Multiplicity, Selection, and Modern Statistical Inference

Xianyang Zhang

Draft version, June 2026.

Comments, corrections, and suggestions are welcome.

Contents

Preface

To the Reader

I wrote this book for students who have already met probability, mathematical statistics, and linear regression, and who now want a coherent route into modern large-scale inference. The recurring question is simple: after many hypotheses, model choices, or benchmark comparisons have been examined, what claim can still be made honestly?

These chapters grew out of lectures for statistics Ph.D. students. I have tried to make the written version more patient than lectures can be: definitions before use, examples that show what the theory is buying, reproducible numerical experiments, and exercises that turn reading into active checking. Some arguments need to be read twice; that is normal. The goal is not to memorize a catalogue of methods, but to learn how to ask careful inferential questions when modern data analysis can easily outrun intuition.

Prerequisites

To read the technical chapters comfortably, it helps to have seen the following material before opening Chapter 2.

- Graduate probability, including the measure-theoretic language of expectation, conditional expectation, modes of convergence, and the Gaussian and multinomial distributions.
- Mathematical statistics: hypothesis testing, Neyman-Pearson, asymptotic normality, likelihood ratios, contiguity in some form.
- Linear regression in matrix form, including the normal equations and the projection geometry of ordinary least squares.
- Basic real analysis sufficient to read ϵ-style arguments and routine inequalities (Markov, Chebyshev, Cauchy-Schwarz, union bound).

Familiarity with the Lasso, martingales, and empirical processes is helpful but not assumed. When these tools first appear, I introduce what is needed for the chapter and point to standard references for readers who want to go deeper.

Organization

Chapter 1 introduces the book's common language: inferential targets, evidence measures, error rates, dependence, and selection. Chapters 2 and 3 develop global testing: detecting whether any signal is present in a high-dimensional collection of hypotheses. Chapters 4 to 6 cover individual decisions under familywise and false-discovery error rates, including closure, graphical procedures, BH, PRDS, adaptive estimation, and the two-groups empirical-Bayes model. Chapter 7 extends these tools to non-flat families: grouped hypotheses, tree-structured multiplicity, and replicability across studies. Chapter 8 introduces e-values, safe testing, e-BH, online FDR, and confidence sequences. Chapter 9 develops conditional randomization tests, knockoff filters, and multilayer knockoffs for high-dimensional variable selection. Chapters 10 to

12 treat distribution-free prediction with conformal inference (including risk control and beyond-exchangeability extensions), confidence intervals for high-dimensional regression coefficients via the debiased Lasso, and selective inference after data-driven model selection (including data thinning and fission for unsupervised targets). Chapter 13 uses applications in genomics, watermarking, contamination auditing, and conformal factuality to show how the same statistical principles reappear in modern scientific and AI workflows.

The chapters can mostly be read in order, but they are not strictly linear. A reader who wants to learn FDR control quickly may go straight from Chapter 4 to Chapters 5 and 6. A reader interested in e-values may read Chapter 8 immediately after Chapter 5. A reader focused on prediction and selective inference may go directly from Chapter 6 to Chapters 10 and 12. Readers interested specifically in hierarchical genomic or regulatory testing can read Chapter 7 immediately after the FDR chapters.

Acknowledgments

This book grew out of the STAT 689 lecture sequence at Texas A&M University. I am grateful to the students and colleagues whose questions and comments sharpened the exposition. I also benefited a lot from Emmanuel Candès's Stats 300C lecture materials.

Notation

The book uses the conventions below. Chapter-specific notation is introduced where it appears.

Counts

m	number of hypotheses, tests, p-values, or e-values
n	sample size (number of observations)
p	number of features in a regression model
m_0	number of true null hypotheses
m_1	number of nonnulls, $m_1 = m - m_0$

Hypotheses and decisions

H_i, H_{0i}, H_{1i}	ith hypothesis with its null and alternative
$H_I = \bigcap_{i \in I} H_i$	intersection hypothesis on index set I
$\mathcal{I}_0$	index set of true nulls, so $m_0 = \lvert\mathcal{I}_0\rvert$
$\mathcal{R}, R$	rejection set and its size $R = \lvert\mathcal{R}\rvert$
U	true nulls not rejected, $U = m_0 - V$
V	number of false rejections, $V = \lvert\mathcal{R} \cap \mathcal{I}_0\rvert$
T	nonnulls not rejected (missed discoveries), $T = m_1 - S$
S	number of true discoveries, $S = R - V$
$\widehat{S}$	data-driven selected set (Chapters 9, 12)

Error rates

$\mathrm{FWER} = \mathbb{P}(V \geq 1)$	familywise error rate
$\mathrm{FDP} = V/(R \vee 1)$	false discovery proportion
$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$	false discovery rate
$\mathrm{pFDR} = \mathbb{E}[V/R \mid R > 0]$	positive false discovery rate
FCR	false coverage rate (Chapter 12)

Evidence and observables

p_i	p-value for H_i, super-uniform under H_{0i}
E_i	e-value for H_i, satisfying $\mathbb{E}[E_i] \leq 1$ under H_{0i}
Z_i	z-score; in the Gaussian sequence model $Z_i \sim N(\mu_i, 1)$
$\mathrm{lfdr}(z)$	local false discovery rate $\mathbb{P}(\theta = 0 \mid Z = z)$
W_j	knockoff statistic (Chapter 9)

Probability and statistics

$\mathbb{P}$, $\mathbb{E}$, Var, Cov	probability, expectation, variance, covariance
$\mathcal{L}(X)$	law (distribution) of a random variable
Φ, ϕ	standard normal CDF and density
$\mathrm{Unif}(0,1)$	uniform distribution on $[0,1]$
χ^2_k, χ_d	chi-square (k df) and chi (d df, the norm of a standard normal vector in $\mathbb{R}^d$) distributions
$\mathrm{TN}(\mu, \sigma^2, [a,b])$	normal $N(\mu,\sigma^2)$ truncated to $[a,b]$
$\xrightarrow{P}$, $\xrightarrow{d}$	convergence in probability and in distribution
$o_p(\cdot)$, $O_p(\cdot)$	stochastic little-o and big-O: $X_n = o_p(a_n)$ means $X_n/a_n \xrightarrow{P} 0$, and $X_n = O_p(a_n)$ means X_n/a_n is bounded in probability
$\mathbf{1}\{A\}$	indicator of the event A

Regression and high-dimensional models

$X \in \mathbb{R}^{n\times p}$	design matrix (rows are observations, columns are features)
$X_j \in \mathbb{R}^n$	jth column of X (the jth feature across all observations)
$x_i^\top \in \mathbb{R}^{1\times p}$	ith row of X (the feature vector for observation i)
y, Y	response vector
$\beta \in \mathbb{R}^p$	regression coefficients (target)
$\widehat{\beta}$	estimator of β
θ, Θ_0	generic parameter and null parameter set
Σ, Γ	covariance matrix and precision matrix Σ^{-1}
$\widetilde{X}$	knockoff copy of X (Chapter 9)

Conventions

- Boldface is not used for vectors; whether a symbol is scalar or vector is clear from context.
- Uppercase letters denote random variables and lowercase letters denote realizations whenever the distinction matters.
- $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$ are the real, integer, and natural numbers.
- All limits are as $m \to \infty$ or $n \to \infty$ unless otherwise stated; in particular, $o(\cdot)$ and $O(\cdot)$ refer to that limit.
- The level α and FDR target q are reserved symbols for prespecified error budgets. In Chapters 5–7 and Chapter 8, the FDR target is also often written α when no other error budget is in play; q is used to disambiguate when both an FDR target and a coverage level (e.g. conformal miscoverage in Chapter 10) appear together.
- Threshold rules use $R(t) \vee 1$ to avoid division by zero before any rejection has been made. Supremum and infimum rules follow the convention stated where the rule is introduced; when a proof uses the inequality at the selected threshold, the chapter either searches over a finite grid or assumes enough one-sided continuity or closedness for the selected threshold to be attained.
- The blackboard symbols $\mathbb{P}, \mathbb{E}$ denote the abstract probability and expectation operators. Italic $P_0, P_1, P_\theta, E_\theta$ and similar denote *specific* probability measures and the corresponding expectations; the subscript identifies which measure.
- Common distributions are written in roman: $\mathrm{Bin}(n,p)$ for binomial, $\mathrm{Bernoulli}(p)$, and $\mathrm{Unif}(0,1)$ for uniform.

- A few symbols carry one global meaning in this table and a separate chapter-local meaning where the surrounding text makes the role unambiguous. In particular: S (true discoveries) is reused for the true support set in Chapter 11 and for nonconformity/strict-safe scores in Chapters 10 and 13; T (missed discoveries) is reused for the test-statistic vector in Chapter 12 and the contamination score in Chapter 13; ϕ (the standard normal density) is reused as the cluster separation parameter in Chapter 12 and a watermark score in Chapter 13; m (number of hypotheses) is reused as the number of selected interval targets in Chapter 12 and benchmarks in Chapter 13. The local meaning is always declared at the start of the chapter or section that uses it.

CHAPTER 1

Introduction: Multiplicity, Selection, and Modern Inference

The trouble often starts after the first successful plot. A genomic screen has a striking gene near the top of a volcano plot. A high-dimensional regression has a small set of variables selected by the Lasso. A model evaluation table has several cells where a new method appears to beat the baseline. Each finding may be real. Each may also be the best-looking accident among many chances to look.

For example, suppose 20,000 null genes are tested and the smallest p-value is 10^{-5}. On its own, 10^{-5} looks impressive. But if all 20,000 p-values were independent uniforms, then

$$\mathbb{P}\{\min_j p_j \le 10^{-5}\} = 1 - (1 - 10^{-5})^{20000} \approx 0.18.$$

The surprise fades once the family of questions is visible. The calculation is elementary, but the lesson is deep: evidence is interpreted relative to the search that produced it.

The same issue appears in less obvious forms. If a regression analyst fits a Lasso, chooses one selected coefficient, and then reports the ordinary least-squares confidence interval as if that coefficient had been fixed in advance, the interval is answering the wrong question. If a benchmark paper compares many models across many tasks and highlights only the biggest gains, unadjusted p-values exaggerate the evidence. If a prediction method is tuned on many subgroups and then advertised as calibrated, we need to know which coverage statement survived the tuning.

This book develops tools for these situations. Its readers are graduate students in statistics, applied researchers who need reliable large-scale inference, and statisticians who want a coherent path from classical multiple testing to modern methods such as e-values, knockoffs, conformal prediction, debiased Lasso, and selective inference. The level is mathematical, but the organizing question is practical: what claim are we trying to make, and what calibration makes that claim honest?

1. Targets Before Methods

The same data set can support several different inferential targets. A global test asks whether any signal is present. A multiple-testing procedure reports a discovery list while controlling a false-positive error rate. A conformal method reports a prediction set with a coverage guarantee for a future observation. A post-selection method reports an interval or p-value after the object of inference has been chosen using the data.

