, so its two-sided p-value is valid and uniform on $[0,1]$ marginally.

For $c \geq 0$, define $u(c) = 2\overline{\Phi}(c)$. Thus u is continuous and strictly decreasing from 1 to 0, and $P_i \leq u(c)$ exactly when $|X_i| \geq c$. For $a \geq 0$ and $s > 0$, define

$$Q(c; a, s) = \overline{\Phi}\left(\frac{c-a}{s}\right) + \overline{\Phi}\left(\frac{c+a}{s}\right). \tag{5}$$

If $Y \sim \mathcal{N}(m, s^2)$, then $\mathbb{P}(|Y| \geq c) = Q(c; |m|, s)$.

Conditional on $Z = z$, the means and standard deviations in the three blocks are

$$\begin{aligned} M_0(z) &= \frac{3z}{10}, & s_0 &= \frac{\sqrt{91}}{10}, \\ M_1(z) &= \frac{12}{5} - \frac{3z}{10}, & s_1 &= \frac{\sqrt{91}}{10}, \\ M_2(z) &= \frac{22}{5} - \frac{18z}{25}, & s_2 &= \frac{\sqrt{301}}{25}. \end{aligned}$$

Let $F_{g,z}$ be the conditional p-value CDF in block g. Equation (5) gives

$$F_{g,z}(u(c)) = Q(c; |M_g(z)|, s_g).$$

Because the three block proportions are exactly 0.96, 0.01, and 0.03, the conditional limiting CDF of all p-values is

$$G_z(u(c)) = 0.96 \cdot Q(c; |M_0(z)|, s_0) + 0.01 \cdot Q(c; |M_1(z)|, s_1) + 0.03 \cdot Q(c; |M_2(z)|, s_2). \tag{6}$$

At $\alpha = 0.01$, define

$$h_z(c) = G_z(u(c)) - 100u(c). \tag{7}$$

Thus $h_z(c) \geq 0$ is precisely the limiting BH feasibility condition at the p-value threshold $u(c)$.

Conditional on $Z = z$, the p-values are independent and identically distributed within each block. The Glivenko–Cantelli theorem (van der Vaart and Wellner, 1996) applied to each of the three blocks therefore gives $\sup_{0 \leq t \leq 1} |\widehat{G}_N(t) - G_z(t)| \longrightarrow 0$ almost surely under the conditional law given $Z = z$. Applied to the null block, it also gives

$$\sup_{0 \leq t \leq 1} |\widehat{F}_{0,N}(t) - F_{0,z}(t)| \longrightarrow 0.$$

By the existence of regular conditional laws and Fubini's theorem, these two convergences hold jointly with unconditional probability one, with z replaced by the realized value Z. All the limiting CDFs are continuous because the conditional Gaussian laws have strictly positive variances.

For later use, we record two monotonicity properties of Q. For $c, a \geq 0$ and $s > 0$,

$$\frac{\partial}{\partial c} Q(c; a, s) = -\frac{1}{s} \left\{ \phi\left(\frac{c-a}{s}\right) + \phi\left(\frac{c+a}{s}\right) \right\} < 0,$$

so Q is strictly decreasing in c. Also,

$$\frac{\partial}{\partial a} Q(c; a, s) = \frac{1}{s} \left\{ \phi\left(\frac{c-a}{s}\right) - \phi\left(\frac{c+a}{s}\right) \right\} \geq 0.$$

Indeed, $|c-a| \leq c+a$, and the standard normal density is decreasing as a function of the absolute value of its argument. Hence Q is nondecreasing in $|m|$.

5 A finite collection of inequalities suffices

Let $c_\alpha = \Phi^{-1}(1-\alpha/2)$. Since BH thresholds never exceed α, only $c \geq c_\alpha$ is relevant. Partition $[-5, 5]$ into the one thousand intervals

$$B_k = \left[\frac{k}{100}, \frac{k+1}{100}\right], \qquad -500 \leq k \leq 499.$$

For $g \in \{0, 1, 2\}$, define the exact extrema

$$m_{g,k}^- = \min_{z \in B_k} |M_g(z)|, \qquad m_{g,k}^+ = \max_{z \in B_k} |M_g(z)|.$$

Because each M_g is affine, these extrema are obtained exactly from the two endpoints and, when the affine function changes sign on the interval, from the value zero.

Use the rational c-grid $c_j = \frac{j}{1000}$, $2575 \leq j \leq 10000$. We provide below a numerical certificate that verifies $u(c_{2575}) > 0.01$, equivalently $c_{2575} < c_\alpha$, so this grid starts below the entire relevant c-domain.

For $2575 \leq j < 10000$, define

$$U_{j,k} = 0.96 \cdot Q(c_j; m_{0,k}^+, s_0) + 0.01 \cdot Q(c_j; m_{1,k}^+, s_1) + 0.03 \cdot Q(c_j; m_{2,k}^+, s_2) - 100u(c_{j+1}). \quad (8)$$

If $z \in B_k$ and $c \in [c_j, c_{j+1}]$, the monotonicities just proved imply

$$Q(c; |M_g(z)|, s_g) \leq Q(c_j; m_{g,k}^+, s_g),$$

while the decrease of u gives $-100u(c) \leq -100u(c_{j+1})$. Therefore

$$h_z(c) \leq U_{j,k} \quad \text{on } B_k \times [c_j, c_{j+1}]. \qquad (9)$$

Let j_k be the first grid index for which $U_{j,k}$ is not certified to be strictly negative, and put $a_k = c_{j_k}$. Our numerical certificate verifies $u(a_k) < \alpha$. In particular, $j_k > 2575$. Every preceding rectangle has $U_{j,k} < 0$, so

$$h_z(c) < 0 \quad \text{for every } z \in B_k \text{ and every } c \in [c_\alpha, a_k]. \qquad (10)$$

The endpoint a_k is included because it is the right endpoint of the last strictly certified rectangle.

For $j \geq j_k$, define the pointwise lower bound

$$L_{j,k} = 0.96 \cdot Q(c_j; m_{0,k}^-, s_0) + 0.01 \cdot Q(c_j; m_{1,k}^-, s_1) + 0.03 \cdot Q(c_j; m_{2,k}^-, s_2) - 100u(c_j). \quad (11)$$

Let $\ell_k \geq j_k$ be the first index for which $L_{\ell_k,k}$ is certified to be strictly positive, and put $b_k = c_{\ell_k}$. For every $z \in B_k$, monotonicity in the absolute mean gives

$$h_z(b_k) \geq L_{\ell_k,k} > 0. \quad (12)$$

Because $b_k \geq a_k$, one has $u(b_k) \leq u(a_k) < \alpha$.

Fix an outcome in the probability-one event on which both empirical-CDF convergences hold, and suppose that its realized factor value $z = Z$ lies in B_k. In Lemma 2, take

$$v = u(b_k), \qquad w = u(a_k), \qquad G = G_z.$$

As u is decreasing, condition (10) is exactly $G_z(t) < 100t = t/\alpha$ for every $t \in [u(a_k), \alpha]$, and (12) is exactly $G_z(u(b_k)) > 100u(b_k)$. Lemma 3, with $\pi_0 = 0.96$, yields

$$\liminf_{N\to\infty} \mathrm{FDP}_N \geq \frac{0.01 \cdot 0.96 \cdot F_{0,z}(u(b_k))}{u(a_k)}.$$

Now

$$F_{0,z}(u(b_k)) = Q(b_k; |M_0(z)|, s_0) \geq Q(b_k; m_{0,k}^-, s_0).$$

Consequently, on this probability-one event, whenever $Z \in B_k$,

$$\liminf_{N\to\infty} \mathrm{FDP}_N \geq d_k, \qquad d_k = \frac{0.0096Q(b_k; m_{0,k}^-, s_0)}{u(a_k)}. \quad (13)$$

The finite reduction for one z-bin is shown in Figure 3. The upper bounds $U_{j,k}$ certify a whole prefix of infeasible BH thresholds, whereas one lower bound $L_{\ell_k,k}$ provides a feasible point farther out in the Gaussian-tail coordinate.