These targets are related, but they are not interchangeable. A valid global test can tell us that a genomic experiment contains signal, but it does not identify the non-null genes. A Benjamini–Hochberg discovery list controls an average false discovery proportion, but it does not promise that every selected gene is real. A conformal prediction interval gives a coverage statement for a future response, but it does not rank variables by importance. Selective inference can repair a confidence statement after model selection, but only relative to the selection event or randomization that is actually analyzed.

This is the book's first habit of mind: state the target before choosing the procedure. The mathematics then has a job to do. It must connect the reported object to the error rate or coverage statement that the reader will interpret.

2. A Common Grammar

Most chapters follow the same grammar:

$$\text{define the claim} \quad \longrightarrow \quad \text{calibrate evidence} \quad \longrightarrow \quad \text{report a guaranteed object.}$$

The evidence may be a p-value, an e-value, a residual score, a feature statistic, or a randomized comparison. The guarantee may concern FWER, FDR, finite-sample coverage, optional stopping, or a post-selection law. What changes from chapter to chapter is the symmetry or null structure that makes calibration possible.

This matching is where many statistical mistakes enter. A method calibrated for independent p-values should not be casually used under strong dependence. An interval for a pre-specified coefficient should not be read as an interval for the coefficient selected by an algorithm. A discovery procedure designed for marginal hypotheses may not answer a conditional variable-importance question. Later chapters will state assumptions carefully because, in large-scale inference, the assumptions often define the claim.

3. How the Chapters Fit Together

Figure 1 summarizes the book's structure. The arrows are conceptual dependencies, not a rigid reading order. The early chapters develop global testing and multiple-testing error rates; the middle chapters address dependence, structure, adaptivity, and safe evidence; the later chapters handle algorithmic targets, prediction, high-dimensional estimation, and selection.

Chapters 2 and 3 begin with the broadest question: is there signal anywhere? Chapters 4, 5, and 6 move from global evidence to discovery lists, contrasting familywise error control with false-discovery control and then studying dependence, adaptive estimation, and empirical Bayes interpretations. Chapter 7 extends the FDR framework to non-flat families: grouped hypotheses, tree-structured genomic and benchmark testing, and replicability across studies. Chapter 8 introduces e-values and safe testing, where evidence can be accumulated under optional continuation, and develops online FDR and confidence sequences for sequential workflows.

The remaining chapters move closer to modern data-analysis workflows. Chapter 9 studies conditional feature importance through conditional randomization tests, knockoff filters, and multilayer knockoffs. Chapter 10 develops distribution-free prediction through conformal methods, including conformal risk control and beyond-exchangeability extensions. Chapters 11 and 12 treat inference after high-dimensional modeling and data-dependent selection, including data thinning and fission for valid inference after unsupervised discovery steps such as clustering. Chapter 13 closes the book by showing how the same ideas reappear in genomics and language-model evaluation.

4. How to Read the Book

A graduate course can read the chapters nearly in order. Applied researchers may instead start from the target: global signal detection in Chapters 2–3, discovery lists in Chapters 4–7, sequential evidence in Chapter 8, and algorithmic or post-selection inference in Chapters 9–12. Statisticians who already know classical multiple testing may use the later chapters as a bridge to knockoffs, conformal prediction, and selective inference.

Before applying any method, keep five questions in view. What is the reported object? What family or selection rule does the guarantee cover? What null, conditional null, or exchangeability

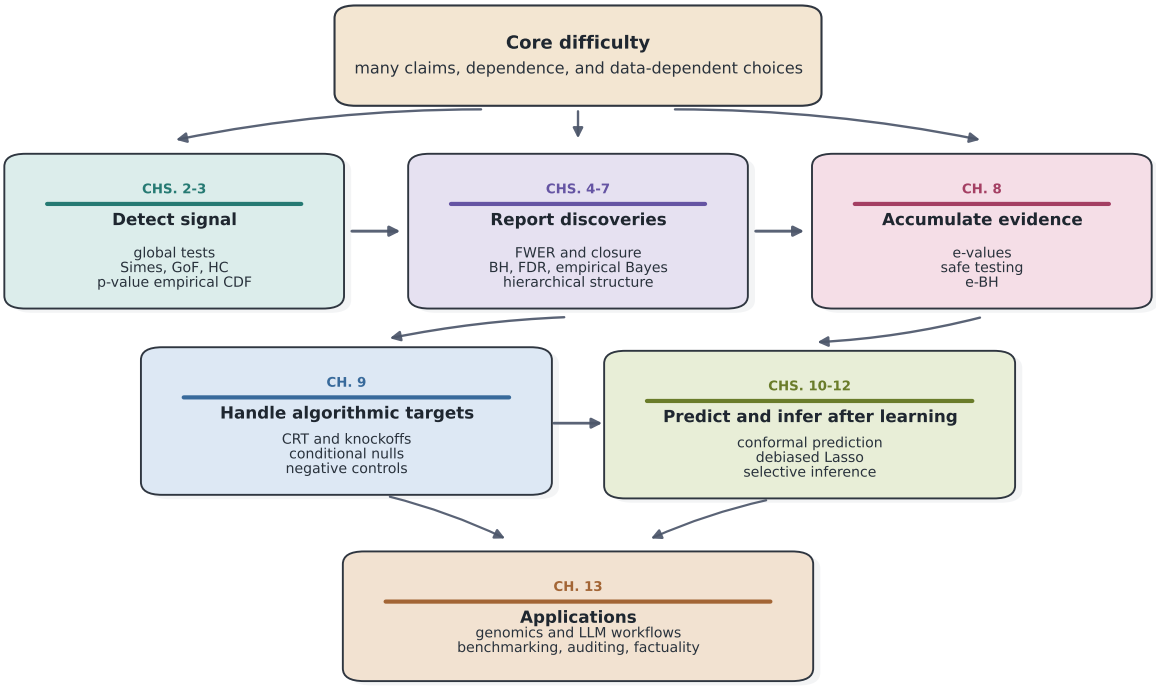

FIGURE 1. Conceptual map of the book. Global testing asks whether signal is present. FWER, FDR, hierarchical, and e-value methods turn calibrated evidence into discovery rules. Knockoffs, conformal prediction, debiased Lasso, and selective inference handle settings in which algorithms or data-dependent choices determine the object being reported. The applications chapter revisits the same questions in genomics and language-model evaluation.

condition justifies the calibration? How does dependence affect the argument? And is the stated error rate the one the scientific claim actually needs?

CHAPTER 2

Global Testing: Extremes, Aggregates, and Detection Boundaries

Large-scale inference often begins before selection, estimation, or ranking. The first question is global: does the data set contain any signal at all? In a genomic screen, this asks whether at least one gene behaves differently between cases and controls. In a high-dimensional regression, it asks whether at least one coefficient is nonzero. In a model-evaluation study, it asks whether a collection of discrepancies can plausibly be explained by noise.

This chapter develops the global testing material that will be used throughout the book. The main theme is that different global tests search for different shapes of evidence. Bonferroni and maximum tests are tuned to a few strong signals. Fisher, Stouffer, and chi-square tests accumulate many weak signals. The Cauchy combination test sits closer to the extreme-value side while retaining a useful robustness to dependence. The chapter ends by making these statements quantitative in the Gaussian sequence model.

1. From Many Hypotheses to One Question

Suppose a study produces m null hypotheses

$$H_{01}, \dots, H_{0m}.$$

The global null is the intersection

$$H_0 = \bigcap_{j=1}^{m} H_{0j}.$$

Rejecting H_0 means that at least one individual null is false. It does not identify which one. The localization problem, where we decide which hypotheses to reject while controlling familywise error or false discovery rate, begins in later chapters.

Example 2.1 (Benchmark-comparison screen). Suppose two prediction systems are compared on m related benchmark tasks, as in modern model-evaluation studies [39, 85]. For task j, the systems are evaluated on the same n_j test cases. Let L_{jk}^A and L_{jk}^B be losses on test case k, and define the paired improvement

$$D_{jk} = L_{jk}^B - L_{jk}^A,$$

so positive values favor system A. A one-sided individual null is

$$H_{0j} : \mathbb{E}D_{j1} \le 0,$$

meaning that task j does not provide evidence that system A improves on system B. With

$$\bar{D}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} D_{jk}, \qquad s_j^2 = \frac{1}{n_j - 1}\sum_{k=1}^{n_j}(D_{jk} - \bar{D}_j)^2,$$

the standardized statistic

$$Z_j = \frac{\sqrt{n_j}\,\bar{D}_j}{s_j}$$

is approximately $N(\mu_j, 1)$ when n_j is moderate. The boundary case $\mu_j = 0$ corresponds to no mean improvement on task j, while $\mu_j > 0$ corresponds to a real improvement. The global

question is whether any benchmark task carries signal. Throughout this chapter we work directly with the Gaussian z-score model, viewing it as an idealized summary that already absorbs whatever distributional approximation produced it.

The common input to many global tests is a collection of p-values $p_1, \dots, p_m$. A p-value for H_{0j} is valid if it is super-uniform under the null:

$$\mathbb{P}_{H_{0j}}(p_j \le t) \le t, \qquad 0 \le t \le 1.$$

Exact uniformity is convenient but not necessary for Type I error control.

For the Gaussian model $Z_j \sim N(\mu_j, 1)$, the two-sided p-value is

$$p_j = 2\{1 - \Phi(|Z_j|)\},$$

whereas the one-sided upper-tail p-value is $p_j = 1 - \Phi(Z_j)$. The direction of the alternative is part of the scientific model and should be fixed before looking at the data.

2. Tests Based on Small P-Values

2.1. Bonferroni and Sidak. The Bonferroni global test rejects H_0 at level α if

$$\min_{1 \le j \le m} p_j \le \frac{\alpha}{m}.$$

PROPOSITION 2.2 (Bonferroni validity). *If all null p-values are super-uniform, then the Bonferroni global test has Type I error at most α, with no assumptions on the dependence among the p-values.*

PROOF. Under the global null,

$$\mathbb{P}\left(\min_j p_j \le \frac{\alpha}{m}\right) = \mathbb{P}\left(\bigcup_{j=1}^{m} \{p_j \le \alpha/m\}\right) \le \sum_{j=1}^{m} \mathbb{P}(p_j \le \alpha/m) \le m \cdot \frac{\alpha}{m} = \alpha.$$

□

This proof is only a union bound, but that simplicity is exactly why the guarantee is so robust. Bonferroni remains valid under arbitrary dependence. When the p-values are independent, the Sidak threshold

$$1 - (1 - \alpha)^{1/m}$$

gives exact level α. Indeed,

$$\mathbb{P}\left(\min_j p_j \le 1 - (1 - \alpha)^{1/m}\right) = \alpha.$$

Bonferroni is often described as conservative. This is sometimes true, but it is not the right mental model in independent large-scale problems. If $p_1, \dots, p_m$ are independent uniforms under the global null, then the *size* of the Bonferroni test — the actual Type I error probability — is

$$1 - \left(1 - \frac{\alpha}{m}\right)^m \longrightarrow 1 - e^{-\alpha}.$$

For $\alpha = 0.05$, the limit is 0.0488, within half a percentage point of the nominal level. So under independence Bonferroni gives up almost nothing; it is a near-exact level-α test. Positive dependence is what makes Bonferroni conservative: the p-values move together, the effective number of independent chances to cross α/m is smaller, and the realized rejection probability falls below $1 - e^{-\alpha}$. Figure 1 shows this effect under an equicorrelated null.

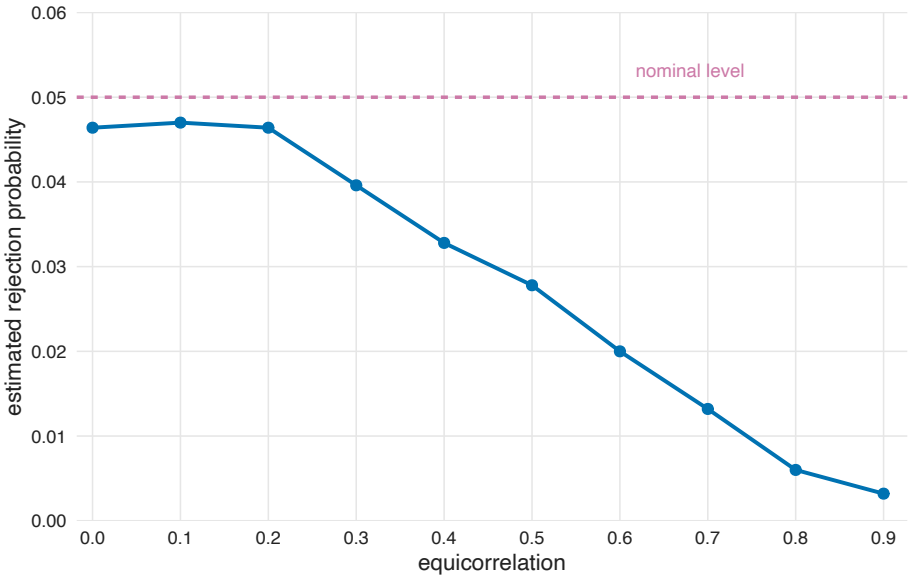

FIGURE 1. Estimated size of Bonferroni's two-sided global test under an equicorrelated Gaussian null with $m = 200$, $\alpha = 0.05$, and 5000 Monte Carlo repetitions. The test remains valid, but positive correlation makes it increasingly conservative.

3. Tests That Combine Evidence

Bonferroni reacts to the smallest p-value. If the alternative consists of many weak effects, no individual p-value may be spectacular, even though the combined evidence is overwhelming. Combination tests are designed for this regime.

3.1. Fisher's Combination Test. If $p \sim \mathrm{Unif}(0,1)$, then $-2\log p \sim \chi^2_2$. Therefore, if the p-values are independent under the global null,

$$T_F = -2\sum_{j=1}^{m} \log p_j \sim \chi^2_{2m}.$$

Fisher's test rejects for large T_F. The exact chi-square calibration is an independence statement; under dependence, the same statistic may still be useful, but its null distribution must be justified separately.

Fisher's statistic is not the only classical combination rule. Pearson's statistic

$$T_P = -2\sum_{j=1}^{m} \log(1 - p_j)$$

is also χ^2_{2m} under independent uniform p-values and rejects for large values; it is sensitive when unusually *large* p-values are evidence against the null (the opposite tail from Fisher's combination of unusually small p-values). Stouffer's method maps p-values to z-scores. For one-sided p-values, define $X_j = \Phi^{-1}(1 - p_j)$. Under independent nulls,

$$Z_S = \frac{\sum_{j=1}^{m} X_j}{\sqrt{m}} \sim N(0,1).$$

A weighted version,

$$Z_w = \frac{\sum_{j=1}^m w_j X_j}{\sqrt{\sum_{j=1}^m w_j^2}},$$

is valid under independent standard normal null scores for fixed weights. This is the first appearance of a theme that will return later: external information can be used to tilt power toward hypotheses that are more promising or measured more precisely.

For readers coming from classical one-dimensional testing, the important distinction is not the name of the combination rule but the functional being applied to the p-value vector. Bonferroni uses the minimum p-value. Fisher uses an average of $-\log p_j$. Stouffer uses an average of normal scores. Pearson uses the opposite tail and is most natural when unusually large p-values are evidence. These choices encode assumptions about the shape of the alternative before any asymptotic theory is invoked. This opposite-tail case is not just a mathematical curiosity: if the p_j's are one-sided upper-tail p-values for beneficial treatment effects, then many unusually large p-values are evidence that the treatment may be harmful rather than beneficial.

3.2. Cauchy Combination. The Cauchy combination test transforms p-values by

$$W = \sum_{j=1}^m w_j \tan\{\pi(1/2 - p_j)\}, \qquad w_j \geq 0, \quad \sum_{j=1}^m w_j = 1.$$

If p_j is uniform, then $\tan\{\pi(1/2 - p_j)\}$ has the standard Cauchy distribution. The heavy tail means that a single very small p-value can dominate the sum. For p-values obtained from jointly normal test statistics, the upper tail of W is asymptotically standard Cauchy under broad dependence conditions [87]. Thus the Cauchy test is often useful when one wants an analytic p-value calculation without assuming independence.

The same idea extends beyond the Cauchy distribution: one can combine p-values through other heavy-tailed calibrators. The point is not that Cauchy is magic; the point is that heavy tails preserve sensitivity to extremes while allowing a tractable tail approximation.

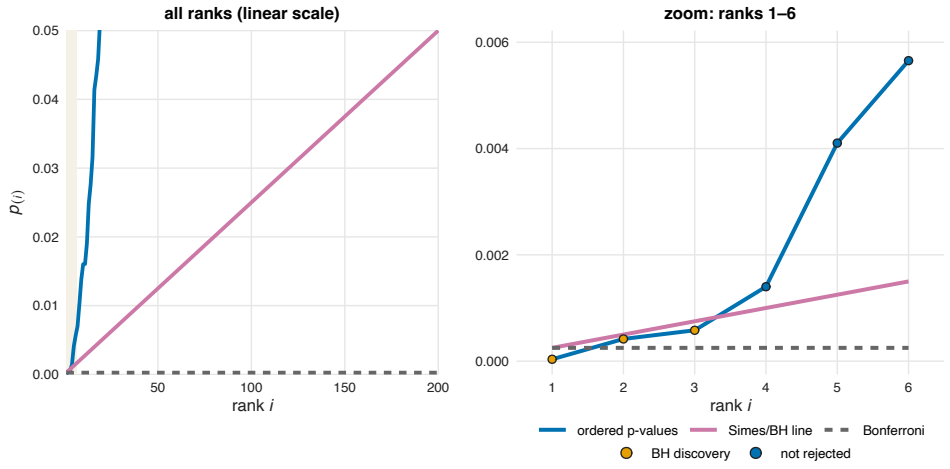

FIGURE 2. Sorted p-values from one simulated screen with $m = 200$ two-sided z-score tests at level $\alpha = 0.05$; six coordinates have mean 3.2, the rest are null. Bonferroni rejects p-values below the horizontal threshold $\alpha/m = 2.5 \times 10^{-4}$. Rank-based tests, such as Simes and BH, compare ordered p-values with a line increasing in the rank $\alpha i/m$; these appear in later chapters.

4. The Gaussian Sequence Model

To compare global tests sharply, we now work in the Gaussian sequence model

$$Z_j = \mu_j + \xi_j, \qquad \xi_j \overset{\text{iid}}{\sim} N(0,1), \qquad j = 1, \dots, m.$$

The global null is $\mu = 0$. This model is deliberately simple: it removes nuisance parameters so that the geometry of the alternatives is visible.

4.1. Bonferroni as a Maximum Test. For two-sided p-values $p_j = 2\{1 - \Phi(|Z_j|)\}$, Bonferroni rejects when

$$\max_{1 \le j \le m} |Z_j| \ge \Phi^{-1}\left(1 - \frac{\alpha}{2m}\right).$$

For the one-sided upper-tail problem, the analogous maximum test rejects when

$$\max_j Z_j \ge \Phi^{-1}(1 - \alpha/m).$$

PROPOSITION 2.3 (Null scale of the maximum). *If $Z_1, \dots, Z_m$ are independent $N(0,1)$, then*

$$\frac{\max_{1 \le j \le m} Z_j}{\sqrt{2 \log m}} \xrightarrow{P} 1.$$

PROOF. Both directions use Mills' ratio, which states that for $x > 0$,

$$\frac{x}{x^2+1}\,\phi(x) \le 1 - \Phi(x) \le \frac{\phi(x)}{x}, \tag{1}$$

where ϕ is the standard normal density. For the upper tail, use the union bound:

$$\mathbb{P}\left(\max_j Z_j > (1+\epsilon)\sqrt{2\log m}\right) \le m\{1 - \Phi((1+\epsilon)\sqrt{2\log m})\}.$$

The upper bound in (1) gives $m\{1 - \Phi((1+\epsilon)\sqrt{2\log m})\} \le m^{1-(1+\epsilon)^2}/\sqrt{4\pi \log m} \to 0$. Here the sharper Mills prefactor contains an additional factor $(1+\epsilon)^{-1}$, which is at most one and has been dropped. For the lower tail, independence gives

$$\mathbb{P}\left(\max_j Z_j \le (1-\epsilon)\sqrt{2\log m}\right) = \Phi((1-\epsilon)\sqrt{2\log m})^m = \left[1 - \{1 - \Phi((1-\epsilon)\sqrt{2\log m})\}\right]^m.$$

The lower bound in (1) gives

$$m\{1 - \Phi((1-\epsilon)\sqrt{2\log m})\} \ge c_\epsilon \frac{m^{1-(1-\epsilon)^2}}{\sqrt{\log m}} \to \infty$$

for a constant $c_\epsilon > 0$. If $u_m = 1 - \Phi((1-\epsilon)\sqrt{2\log m})$, then $(1-u_m)^m \le \exp(-m u_m) \to 0$, so the lower-tail probability tends to zero. □

The threshold $\Phi^{-1}(1-\alpha/m)$ is also asymptotic to $\sqrt{2\log m}$: from $1 - \Phi(t) \sim \phi(t)/t$ one solves $\phi(t)/t = \alpha/m$ to find $t = \sqrt{2\log m}\{1 + o(1)\}$. Thus the maximum test looks for at least one coordinate whose mean is on the extreme-noise scale.

4.2. One Sparse Signal. Suppose exactly one coordinate has mean μ^*, while all others have mean zero. If

$$\mu^* = (1+\epsilon)\sqrt{2\log m},$$

then the maximum test has power tending to one. The signal coordinate alone crosses the null extreme-value threshold with probability tending to one.