6 The certified lower bound and completion of the proof

Every rational input to the certificate—the means, factor loadings, block weights, z-bin endpoints, and c-grid points—is specified using ratios of integers, without binary floating-point literals. The square roots defining the residual standard deviations and all Gaussian tails are evaluated as Arb balls. Arb performs outward-rounded midpoint-radius ball arithmetic (Johansson, 2017): every computed ball contains the exact real value. A strict comparison is accepted only when the entire resulting ball lies on the stated side of zero. Thus the tests $U_{j,k} < 0$, $L_{j,k} > 0$, $u(c_{2575}) > \alpha$, and $u(a_k) < \alpha$ are rigorous interval statements, not floating-point heuristics.

The certificate computes the right side of

$$\sum_{k=-500}^{499} d_k \left\{ \Phi\left(\frac{k+1}{100}\right) - \Phi\left(\frac{k}{100}\right) \right\}. \quad (14)$$

For readability, the one thousand terms are grouped into ten unit intervals. The following are certified strict lower bounds; every displayed decimal is rounded downward:

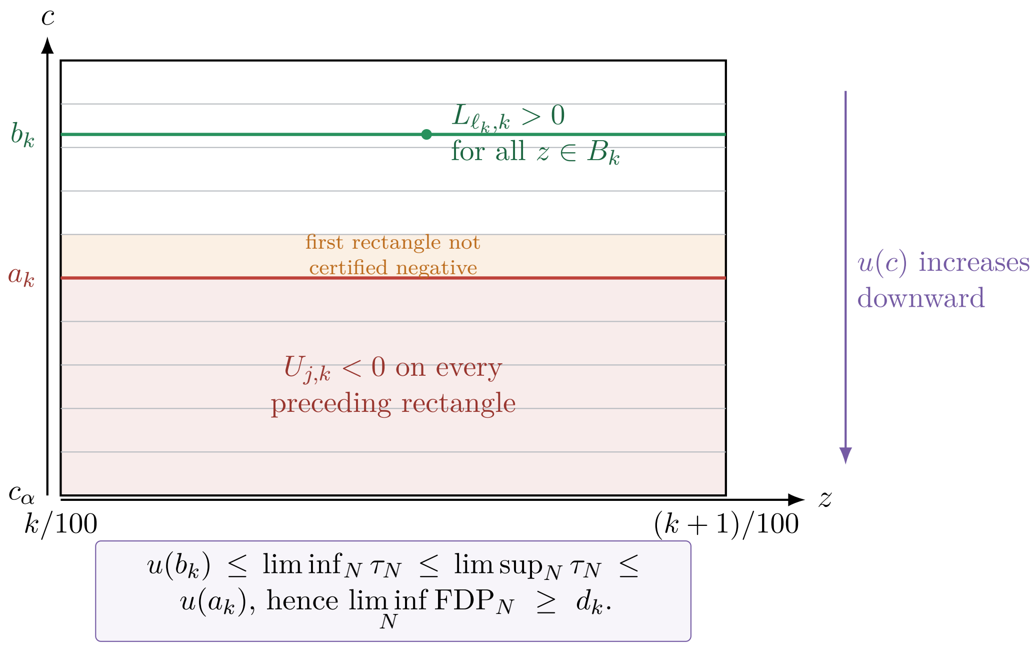

Figure 3: One certified rectangle column $B_k \times [c_\alpha, 10]$. Monotonicity in c and in the absolute conditional means makes the displayed finite sign checks valid uniformly over the entire bin.

Range of Z	Contribution to (14)
$[-5, -4]$	0.0000062542270572156952920154711845188 27
$[-4, -3]$	0.000116801968722543993192086605837856445
$[-3, -2]$	0.000762743726482023098968624717496448543
$[-2, -1]$	0.001839640615452850387738159572547846549
$[-1, 0]$	0.002006865275162943558353752660173437476
$[0, 1]$	0.001953237173075828884087571780890001262
$[1, 2]$	0.001827735787140664069157918588581283062
$[2, 3]$	0.001367544916156818165023914540138520994
$[3, 4]$	0.000508812722703075293268312075251660515
$[4, 5]$	0.000027192658519749972390210373148917831
Total on $[-5, 5]$	0.010416829070473713117472566385250491510

Since $0 \le \mathrm{FDP}_N \le 1$, Fatou's lemma applies. The bins cover $[-5, 5]$ up to endpoints of Gaussian probability zero. Equation (13) and the nonnegativity of the contribution from $|Z| > 5$ give

$$\liminf_{N\to\infty} \mathrm{FDR}_N = \liminf_{N\to\infty} \mathbb{E}[\mathrm{FDP}_N] \ge \mathbb{E}\left[\liminf_{N\to\infty} \mathrm{FDP}_N\right] \ge \sum_{k=-500}^{499} d_k \mathbb{P}(Z \in B_k) > 0.0104.$$

This proves Theorem 1.

7 A Monte Carlo experiment

Here we provide a Monte Carlo experiment to support the theoretical analysis. However, a naive Monte Carlo experiment is inefficient: the conditional FDP is typically modest for central values of the common factor Z, whereas uncommon factor values can produce much larger FDPs and make a nonnegligible contribution to the expectation. We therefore stratify on Z while simulating every residual coordinate and recomputing the BH rule without approximation.

Let $K = 1000$ and define the equiprobable standard-normal strata

$$I_k = \left(\Phi^{-1}\left(\frac{k-1}{K}\right), \Phi^{-1}\left(\frac{k}{K}\right)\right], \qquad 1 \le k \le K,$$

with the usual interpretations at probabilities zero and one. In macro-replication b, draw independently

$$U_{b,k} \sim \mathrm{Unif}\left(\frac{k-1}{K}, \frac{k}{K}\right), \qquad Z_{b,k} = \Phi^{-1}(U_{b,k}),$$

one draw from each stratum. Conditional on each $Z_{b,k}$, generate all $100N$ Gaussian coordinates from (1)–(3), form the two-sided p-values, sort them, apply ordinary BH at $\alpha = 0.01$ exactly, and record the resulting $D_{b,k,N} = \mathrm{FDP}_{b,k,N}$. The macro-replication estimate is $Y_{b,N} = \frac{1}{K}\sum_{k=1}^{K} D_{b,k,N}$. Because every stratum has probability $1/K$ and $Z_{b,k}$ has the conditional law of Z given $Z \in I_k$,

$$\mathbb{E}[Y_{b,N}] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\mathrm{FDP}_N \mid Z \in I_k] = \mathbb{E}[\mathrm{FDP}_N] = \mathrm{FDR}_N\,.$$

Thus stratification changes the Monte Carlo variance but not the estimand. We used $B = 100$ independent macro-replications and reported $\widehat{\mathrm{FDR}}_N = \frac{1}{B}\sum_{b=1}^{B} Y_{b,N}$, $\widehat{\mathrm{MCSE}} = \frac{s_Y}{\sqrt{B}}$, where s_Y is the sample standard deviation of the $Y_{b,N}$. The intervals below are conventional Student-t Monte Carlo intervals with $B - 1 = 99$ degrees of freedom.

The experiment used 100,000 complete Gaussian data sets at each dimension. The retained reproducibility bundle uses Python 3.12.3, NumPy 1.26.4, SciPy 1.14.1, and python-flint 0.8.0. The complete executable and the macro-replication outputs are available in the GitHub repository at `https://github.com/dobriban/BH`.

N	m	$\widehat{\mathrm{FDR}}_N$	MCSE	95% MC interval	p_+
50	5,000	0.009936	0.000113	[0.009711, 0.010161]	0.713
100	10,000	0.010129	0.000096	[0.009939, 0.010320]	0.0905
200	20,000	0.010359	0.000103	[0.010155, 0.010563]	3.56×10^{-4}

Here p_+ denotes the one-sided Student-t Monte Carlo p-value for the null inequality $\mathrm{FDR}_N \le 0.01$.