If instead

$$\mu^* = (1-\epsilon)\sqrt{2\log m},$$

the signal is below the maximum-noise scale. Let j^* be the index of the signal coordinate and write $Z_{j^*} = \mu^* + \xi_{j^*}$ with $\xi_{j^*} \sim N(0,1)$. Using independence,

$$\begin{aligned}\mathbb{P}(\max_j Z_j \le \Phi^{-1}(1-\alpha/m)) &= \mathbb{P}(\xi_{j^*} \le \Phi^{-1}(1-\alpha/m) - \mu^*) \prod_{j\neq j^*} \mathbb{P}(Z_j \le \Phi^{-1}(1-\alpha/m)) \\ &= \mathbb{P}(\xi_{j^*} \le \epsilon\sqrt{2\log m}\{1+o(1)\}) \cdot (1-\alpha/m)^{m-1} \\ &\longrightarrow 1 \cdot e^{-\alpha} = e^{-\alpha}.\end{aligned}$$

The first factor tends to one because $\Phi^{-1}(1-\alpha/m) - \mu^* = \epsilon\sqrt{2\log m}\{1+o(1)\} \to \infty$; the second factor uses the same limit as under the null. Hence the rejection probability tends to $1-e^{-\alpha}$, the same limit as under the null. For small α, $1-e^{-\alpha} = \alpha + O(\alpha^2)$, so this limiting power is only the nominal size up to the usual Bonferroni conservativeness. In this sense, the maximum test cannot reliably detect a single signal below the $\sqrt{2\log m}$ scale.

4.3. No Other Test Can Beat This Boundary. The preceding statement is about one test. We now show that no test of level α can have asymptotic power against the one-sparse alternative when $\mu^* < (1-\epsilon)\sqrt{2\log m}$, so the boundary is not merely a limitation of Bonferroni. The standard route is a Bayesian least-favorable prior: an alternative under which the data are nearly indistinguishable from the null.

Choose J uniformly from $\{1,\dots,m\}$, set $\mu_J = \mu^*$, and set all other means to zero. Under the null, the density of $Z = (Z_1,\dots,Z_m)$ is f_0. Under this mixture alternative, the likelihood ratio is

$$L_m(Z) = \frac{1}{m}\sum_{j=1}^{m} \exp\left(\mu^* Z_j - \frac{(\mu^*)^2}{2}\right).$$

Proposition 2.4 (Sparse one-signal lower bound). *If $\mu^* \le (1-\epsilon)\sqrt{2\log m}$ for some fixed $\epsilon > 0$, then $L_m(Z) \xrightarrow{P_0} 1$. Consequently, for every sequence of level-α rejection regions A_m, the asymptotic power under the mixture alternative satisfies*

$$P_1(A_m) \le \alpha + o(1).$$

Proof. The mean $E_0 L_m = 1$ is automatic from the likelihood-ratio form. For convergence in probability, a direct second-moment calculation gives

$$E_0 L_m^2 = 1 - \frac{1}{m} + \frac{e^{(\mu^*)^2}}{m} \le 1 + \frac{m^{2(1-\epsilon)^2} - 1}{m},$$

because the $m(m-1)$ off-diagonal terms have expectation 1, while the m diagonal terms have expectation $e^{(\mu^*)^2}$. The displayed bound is $1+o(1)$ when $2(1-\epsilon)^2 < 1$, i.e., when ϵ is sufficiently large: $\epsilon > 1 - 1/\sqrt{2}$. For smaller ϵ, μ^* is close to $\sqrt{2\log m}$ and the rare extreme terms $\exp(\mu^* Z_j - (\mu^*)^2/2)$ have heavy upper tails that inflate the second moment. We now carry out the truncation.

Write

$$r_m = \frac{(\mu^*)^2}{2\log m} \le (1-\epsilon)^2 < 1, \qquad Y_{mj} = \exp\{\mu^* Z_j - (\mu^*)^2/2\},$$

so $L_m = m^{-1}\sum_j Y_{mj}$. Let $a_m = \sqrt{2\log m}$ and truncate at the null maximum scale:

$$\widetilde{Y}_{mj} = Y_{mj}\mathbf{1}\{Z_j \le a_m\}, \qquad \widetilde{L}_m = m^{-1}\sum_{j=1}^{m}\widetilde{Y}_{mj}.$$

The likelihood-ratio identity for a single shifted normal gives

$$E_0\{Y_{m1}\mathbf{1}(Z_1 > a_m)\} = \mathbb{P}\{N(\mu^*,1) > a_m\} \le 1 - \Phi\{(1-\sqrt{r_m})\sqrt{2\log m}\} \to 0.$$

Hence $E_0\widetilde{Y}_{m1} = 1 - o(1)$, and

$$E_0|L_m - \widetilde{L}_m| = E_0\{Y_{m1}\mathbf{1}(Z_1 > a_m)\} \to 0.$$

It remains to show that $\widetilde{L}_m$ concentrates around its mean. A second likelihood-ratio calculation gives

$$E_0\widetilde{Y}_{m1}^2 = e^{(\mu^*)^2}\mathbb{P}\{N(2\mu^*,1) \le a_m\}.$$

If $r_m \le 1/4$, this is at most $e^{(\mu^*)^2} \le m^{1/2+o(1)}$, so $m^{-1}E_0\widetilde{Y}_{m1}^2 \to 0$. If $r_m > 1/4$, Mills' ratio gives

$$\mathbb{P}\{N(2\mu^*,1) \le a_m\} \le \exp\{-(2\sqrt{r_m}-1)^2\log m\}\cdot O(1),$$

up to a polynomial factor in $\log m$. Therefore

$$\frac{1}{m}E_0\widetilde{Y}_{m1}^2 \le m^{-1+2r_m-(2\sqrt{r_m}-1)^2+o(1)} = m^{-2(1-\sqrt{r_m})^2+o(1)} \to 0.$$

Since the truncated summands are independent,

$$\mathrm{Var}_0(\widetilde{L}_m) \le \frac{1}{m}E_0\widetilde{Y}_{m1}^2 \to 0.$$

Thus $\widetilde{L}_m - E_0\widetilde{L}_m \to 0$ in P_0-probability, while $E_0\widetilde{L}_m = 1 - o(1)$. Combining this with $E_0|L_m - \widetilde{L}_m| \to 0$ yields $L_m \xrightarrow{P_0} 1$. The same L^1 bound also gives $E_0|L_m - 1| \to 0$.

For any rejection region A_m,

$$P_1(A_m) - P_0(A_m) = E_0[(L_m - 1)\mathbf{1}_{A_m}] \longrightarrow 0,$$

which gives the claim. □

The maximum test is therefore rate-optimal for one sparse coordinate: no other procedure can asymptotically distinguish the null from the alternative when $\mu^* \le (1-\epsilon)\sqrt{2\log m}$.

5. L_2 Tests for Dense Signals

Sparse alternatives are not the only scientifically important regime. Many studies produce weak shifts in many coordinates: a small but real effect on a pathway of correlated genes, a fractional shift on most subgroups of a treatment effect, or a modest improvement of a model over a baseline on most of a benchmark. A maximum test may miss such a pattern because no single coordinate is large.

The natural competing statistic is the squared Euclidean norm

$$T_m = \|Z\|_2^2 = \sum_{j=1}^{m} Z_j^2,$$

the classical chi-square statistic for testing $\mu = 0$ against $\mu \ne 0$ in the Gaussian sequence model. Whereas Bonferroni rewards a single extreme coordinate, T_m rewards aggregate energy: any μ of large enough ℓ_2 norm will inflate T_m above its null mean, regardless of how that norm is distributed across coordinates.

Under the global null, $T_m \sim \chi^2_m$, and

$$\frac{T_m - m}{\sqrt{2m}} \xrightarrow{d} N(0,1).$$

Thus an asymptotic level-α L_2 test rejects when

$$\frac{T_m - m}{\sqrt{2m}} \geq z_{1-\alpha}.$$

Under a fixed mean vector μ,

$$T_m = \sum_{j=1}^{m} \xi_j^2 + \|\mu\|_2^2 + 2\sum_{j=1}^{m} \mu_j \xi_j.$$

If $\|\mu\|_2 = o(\sqrt{m})$, then

$$\frac{T_m - \{m + \|\mu\|_2^2\}}{\sqrt{2m}} \xrightarrow{d} N(0,1).$$

More generally,

$$\frac{T_m - \{m + \|\mu\|_2^2\}}{\sqrt{2m + 4\|\mu\|_2^2}} \xrightarrow{d} N(0,1).$$

Define

$$\theta_m = \frac{\|\mu\|_2^2}{\sqrt{2m}}.$$

Using the preceding normal approximation, the power of the L_2 test is approximately

$$1 - \Phi\left(\frac{z_{1-\alpha} - \theta_m}{\sqrt{1 + \theta_m/\sqrt{m/8}}}\right).$$

Consequently:

$$\begin{aligned}
\theta_m = o(1) \quad &\Longrightarrow \quad \text{power} \to \alpha, \\
\theta_m \to c < \infty \quad &\Longrightarrow \quad \text{power} \to 1 - \Phi(z_{1-\alpha} - c), \\
\theta_m \to \infty \quad &\Longrightarrow \quad \text{power} \to 1.
\end{aligned}$$

5.1. A Lower Bound for Dense Alternatives. The L_2 detection threshold is also optimal for dense, direction-unknown signals. Consider the testing problem

$$H_0 : \mu = 0 \qquad \text{versus} \qquad H_1 : \mu = \rho U,$$

where U is drawn from the uniform distribution on the unit sphere $\mathcal{S}^{m-1}$. After integrating over U, the likelihood ratio is

$$L_m(Z) = \int_{\mathcal{S}^{m-1}} \exp\left(-\frac{\rho^2}{2} + \rho Z^\top u\right) d\pi_1^{(m)}(u),$$

where $\pi_1^{(m)}$ is uniform measure on the unit sphere.

If

$$\frac{\rho^2}{\sqrt{2m}} \to 0,$$

then $L_m \to 1$ in probability under the null. The calculation is the same second-moment argument used in Le Cam's first lemma. Let P_m denote the null law, Q_m the spherical mixture alternative, and $L_m = dQ_m/dP_m$. Clearly $E_0 L_m = 1$. We now show that $E_0 L_m^2 \to 1$, which implies $L_m \to 1$ in $L^2(P_m)$, hence also in probability under P_m.

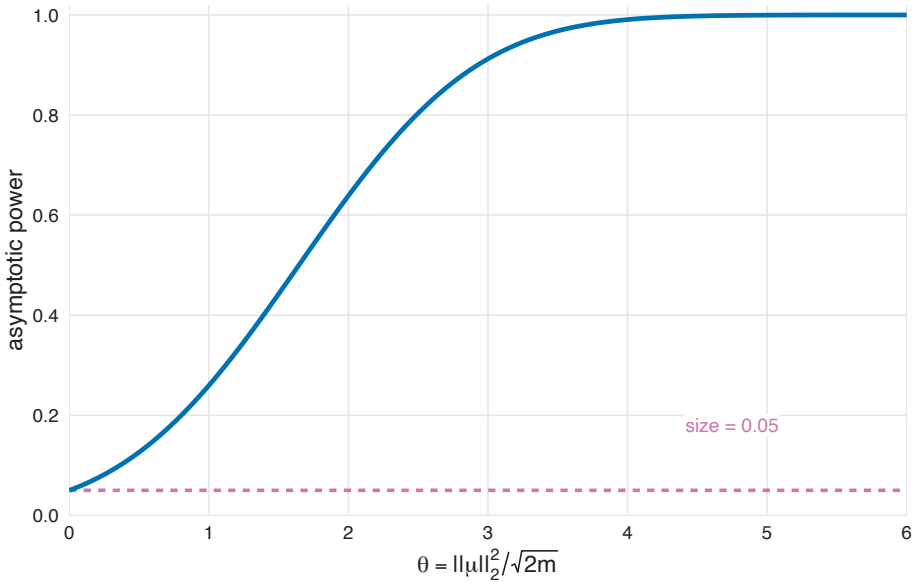

FIGURE 3. Asymptotic power curve of the L_2 test as a function of $\theta = \|\mu\|_2^2/\sqrt{2m}$, for $\alpha = 0.05$. Dense alternatives are visible once the total squared signal is on the $\sqrt{m}$ scale.