At $N = 200$, the excess is $\widehat{\mathrm{FDR}}_{200} - \alpha = 0.0003589$, or about 3.59% of the nominal level. The corresponding statistic is 3.495 Monte Carlo standard errors above α, giving the one-sided p-value in the table. Thus the direct finite-dimensional experiment provides evidence of the same failure of control established asymptotically. At $N = 50$ and $N = 100$, the intervals still overlap the nominal level. This may be because more Monte Carlo replications are needed, or because the true finite-sample FDR does not exceed the nominal level at those dimensions.

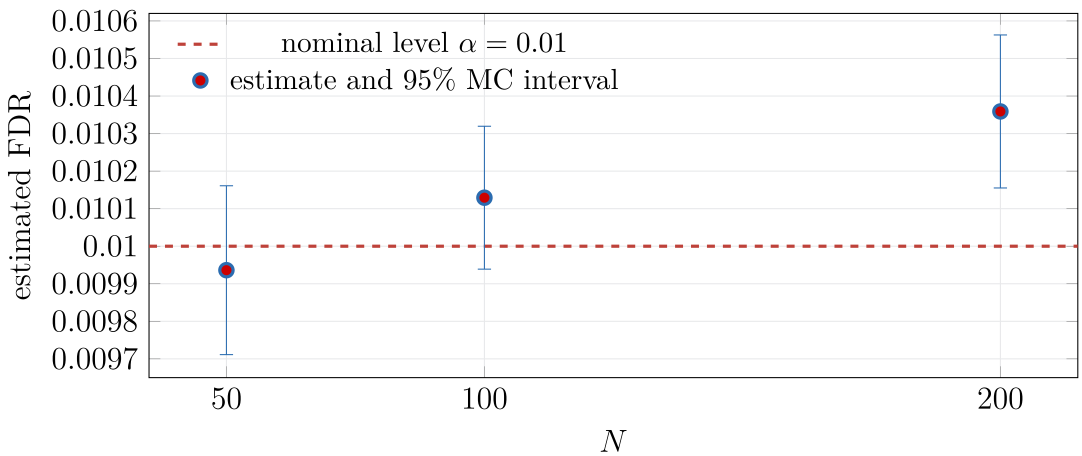

Figure 4: Finite-sample stratified Monte Carlo estimates. The first two intervals do not resolve the sign of $\mathrm{FDR}_N - \alpha$, while the interval at $N = 200$ lies wholly above the nominal level.

8 Discussion

Several points merit further study. The example above violates the nominal level only slightly, and additional numerical searches over related models have found similarly small violations. This raises the question of whether a universal bound exists on the possible inflation of the FDR above its nominal level.

Moreover, the example uses a large number of tests. It is therefore important to determine whether the BH procedure is guaranteed to control the FDR for smaller numbers of tests and, if not, to obtain bounds that depend explicitly on the number of tests. Both questions are directions for future research.

References

Mohammed Abouzaid, Andrew J. Blumberg, Martin Hairer, Joe Kileel, Tamara G. Kolda, Paul D. Nelson, Daniel Spielman, Nikhil Srivastava, Rachel Ward, Shmuel Weinberger, and Lauren Williams. First proof solutions and comments. `https://1stproof.org/documents/FirstProofSolutionsComments.pdf`, February 2026. Manuscript dated February 14, 2026.

Boris Alexeev and Dustin G Mixon. Forbidden sidon subsets of perfect difference sets, featuring a human-assisted proof. *arXiv preprint arXiv:2510.19804*, 2025.

Boris Alexeev, John Jasper, and Dustin G Mixon. Asymptotically optimal approximate hadamard matrices. *arXiv preprint arXiv:2511.14653*, 2025.

Y. Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B*, 72(4):405–416, 2010. doi:`10.1111/j.1467-9868.2010.00746.x`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi:`10.1111/j.2517-6161.1995.tb02031.x`.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi:10.1214/aos/1013699998.

G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008. doi:10.1214/08-EJS180.

Sébastien Bubeck, Christian Coester, Ronen Eldan, Timothy Gowers, Yin Tat Lee, Alexandru Lupsasca, Mehtaab Sawhney, Robert Scherrer, Mark Sellke, Brian K. Spears, Derya Unutmaz, Kevin Weil, Steven Yin, and Nikita Zhivotovskiy. Early science acceleration experiments with gpt-5, 2025. URL https://arxiv.org/abs/2511.16072.

Edgar Dobriban. Solving a research problem in mathematical statistics with ai assistance. *arXiv preprint arXiv:2511.18828*, 2025.

Alessio Farcomeni. More powerful control of the false discovery rate under dependence. *Statistical Methods and Applications*, 15(1):43–73, 2006.

Moran Feldman and Amin Karbasi. G\" odel test: Can large language models solve easy conjectures? *arXiv preprint arXiv:2509.18383*, 2025.

H. Finner, T. Dickhaus, and M. Roters. Dependency and false discovery rate: Asymptotics. *The Annals of Statistics*, 35(4):1432–1455, 2007. doi:10.1214/009053607000000046.

W. Fithian and L. Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118, 2022. doi:10.1214/21-AOS2137.

C. R. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B*, 64(3):499–517, 2002. doi:10.1111/1467-9868.00347.

C. R. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004. doi:10.1214/009053604000000283.

D. Ghosh and S. K. Sarkar. Dependence-aware false discovery rate control in two-sided Gaussian mean testing, 2025. Preprint.

Uijeong Jang and Ernest K Ryu. Point convergence of nesterov's accelerated gradient method: An ai-assisted proof. *arXiv preprint arXiv:2510.23513*, 2025.

F. Johansson. Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017. doi:10.1109/TC.2017.2690633.

Kyung In Kim and Mark A van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC bioinformatics*, 9(1):114, 2008.

OpenAI. A proof of the cycle double cover conjecture. https://cdn.openai.com/pdf/04d1d1e4-bc75-476a-97cf-49055cd98d31/cdc_proof.pdf, 2026a. Accessed 2026-07-13.

OpenAI. Planar point sets with many unit distances. https://cdn.openai.com/pdf/74c24085-19b0-4534-9c90-465b8e29ad73/unit-distance-proof.pdf, 2026b. Accessed 2026-07-13.

A. Reiner-Benaim. FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal*, 49(1):107–126, 2007. doi:10.1002/bimj.200510313.

Adil Salim. Accelerating mathematical research with language models: A case study of an interaction with gpt-5-pro on a convex analysis problem. *arXiv preprint arXiv:2510.26647*, 2025.

S. K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30(1):239–257, 2002. doi:10.1214/aos/1015362192.

S. K. Sarkar. On controlling the false discovery rate in multiple testing of the means of correlated normals against two-sided alternatives, 2023. Preprint.

S. K. Sarkar and S. Zhang. Shifted BH methods for controlling false discovery rate in multiple testing of the means of correlated normals against two-sided alternatives. *Journal of Statistical Planning and Inference*, 236:106238, 2025. doi:10.1016/j.jspi.2024.106238.

J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B*, 66(1):187–205, 2004. doi:10.1111/j.1467-9868.2004.00439.x.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

Erik Y. Wang. Adaboost does not always cycle: A computer-assisted counterexample, 2026. URL https://arxiv.org/abs/2604.07055.

A Proofs

A.1 Proof of Lemma 2

Proof. Set $H(t) = G(t) - t/\alpha$. By continuity and the strict inequality (4), the maximum of H on the compact interval $[w, \alpha]$ is a strictly negative number. Hence there is an $\varepsilon > 0$ such that $H(t) \le -2\varepsilon \quad$ for all $t \in [w, \alpha]$. For all sufficiently large N, uniform convergence gives $\|\widehat{G}_N - G\|_\infty < \varepsilon$. Therefore

$$\widehat{G}_N(t) - \frac{t}{\alpha} \le H(t) + \varepsilon \le -\varepsilon < 0$$

for every $t \in [w, \alpha]$. No BH grid point in that interval is feasible, which proves $\limsup_N \tau_N \le w$.