Under P_m, $Z \sim N(0, I_m)$. Using $E_0 \exp(t^\top Z) = \exp(\|t\|_2^2/2)$,

$$\begin{aligned} E_0 L_m^2 &= \int_{\mathcal{S}^{m-1}} \int_{\mathcal{S}^{m-1}} E_0 \exp\left(-\rho^2 + \rho Z^\top u + \rho Z^\top v\right) d\pi_1^{(m)}(u) d\pi_1^{(m)}(v) \\ &= \int_{\mathcal{S}^{m-1}} \int_{\mathcal{S}^{m-1}} \exp\left(-\rho^2 + \frac{\rho^2}{2}\|u+v\|_2^2\right) d\pi_1^{(m)}(u) d\pi_1^{(m)}(v) \\ &= \int_{\mathcal{S}^{m-1}} \int_{\mathcal{S}^{m-1}} \exp(\rho^2 u^\top v) d\pi_1^{(m)}(u) d\pi_1^{(m)}(v). \end{aligned}$$

For fixed v, choose an orthogonal matrix B such that $Bv = e_1$. Since uniform measure on the sphere is rotation invariant, Bu is again uniform on $\mathcal{S}^{m-1}$. Therefore

$$\int_{\mathcal{S}^{m-1}} \exp(\rho^2 u^\top v) d\pi_1^{(m)}(u) = \int_{\mathcal{S}^{m-1}} \exp(\rho^2 u_1) d\pi_1^{(m)}(u),$$

which does not depend on v. Hence

$$E_0 L_m^2 = E \exp(\rho^2 U_1),$$

where U_1 is the first coordinate of a uniform point on $\mathcal{S}^{m-1}$. This is the origin of the displayed identity.

It remains to expand the last expectation. By symmetry, odd moments of U_1 vanish, and

$$E[U_1^{2k}] = \frac{(2k-1)!!}{m(m+2)\cdots(m+2k-2)}, \qquad k \geq 1.$$

Thus

$$E\exp(\rho^2 U_1) = 1 + \frac{\rho^4}{2}E[U_1^2] + \sum_{k\geq 2}\frac{\rho^{4k}}{(2k)!}E[U_1^{2k}]$$
$$= 1 + \frac{\rho^4}{2m} + \sum_{k\geq 2}\frac{\rho^{4k}}{(2k)!}\frac{(2k-1)!!}{m(m+2)\cdots(m+2k-2)}.$$

Writing $\eta_m = \rho^4/m$, the remainder is bounded by

$$\sum_{k\geq 2}\frac{\eta_m^k}{2^k k!} = O(\eta_m^2),$$

because $(2k)! = 2^k k!(2k-1)!!$. Therefore, if $\eta_m = \rho^4/m \to 0$, equivalently $\rho^2/\sqrt{2m} \to 0$ up to constants, then

$$E_0 L_m^2 = 1 + \frac{\rho^4}{2m}\{1 + o(1)\} = 1 + o(1).$$

Consequently $E_0(L_m - 1)^2 = E_0 L_m^2 - 1 \to 0$. For any rejection region A_m,

$$|Q_m(A_m) - P_m(A_m)| = |E_0\{(L_m - 1)\mathbf{1}_{A_m}\}| \leq \{E_0(L_m - 1)^2\}^{1/2} \to 0.$$

Hence no test can reliably separate the null from these spherical alternatives below the L_2 threshold.

The next proposition records exactly which Le Cam facts are being used. The first item is a convenient sufficient form of Le Cam's first lemma. In the usual statement, if $L_m = dQ_m/dP_m$ converges in distribution under P_m to a nonnegative random variable V, then Q_m is contiguous with respect to P_m precisely when $EV = 1$. Here $L_m \to 1$ in P_m-probability, so the only possible limit is $V = 1$, and the condition is automatic. In the dense lower bound above we proved the stronger fact $L_m \to 1$ in $L^2(P_m)$, which even implies that the total variation distance between P_m and Q_m tends to zero. The second item is Le Cam's third lemma.

PROPOSITION 2.5 (Contiguity tools used in the lower bounds). *Let P_m and Q_m be sequences of probability measures, and let $L_m = dQ_m/dP_m$.*

1. *If $L_m \to 1$ in P_m-probability, then Q_m is contiguous with respect to P_m. Thus every event whose P_m-probability tends to zero also has Q_m-probability tending to zero.*
2. *If $(T_m, \log L_m)$ has a bivariate normal joint limit under P_m, then Le Cam's third lemma describes the limiting distribution of T_m under Q_m by exponentially tilting that joint Gaussian limit.*

In the dense lower bound above, $L_m \to 1$ under P_m. The tilt in Le Cam's third lemma is therefore trivial: any statistic with a null limiting distribution has the same limiting distribution under the spherical alternative. Such a statistic cannot yield asymptotic power above its size. This is the formal version of the phrase "the two experiments are asymptotically indistinguishable."

6. Comparing the Maximum and L_2 Tests

The maximum test and the L_2 test measure different features of μ:

$$\text{maximum test:} \quad \|\mu\|_\infty, \qquad L_2 \text{ test:} \quad \frac{\|\mu\|_2^2}{\sqrt{2m}}.$$

Neither dominates the other.

EXAMPLE 2.6 (Many strong but still sparse signals). Suppose $m^{3/8}$ coordinates of μ equal $1.1\sqrt{2\log m}$, and all other coordinates are zero. (The exponent $3/8$ is chosen so that the signal count grows fast enough for the maximum to dominate but slowly enough that the L_2 energy stays below the detection threshold; any exponent strictly less than $1/2$ would have the same qualitative effect.) The maximum test has power tending to one because at least one coordinate is above the extreme-noise scale. But

$$\theta_m = \frac{m^{3/8} \cdot 1.1^2(2\log m)}{\sqrt{2m}} = \frac{2.42\ \log m}{\sqrt{2}\, m^{1/8}} \to 0,$$

so the L_2 test is asymptotically powerless.

EXAMPLE 2.7 (Many weak signals). Suppose $\sqrt{2m}$ coordinates of μ equal a fixed constant a, and the rest are zero. Then $\|\mu\|_\infty = a$, which is far below $\sqrt{2\log m}$, so the maximum test has no asymptotic power beyond its null rejection probability. But $\theta_m = a^2$, and the L_2 test has limiting power

$$1 - \Phi(z_{1-\alpha} - a^2).$$

For example, at $\alpha = 0.05$ and $a = 2$, the limiting power is approximately 0.991.

TABLE 1. Monte Carlo rejection probabilities for four global tests.

Regime	Signals	Effect	Bonferroni	Fisher	Cauchy	L_2
Sparse strong	3	3.0	0.476	0.182	0.492	0.216
Moderately sparse	30	2.0	0.606	0.943	0.746	0.963
Dense weak	250	0.5	0.098	0.566	0.126	0.574

7. Assumptions in Plain Language

The p-value procedures in this chapter need valid null p-values. Bonferroni needs no dependence model; Fisher, Pearson, Stouffer, and the exact L_2 calibrations do. The Cauchy combination test uses a tail approximation that is much less sensitive to Gaussian dependence, but it is still a calibration statement that should be checked when the p-values come from a different model. The Gaussian sequence lower bounds are asymptotic statements for simplified models. Their role is to explain why maximum tests are natural for rare strong signals and why energy tests are natural for many weak signals, not to claim that one test is best in every finite-sample study.

8. Bibliographic Notes

Fisher's combination test is classical [46], and Stouffer's normal-score combination was developed in the social-science meta-analysis literature [115]. Sidak's exact independent-threshold calibration is due to Šidák [108]. The Cauchy combination test and its dependence-robust tail approximation are from Liu and Xie [87]. The contiguity arguments used in the lower bounds are Le Cam arguments; see van der Vaart [126, Chapter 6], Le Cam [75], and Le Cam and Yang [76]. Ingster's work on Gaussian sequence testing made precise how sparse alternatives can lead to non-Gaussian and infinitely divisible likelihood-ratio limits [64]; modern high-dimensional detection-boundary treatments include Donoho and Jin [38] and Arias-Castro et al. [4].

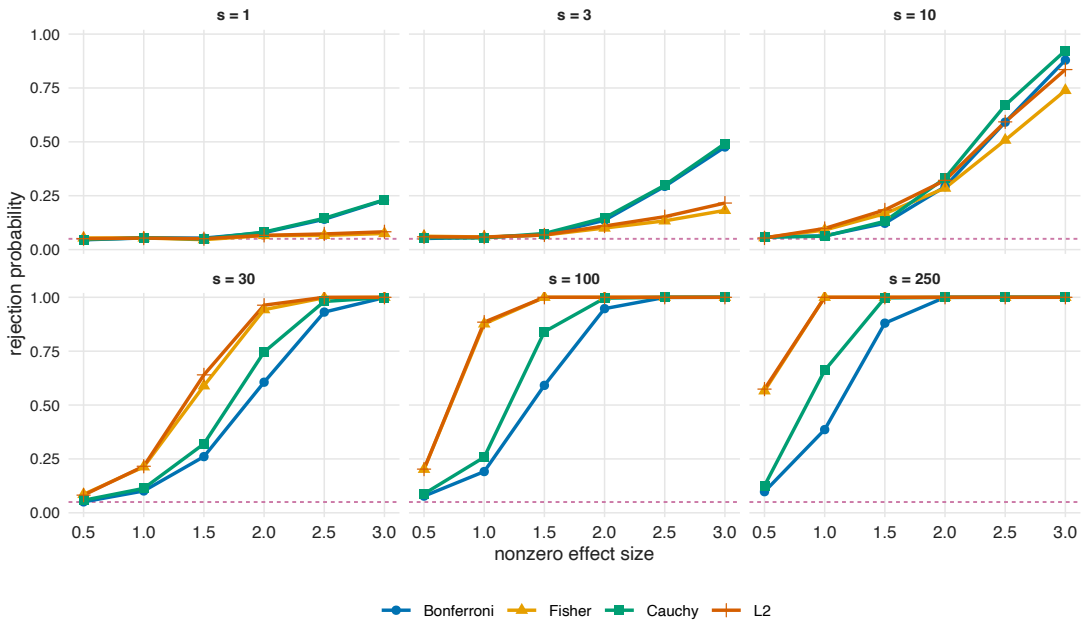

FIGURE 4. Estimated rejection probabilities at level $\alpha = 0.05$ in a Gaussian sequence model with $m = 500$ coordinates. Each panel fixes the number s of nonzero means and varies their common effect size; each curve is based on 1500 Monte Carlo replicates. Bonferroni reacts fastest when s is small and effects are strong. Fisher and L_2-type aggregation gain power as evidence is spread over many coordinates, while Cauchy remains closer to the extreme-value side.