For the lower bound, $G(v) > \frac{v}{\alpha}$ and continuity give a $\delta > 0$ and an open interval J containing v such that $H(t) \ge 2\delta$ for all $t \in J$. The mesh of $\mathcal{T}_N$ is $\alpha/M_N \to 0$, so one can choose $s_N \in \mathcal{T}_N \cap J$ with $s_N \to v$. For all sufficiently large N, uniform convergence gives

$$\widehat{G}_N(s_N) - \frac{s_N}{\alpha} \ge H(s_N) - \delta \ge \delta > 0.$$

Thus s_N is feasible, and the maximal feasible grid point obeys $\tau_N \ge s_N$. Taking lower limits proves $\liminf_N \tau_N \ge v$. □

A.2 Proof of Lemma 3

Proof. Lemma 2 gives $\liminf_N \tau_N \geq v > 0$, so eventually $\tau_N > 0$. By the definition of the BH threshold, $R_N = \frac{M_N \tau_N}{\alpha}$. The number of false rejections is $V_N = \pi_0 M_N \widehat{F}_{0,N}(\tau_N)$. Consequently,

$$\mathrm{FDP}_N = \frac{V_N}{R_N} = \frac{\alpha \pi_0 \widehat{F}_{0,N}(\tau_N)}{\tau_N}. \tag{15}$$

For every $\varepsilon \in (0, v)$, the inequality $\tau_N \geq v - \varepsilon$ holds eventually. Monotonicity of empirical CDFs and uniform convergence then give $\liminf_{N\to\infty} \widehat{F}_{0,N}(\tau_N) \geq F_0(v - \varepsilon)$. Letting $\varepsilon \downarrow 0$ and using continuity of F_0 yields $\liminf_{N\to\infty} \widehat{F}_{0,N}(\tau_N) \geq F_0(v)$. For any $\varepsilon > 0$, these two bounds imply, eventually, $\widehat{F}_{0,N}(\tau_N) \geq F_0(v) - \varepsilon$ and $\tau_N \leq w + \varepsilon$. Substitution in (15) gives

$$\mathrm{FDP}_N \geq \frac{\alpha \pi_0 \{F_0(v) - \varepsilon\}}{w + \varepsilon}.$$