9. Exercises

Basic.

EXERCISE 2.8 (Bonferroni and Sidak thresholds). For $\alpha = 0.05$ and $m \in \{10, 100, 1000\}$, compute the Bonferroni threshold α/m and the Sidak threshold $1-(1-\alpha)^{1/m}$. Verify numerically that they agree to leading order and that the Bonferroni size under independence is close to $1 - e^{-\alpha}$.

EXERCISE 2.9 (Fisher's chi-square distribution). Let $p_1, \dots, p_5$ be independent $\mathrm{Unif}(0, 1)$. Compute the distribution of $T_F = -2\sum_{j=1}^{5} \log p_j$ and find the 0.95 quantile of T_F. For the observed p-values $0.02, 0.18, 0.07, 0.30, 0.12$, decide whether Fisher's test rejects at level 0.05.

EXERCISE 2.10 (Stouffer's z-score). Convert the p-values $0.01, 0.04, 0.20, 0.32, 0.50$ to one-sided z-scores $X_j = \Phi^{-1}(1 - p_j)$ and compute Stouffer's combined statistic $Z_S = \sum_j X_j/\sqrt{5}$. Compare its p-value with Fisher's combined p-value on the same inputs.

EXERCISE 2.11 (Two-sided p-value calibration). Let $Z \sim N(\mu, 1)$ and let $p = 2\{1 - \Phi(|Z|)\}$ be the two-sided p-value. Show that $p \sim \mathrm{Unif}(0, 1)$ under $\mu = 0$. Then compute $\mathbb{P}_\mu(p \le 0.05)$ for $\mu = 1, 2, 3$ and interpret the resulting power values.

Intermediate.

EXERCISE 2.12 (Maximum scale). Use Mills' ratio to prove

$$\frac{\max_{1\le j\le m} Z_j}{\sqrt{2\log m}} \xrightarrow{P} 1$$

for independent standard normal Z_j's. Then show that $\Phi^{-1}(1-\alpha/m) = \sqrt{2\log m}\{1+o(1)\}$.

EXERCISE 2.13 (L_2 normal approximation). In the Gaussian sequence model, prove that if $\|\mu\|_2 = o(\sqrt{m})$, then

$$\frac{T_m - \{m + \|\mu\|_2^2\}}{\sqrt{2m}} \xrightarrow{d} N(0,1).$$

Then prove the more general approximation with denominator $\sqrt{2m + 4\|\mu\|_2^2}$. Use the fact that $E\xi_j^3 = 0$.

EXERCISE 2.14 (Sparse mixture likelihood ratio). For the one-sparse mixture alternative with $\mu^* = (1-\epsilon)\sqrt{2\log m}$, derive the likelihood ratio

$$L_m(Z) = m^{-1}\sum_{j=1}^{m} \exp\{\mu^* Z_j - (\mu^*)^2/2\}.$$

Show that $E_0 L_m = 1$. For which values of ϵ does the elementary second-moment argument prove $L_m \to 1$ in $L^2(P_0)$? Why is a truncation argument needed for the full range $\epsilon > 0$?

EXERCISE 2.15 (Cauchy combination under independence and under dependence). Suppose $p_1, \dots, p_m$ are independent Unif$(0,1)$, so that the transformed quantities $C_j = \tan\{\pi(1/2 - p_j)\}$ are iid standard Cauchy. For any nonnegative weights w_j with $\sum_j w_j = 1$, use the stability of the Cauchy distribution under positive-weight linear combinations to prove that $W = \sum_j w_j C_j$ is *exactly* standard Cauchy.

Now suppose $p_1, \dots, p_m$ come from jointly normal test statistics with arbitrary covariance. Following Liu and Xie [87], illustrate numerically that W is no longer exactly Cauchy by comparing its empirical distribution to the standard Cauchy in the body of the distribution, and verify that the upper-tail approximation $\mathbb{P}(W \ge t) \approx 1/(\pi t)$ becomes accurate as the threshold t grows large. (A simulation cannot prove an asymptotic tail equivalence; for the proof of $\mathbb{P}(W \ge t) = \{1/(\pi t)\}\{1+o(1)\}$ under broad dependence, see Liu and Xie [87].) Explain why this tail equivalence is the genuinely useful result: it justifies the Cauchy calibration of W under broad dependence, where the exact distribution is intractable.

Computational.

EXERCISE 2.16 (Bonferroni under equicorrelation). Let $Z \sim N(0, \Sigma_\rho)$, where $(\Sigma_\rho)_{jj} = 1$ and $(\Sigma_\rho)_{jk} = \rho$ for $j \ne k$. For $\rho \in \{0, 0.1, \dots, 0.9\}$, estimate the size of the two-sided Bonferroni global test at level 0.05 by Monte Carlo with at least 5000 replicates. Compare your output with Figure 1. Explain why positive correlation changes the observed size but not the validity guarantee.

EXERCISE 2.17 (Sparse versus dense simulation). Reproduce Figure 4. Vary the number of nonzero coordinates and the common effect size. Compare Bonferroni, Fisher, Cauchy, and the L_2 test. Summarize the region in which each procedure is strongest.

EXERCISE 2.18 (Fisher under dependence). Simulate equicorrelated Gaussian z-scores under the global null and compute two-sided p-values. Apply Fisher's test using the independent chi-square calibration. How does the Type I error change with correlation? Compare this behavior with Bonferroni.

Advanced.

EXERCISE 2.19 (Dense lower bound). For the spherical prior alternative $\mu = \rho U$, derive the likelihood ratio and verify that

$$E_0 L_m^2 = E\exp(\rho^2 U_1),$$

where U_1 is the first coordinate of a uniform point on $\mathcal{S}^{m-1}$. Use this to show that the likelihood ratio converges to one in P_0-probability when $\rho^2/\sqrt{2m} \to 0$, so no test can have asymptotic power above its size against the spherical mixture. Then show the converse: when $\rho^2/\sqrt{2m} \to \infty$, the L_2 test itself has power tending to one. (Note that the likelihood ratio cannot diverge under P_0 — its P_0-mean equals 1; divergence is a statement under the alternative P_1.)

Exercise 2.20 (Truncation for the sparse lower bound). Complete the truncation step in Proposition 2.4. Define the truncated likelihood ratio $\widetilde{L}_m = m^{-1} \sum_j \exp(\mu^* Z_j - (\mu^*)^2/2)\, \mathbf{1}\{Z_j \le c_m\}$ for a sequence $c_m \to \infty$, choose c_m so the second moment is $1 + o(1)$, and bound the tail contribution $E_0|L_m - \widetilde{L}_m|$.

CHAPTER 3

Simes, Goodness-of-Fit, and Higher Criticism

Chapter 2 compared global tests by asking whether evidence is concentrated in one coordinate or spread across many coordinates. This chapter gives a second view of the same question. Instead of starting from z-scores, we start from the empirical distribution of p-values. Under a clean global null, continuous p-values look uniform. A global test can therefore be read as a goodness-of-fit test for the uniform distribution, with different procedures looking at different parts of the empirical CDF.

This viewpoint is useful because it makes the geometry of rare signals visible. If a few alternatives produce very small p-values, the empirical CDF rises above the diagonal near zero. If many weak alternatives are present, the deviation may be spread over a wider range. Simes, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, higher criticism, Berk-Jones, and average likelihood ratio tests are different ways of measuring this upward departure.

1. A Uniformity Problem

Suppose a large monitoring system produces m one-sided p-values

$$p_1, \dots, p_m.$$

The system might be a scientific screen, an anomaly detector, or a benchmark audit. For each coordinate j, small p_j is evidence against the coordinate null. The global null says that all coordinates are null. Under the idealized version used in this chapter,

$$p_1, \dots, p_m \overset{\text{iid}}{\sim} \text{Unif}(0,1) \qquad \text{under } H_0.$$

Write the order statistics as

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$$

and define the empirical CDF

$$F_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{p_j \le t\}, \qquad 0 \le t \le 1.$$

Under the global null, $F_m(t)$ fluctuates around t. Under alternatives that create too many small p-values, $F_m(t)$ exceeds t for small or moderate t.

This picture also clarifies why no single global test is uniformly best. A test that averages deviations over the whole unit interval can be powerful for broad departures but inefficient for a tiny bump near zero. A test that looks only at the smallest p-value is excellent for one extreme coordinate but can miss many modest ones. Tail-adaptive procedures try to avoid choosing the scale of the departure in advance.

2. Simes as a Crossing Rule

The Simes test rejects the global null when at least one ordered p-value crosses the line $i\alpha/m$:

$$p_{(i)} \le \frac{i\alpha}{m} \quad \text{for some } 1 \le i \le m.$$

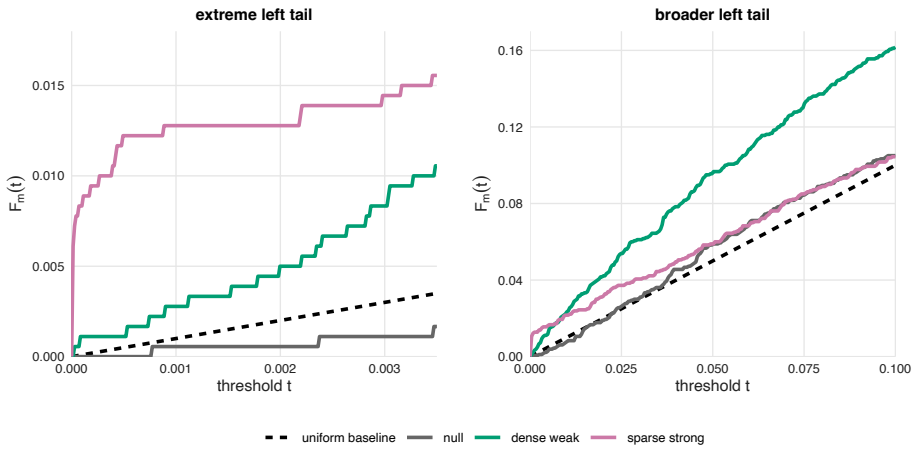

FIGURE 1. Left-tail empirical CDFs of one-sided p-values under a null model, a dense weak alternative, and a sparse stronger alternative. In this stylized draw, $m = 1800$, the dense weak case shifts 360 coordinates by 1.0, and the sparse strong case shifts 24 coordinates by 4.2. The diagonal is the uniform baseline. The extreme-left panel shows the sparse alternative creating the sharpest bump very near zero; the broader-left panel shows the dense weak alternative lifting the CDF over a wider range.

Equivalently, define the Simes p-value

$$p_{\text{Simes}} = \min_{1 \le i \le m} \frac{m p_{(i)}}{i}.$$

The level-α Simes test rejects when $p_{\text{Simes}} \le \alpha$.

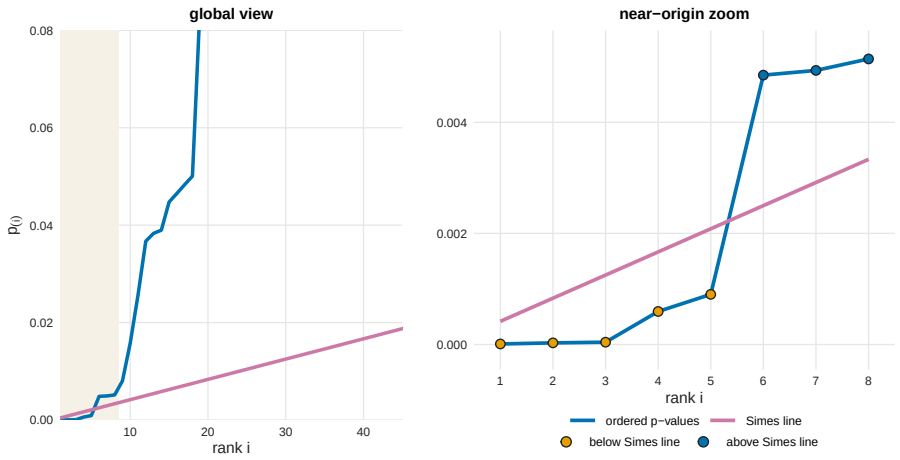

FIGURE 2. The Simes crossing rule. Ordered p-values are compared with the line $\alpha i/m$. A crossing gives a global rejection. Later, the same linear boundary becomes the basis of the BH multiple-testing procedure.

Simes is a global test in this chapter. It does not by itself say which hypotheses are false. The same boundary will reappear in Chapters 4 and 5, where it is used inside closure and FDR procedures to produce individual discoveries.

THEOREM 3.1 (Simes under independent uniforms). *If $p_1, \dots, p_m$ are independent* Unif$(0,1)$ *random variables, then*

$$\mathbb{P}\left\{\min_{1\le i\le m} \frac{mp_{(i)}}{i} \le \alpha\right\} = \alpha, \qquad 0 \le \alpha \le 1.$$

PROOF. Let

$$T_m = \min_{1\le i\le m} \frac{mp_{(i)}}{i}.$$

We prove by induction that $T_m \sim \text{Unif}(0,1)$. The case $m = 1$ is immediate.

Assume the result is true for $m-1$. Fix $0 < \alpha < 1$ and split the event $\{T_m \le \alpha\}$ according to whether the largest order statistic is at most α:

$$\mathbb{P}(T_m \le \alpha) = \mathbb{P}\left(\min_{1\le i\le m-1} \frac{mp_{(i)}}{i} \le \alpha,\ p_{(m)} > \alpha\right) + \mathbb{P}\{p_{(m)} \le \alpha\}.$$

The second term is α^m. The density of $p_{(m)}$ is mt^{m-1}, $0 < t < 1$. Conditional on $p_{(m)} = t$, the remaining $m-1$ order statistics have the same distribution as the order statistics of $m-1$ independent uniforms on $[0,t]$. Equivalently,

$$\left(\frac{p_{(1)}}{t}, \dots, \frac{p_{(m-1)}}{t}\right)$$

has the order-statistic distribution of $m-1$ independent Unif$(0,1)$ variables. Therefore, by the induction hypothesis,

$$\min_{1\le i\le m-1} \frac{(m-1)p_{(i)}}{t\,i} \sim \text{Unif}(0,1) \qquad \text{conditional on } p_{(m)} = t.$$

On the event $p_{(m)} > \alpha$, the crossing among the first $m-1$ order statistics is equivalent to

$$\min_{1\le i\le m-1} \frac{(m-1)p_{(i)}}{t\,i} \le \frac{(m-1)\alpha}{mt}.$$

For $t \in (\alpha, 1)$, the right-hand side lies in $(0,1)$. Hence

$$\begin{aligned}&\mathbb{P}\left(\min_{1\le i\le m-1} \frac{mp_{(i)}}{i} \le \alpha,\ p_{(m)} > \alpha\right)\\ &\quad= \int_\alpha^1 mt^{m-1}\,\frac{(m-1)\alpha}{mt}\,dt = (m-1)\alpha\int_\alpha^1 t^{m-2}\,dt = \alpha(1-\alpha^{m-1}).\end{aligned}$$

Adding the case $p_{(m)} \le \alpha$ gives

$$\mathbb{P}(T_m \le \alpha) = \alpha(1-\alpha^{m-1}) + \alpha^m = \alpha.$$

Thus T_m is exactly uniform on $[0,1]$. □

The exact statement uses independence and continuity. Under certain positive dependence conditions, Simes remains conservative rather than exact. That dependence issue is one of the main topics of Chapters 4 and 6.

3. Empirical Processes and Classical Goodness-of-Fit Tests

Under independent uniform p-values, the centered empirical process

$$\alpha_m(t) = \sqrt{m}\{F_m(t) - t\}$$

converges weakly to a Brownian bridge $B(t)$, a mean-zero Gaussian process with covariance

$$\mathrm{Cov}\{B(s), B(t)\} = \min(s,t) - st.$$

This limit is the organizing principle behind classical uniformity tests.

The one-sided Kolmogorov-Smirnov statistic is

$$D_m^+ = \sup_{0\le t\le 1} \{F_m(t) - t\}.$$

It looks for the largest vertical gap above the diagonal. The two-sided version uses $\sup_t |F_m(t)-t|$, but for detecting an excess of small p-values the one-sided statistic is the relevant form.

The Cramer-von Mises statistic averages squared deviations:

$$W_m = m \int_0^1 \{F_m(t) - t\}^2 \, dt.$$

A one-sided version may replace the squared deviation by $\{F_m(t)-t\}_+^2$. This averaging makes the test less dependent on a single rank than KS, but it can dilute a very narrow left-tail departure.

The Anderson-Darling statistic adds tail weights:

$$A_m = m \int_0^1 \frac{\{F_m(t) - t\}^2}{t(1-t)} \, dt.$$

Near zero, the binomial variance of $F_m(t)$ is approximately t/m, so division by $t(1-t)$ gives more attention to the small-p-value region. This is the first hint of higher criticism: if the alternatives are rare, the left tail should be standardized more aggressively than the center of the distribution.

For computation, A_m has a useful ordered-p-value form. Splitting the integral over the intervals between order statistics gives

$$A_m = -m - \frac{1}{m}\sum_{i=1}^{m}(2i-1)\left\{\log p_{(i)} + \log(1 - p_{(m+1-i)})\right\}.$$

This formula is also good intuition: the logarithms heavily penalize unusually small $p_{(i)}$'s and unusually large $p_{(m+1-i)}$'s. In the one-sided signal-detection setting we mostly care about the small-p-value contribution, but the classical Anderson-Darling statistic is symmetric in the two tails.

PROPOSITION 3.2 (Brownian bridge covariance). *If $p_1, \ldots, p_m$ are independent uniforms, then for fixed $s, t \in [0,1]$,*

$$\mathrm{Cov}\{\alpha_m(s), \alpha_m(t)\} = \min(s,t) - st.$$

PROOF. For each j, set $I_j(t) = \mathbf{1}\{p_j \le t\}$. Since the observations are independent,

$$\mathrm{Cov}\{F_m(s), F_m(t)\} = \frac{1}{m}\mathrm{Cov}\{I_1(s), I_1(t)\}.$$

For $s \le t$, $I_1(s)I_1(t) = I_1(s)$, so

$$\mathrm{Cov}\{I_1(s), I_1(t)\} = s - st.$$

Multiplying by m gives the covariance of α_m. The case $t < s$ is symmetric. □

4. Higher Criticism, Berk-Jones, and Average Likelihood Ratios

The empirical CDF at a fixed threshold has variance $t(1-t)/m$. This suggests the standardized statistic

$$\frac{\sqrt{m}\{F_m(t) - t\}}{\sqrt{t(1-t)}}.$$

Higher criticism maximizes this quantity over left-tail thresholds, taking only the positive part so that empty crossings do not contribute negative values. A common ordered-p-value form is

$$\mathrm{HC}_m = \max_{1\le i\le \lfloor m/2\rfloor} \left[\frac{\sqrt{m}\{i/m - p_{(i)}\}}{\sqrt{p_{(i)}(1-p_{(i)})}}\right]_+ = \max_{\substack{1\le i\le \lfloor m/2\rfloor \\ p_{(i)}<i/m}} \frac{\sqrt{m}\{i/m - p_{(i)}\}}{\sqrt{p_{(i)}(1-p_{(i)})}}.$$

The two forms agree because only ranks with $p_{(i)} < i/m$ contribute a positive deviation; ranks where the empirical CDF is at or below the diagonal are skipped.

Under the global null, the maximum is not $O(1)$. Each standardized deviation $\sqrt{m}\{F_m(t) - t\}/\sqrt{t(1-t)}$ is approximately $N(0,1)$ for fixed t, but the supremum over the left tail picks up extreme values from a continuum of t's. A law-of-the-iterated-logarithm calculation in the spirit of Darling-Erdős shows that

$$\mathrm{HC}_m = \sqrt{2\log\log m}\,\{1 + o_P(1)\} \qquad \text{under } H_0.$$

The extra $\sqrt{\log\log m}$ factor over the pointwise normal scale is the price of scanning many tail thresholds. In finite samples, the statistic is usually calibrated by simulation, by permutation, or by a conservative asymptotic threshold such as $\sqrt{2(1+\epsilon)\log\log m}$. The important point is that HC scans over many possible tail thresholds without committing to one sparsity level.

Berk-Jones replaces the normal approximation in HC by an exact binomial likelihood ratio. Let

$$h(q,p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}, \qquad 0 < p < q < 1.$$

This is the Kullback-Leibler divergence $D(\mathrm{Bernoulli}(q) \,\|\, \mathrm{Bernoulli}(p))$, so $m\,h(q,p)$ is the log likelihood ratio of $\mathrm{Bin}(m,q)$ versus $\mathrm{Bin}(m,p)$ at the observed count mq. The one-sided Berk-Jones statistic is

$$\mathrm{BJ}_m = \max_{1\le i\le \lfloor m/2\rfloor} m\,h\left(\frac{i}{m}, p_{(i)}\right)\mathbf{1}\left\{p_{(i)} < \frac{i}{m}\right\}.$$

The natural reading of this scan: think first of a *fixed* threshold $t \in (0,1)$, under which $N(t) = \#\{j : p_j \le t\}$ is binomial with success probability t under the global null. Then $N(t)/m$ is its sample proportion, and the binomial log-likelihood ratio comparing the data-driven proportion to the null t is $m\,h(N(t)/m,\, t)$. The BJ statistic plugs in the data-dependent threshold $t = p_{(i)}$ (so that $N(t)/m = i/m$) at each rank i and takes the maximum, scanning over which rank gives the most extreme binomial likelihood-ratio departure.