Letting $\varepsilon \downarrow 0$ proves the result. □

B Complete outward-rounded certificate

The following Python program is the complete numerical certificate used above. It requires `python-flint 0.8.0`. The specialization to the exact grids used in the proof deliberately avoids floating-point conversion in all mathematical inputs and all grid-index calculations.

```
#!/usr/bin/env python3
"""Outward-rounded certificate for the two-sided Gaussian BH counterexample.

Dependency: python-flint

All model parameters and all subdivision endpoints are exact rationals.
Every transcendental evaluation is an Arb ball with outward rounding.
"""

from flint import arb, ctx

ctx.dps = 40

ALPHA = arb(1) / 100
PI0 = arb(24) / 25
W1 = arb(1) / 100
W2 = arb(3) / 100
R0 = arb(3) / 10
R1 = -arb(3) / 10
R2 = -arb(18) / 25
MU1 = arb(12) / 5
MU2 = arb(22) / 5
S0 = (1 - R0 * R0).sqrt()
S1 = (1 - R1 * R1).sqrt()
S2 = (1 - R2 * R2).sqrt()
SQRT2 = arb(2).sqrt()

Z_DEN = 100
C_DEN = 1000
K_MIN = -500
K_MAX = 500
J_START = 2575
J_STOP = 10000
```

```python
def normal_upper_tail(x: arb) -> arb:
    return (x / SQRT2).erfc() / 2


def normal_cdf(x: arb) -> arb:
    return 1 - normal_upper_tail(x)


def two_sided_tail(c: arb, abs_mean: arb, sd: arb) -> arb:
    return (
        normal_upper_tail((c - abs_mean) / sd)
        + normal_upper_tail((c + abs_mean) / sd)
    )


def p_threshold(c: arb) -> arb:
    return 2 * normal_upper_tail(c)


def abs_range_of_affine(
    mu: arb, loading: arb, lo: arb, hi: arb
) -> tuple[arb, arb]:
    left = mu + loading * lo
    right = mu + loading * hi
    abs_left = abs(left)
    abs_right = abs(right)
    if (left <= 0 and right >= 0) or (right <= 0 and left >= 0):
        minimum = arb(0)
    else:
        minimum = abs_left if abs_left < abs_right else abs_right
    maximum = abs_left if abs_left > abs_right else abs_right
    return minimum, maximum


def certify_bin(k: int) -> arb:
    z_lo = arb(k) / Z_DEN
    z_hi = arb(k + 1) / Z_DEN

    m0_lo, m0_hi = abs_range_of_affine(arb(0), R0, z_lo, z_hi)
    m1_lo, m1_hi = abs_range_of_affine(MU1, R1, z_lo, z_hi)
    m2_lo, m2_hi = abs_range_of_affine(MU2, R2, z_lo, z_hi)

    c_start = arb(J_START) / C_DEN
    assert p_threshold(c_start) > ALPHA

    j_lower = None
    for j in range(J_START, J_STOP):
        c_j = arb(j) / C_DEN
        c_next = arb(j + 1) / C_DEN
        h_upper = (
            PI0 * two_sided_tail(c_j, m0_hi, S0)
            + W1 * two_sided_tail(c_j, m1_hi, S1)
            + W2 * two_sided_tail(c_j, m2_hi, S2)
            - 100 * p_threshold(c_next)
        )
        if not (h_upper < 0):
            j_lower = j
            break
    if j_lower is None:
        raise RuntimeError(f"No lower bracket in z-bin {k}")

    c_lower = arb(j_lower) / C_DEN
    # This is equivalent to c_lower > c_alpha and ensures that at least
    # one preceding cell was rigorously certified negative.
    assert p_threshold(c_lower) < ALPHA
```

```python
    j_upper = None
    for j in range(j_lower, J_STOP + 1):
        c_j = arb(j) / C_DEN
        h_lower = (
            PI0 * two_sided_tail(c_j, m0_lo, S0)
            + W1 * two_sided_tail(c_j, m1_lo, S1)
            + W2 * two_sided_tail(c_j, m2_lo, S2)
            - 100 * p_threshold(c_j)
        )
        if h_lower > 0:
            j_upper = j
            break
    if j_upper is None:
        raise RuntimeError(f"No feasible point in z-bin {k}")

    c_upper = arb(j_upper) / C_DEN
    fdp_lower = (
        (PI0 / 100)
        * two_sided_tail(c_upper, m0_lo, S0)
        / p_threshold(c_lower)
    )
    gaussian_mass = normal_cdf(z_hi) - normal_cdf(z_lo)
    return fdp_lower * gaussian_mass


def main() -> None:
    total = arb(0)
    unit_totals: dict[int, arb] = {}

    for k in range(K_MIN, K_MAX):
        contribution = certify_bin(k)
        total += contribution
        unit = k // Z_DEN
        unit_totals[unit] = unit_totals.get(unit, arb(0)) + contribution

    for unit in sorted(unit_totals):
        print(f"z in [{unit},{unit + 1}]: {unit_totals[unit]}")
    print(f"certified total over [-5,5]: {total}")

    assert total > arb("0.010416829070473713117472566385250491 5")
    assert total > ALPHA
    print(
        "CERTIFIED: liminf FDR > "
        "0.0104168290704737131174725663852504915 > alpha = 0.01"
    )


if __name__ == "__main__":
    main()
```