The average likelihood ratio statistic smooths this idea across ranks:

$$\mathrm{ALR}_m = \sum_{i=1}^{\lfloor m/2\rfloor} w_i \exp\left[m\,h\left(\frac{i}{m}, p_{(i)}\right)\mathbf{1}\left\{p_{(i)} < \frac{i}{m}\right\}\right].$$

The weights w_i keep the average from being dominated by the many ranks available away from the extreme tail. A convenient way to think about the left tail is by dyadic blocks of ranks:

$$i \in [1,2),\ [2,4),\ [4,8), \ldots.$$

There are $O(\log m)$ such resolution levels up to $m/2$, and the block $[2^k, 2^{k+1})$ contains order 2^k ranks. Weights of order $1/i$ therefore give each dyadic block comparable total weight, while the extra $1/\log m$ normalizes across the logarithmic number of blocks. Thus a choice such as

$w_i \propto 1/(i \log m)$ spreads prior weight over tail locations on a scale-invariant grid. HC takes the largest standardized tail deviation, BJ takes the largest binomial likelihood ratio, and ALR averages likelihood ratios across tail locations. These are not cosmetic differences: they behave similarly in the rare/weak asymptotic theory, but in finite samples averaging can be less sensitive to a single noisy rank or cutoff than a pure maximum.

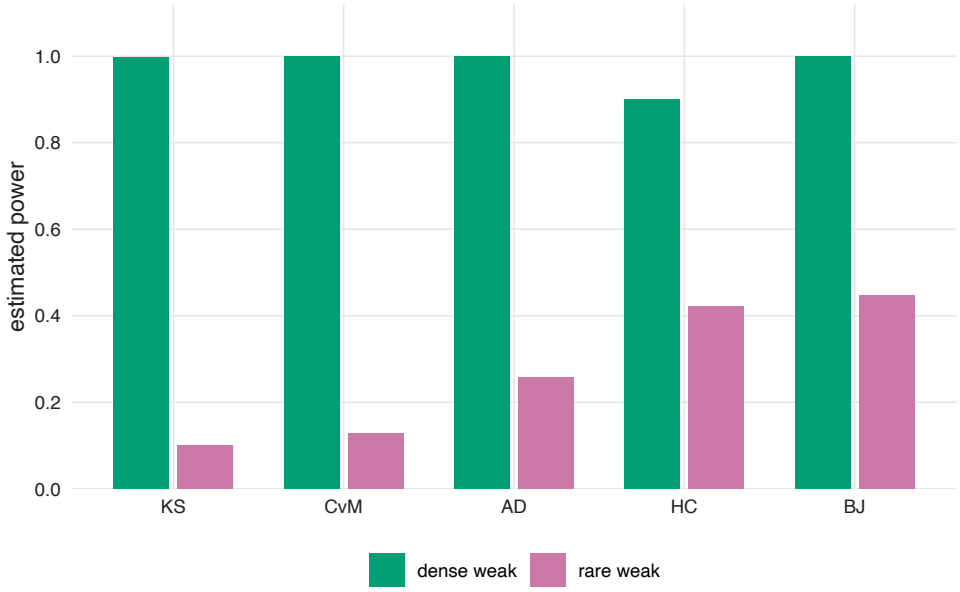

FIGURE 3. Monte Carlo power comparison for one-sided goodness-of-fit tests calibrated under independent uniform p-values. Broad departures favor averaged statistics; rare left-tail departures favor HC and BJ.

5. Rare and Weak Gaussian Mixtures

We now connect the p-value empirical process to the Gaussian sequence model. Let

$$Z_j \sim (1-\varepsilon_m)N(0,1) + \varepsilon_m N(\mu_m, 1), \qquad j = 1, \dots, m,$$

independently, and use one-sided p-values $p_j = 1 - \Phi(Z_j)$. The rare/weak parameterization is

$$\varepsilon_m = m^{-\beta}, \qquad \mu_m = \sqrt{2r \log m},$$

where $1/2 < \beta < 1$ and $r > 0$. The expected number of signals is

$$m\varepsilon_m = m^{1-\beta}.$$

Thus β controls sparsity and r controls signal strength on the extreme-value scale.

The detection boundary for this model is

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2}, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 \le \beta < 1. \end{cases}$$

For $r > \rho^*(\beta)$, there exist level-α tests whose power tends to one. For $r < \rho^*(\beta)$, no sequence of level-α tests is consistent: in the standard formulation of the lower bound, the total-variation distance between the null and the least-favorable mixture under the alternative tends to zero, so the sum of Type I and Type II errors of any sequence of tests cannot fall below one. This is the

Ingster-Donoho-Jin rare/weak boundary; we cite the Donoho-Jin formulation because it is the one most directly connected with higher criticism [38, 64].

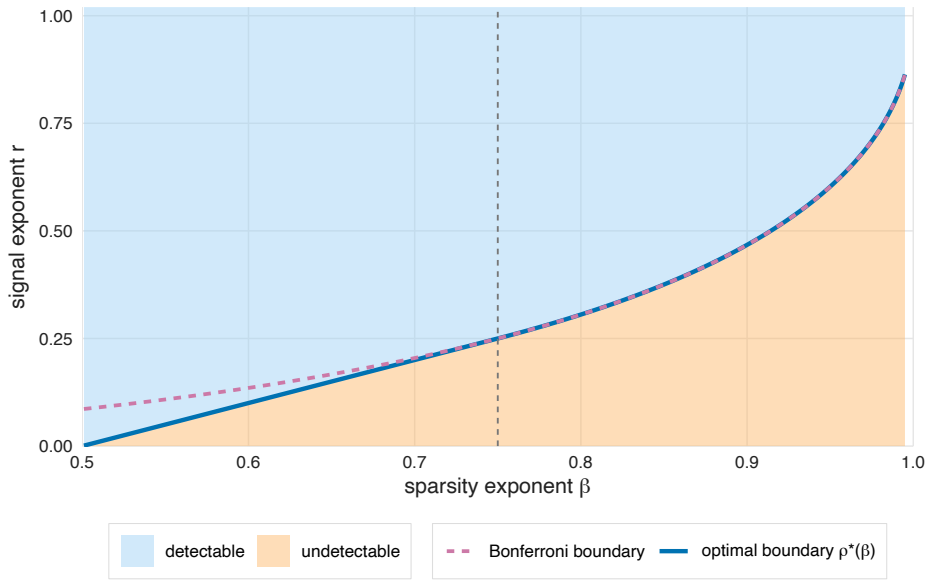

FIGURE 4. Rare/weak detection boundary for Gaussian mixtures. Above $\rho^*(\beta)$, detection is possible; below it, detection is impossible. Bonferroni is rate-optimal in the very sparse region $3/4 \le \beta < 1$, but not throughout the whole rare/weak regime.

The shape of the boundary explains the roles of the tests. When $\beta \ge 3/4$, the alternatives are so sparse that the best evidence is often an extreme coordinate. Bonferroni, or equivalently a maximum test, reaches the optimal boundary because the maximum among the signal coordinates can exceed the maximum-noise scale. The signal count is approximately $m\varepsilon_m = m^{1-\beta}$, each signal mean is $\sqrt{2r\log m}$, and the maximum of $m^{1-\beta}$ independent $N(0,1)$ noise terms concentrates near $\sqrt{2(1-\beta)\log m}$ by the same Mills-ratio calculation used in Chapter 2. Therefore

$$\max_{j\in\text{nonnull}} Z_j = \{\sqrt{2r\log m} + \sqrt{2(1-\beta)\log m}\}\{1+o_P(1)\},$$

which exceeds the Bonferroni threshold $\sqrt{2\log m}\{1+o(1)\}$ exactly when

$$\sqrt{r} + \sqrt{1-\beta} > 1, \qquad \text{equivalently} \qquad r > (1-\sqrt{1-\beta})^2.$$

When $1/2 < \beta < 3/4$, the signals are still rare but not rare enough for the minimum p-value to be the whole story. The optimal boundary is $\beta - 1/2$, which lies below the Bonferroni boundary. Higher criticism and Berk-Jones recover this region by scanning over many tail cutoffs.

Here is the exponent calculation behind that statement. Fix a p-value cutoff $t = m^{-q}$, $0 < q < 1$. Under the null, the expected count below t is

$$mt = m^{1-q}, \qquad \text{sd}\{\#(p_i \le t)\} \asymp m^{(1-q)/2}.$$

For a nonnull observation $Z \sim N(\sqrt{2r\log m}, 1)$,

$$\mathbb{P}(p \le m^{-q}) = \mathbb{P}\{Z \ge \Phi^{-1}(1-m^{-q})\} \approx \mathbb{P}\{N(0,1) \ge \sqrt{2q\log m} - \sqrt{2r\log m}\}.$$

The approximation uses the normal tail quantile $\Phi^{-1}(1-m^{-q}) = \sqrt{2q\log m}\{1+o(1)\}$. If $q \le r$, the threshold lies below the signal mean at the exponent scale, so a nonnull coordinate crosses

with probability tending to one. If $q > r$, Mills' ratio gives

$$\mathbb{P}\{N(0,1) \geq (\sqrt{2q} - \sqrt{2r})\sqrt{\log m}\} = m^{-(\sqrt{q}-\sqrt{r})^2 + o(1)}.$$

Thus the nonnull crossing probability is approximately

$$\begin{cases} 1, & q \leq r, \\ m^{-(\sqrt{q}-\sqrt{r})^2}, & q > r, \end{cases}$$

up to polynomial factors in $\log m$. The excess signal count has exponent

$$1 - \beta - (\sqrt{q} - \sqrt{r})_+^2,$$

whereas the null standard deviation has exponent $(1-q)/2$. At threshold m^{-q}, a standardized tail count such as higher criticism is powerful when

$$1 - \beta - (\sqrt{q} - \sqrt{r})_+^2 > \frac{1-q}{2}.$$

Scanning over q lets the procedure find the cutoff where this exponent gap is largest. To see the optimization, write the left minus right side as

$$G(q; r, \beta) = 1 - \beta - (\sqrt{q} - \sqrt{r})_+^2 - \frac{1-q}{2}.$$

For $q > r$,

$$G(q; r, \beta) = \frac{1}{2} - \beta - r - \frac{q}{2} + 2\sqrt{qr}.$$

The unconstrained maximizer is $q = 4r$. If $4r < 1$, the best exponent is $G(4r; r, \beta) = r - \beta + 1/2$, giving the condition $r > \beta - 1/2$. This is the relevant case when $1/2 < \beta < 3/4$. If the optimizer would lie beyond the search range, the best cutoff is at the endpoint $q = 1$, giving

$$G(1; r, \beta) = 1 - \beta - (1 - \sqrt{r})^2,$$

which is positive exactly when $r > (1 - \sqrt{1-\beta})^2$. This gives the second branch for $3/4 \leq \beta < 1$. The full theorem in Donoho and Jin [38] controls the stochastic fluctuations uniformly over the scanned thresholds; the calculation here is the deterministic signal-to-noise comparison that explains the shape of the phase diagram.

THEOREM 3.3 (Adaptive rare/weak detection, informal). *In the Gaussian mixture model above, higher criticism has asymptotic power tending to one whenever $r > \rho^*(\beta)$, without knowing β or r in advance. If $r < \rho^*(\beta)$, no sequence of level-α tests has power tending to one uniformly over the corresponding rare/weak alternatives.*

Why the boundary has this form. For a fixed threshold $t = m^{-q}$, the null count $mF_m(t)$ has mean approximately m^{1-q} and standard deviation approximately $m^{(1-q)/2}$. Under the mixture, the nonnull coordinates add an excess count whose exponent depends on β, r, and q. HC standardizes the excess by the null standard deviation and then maximizes over q. The best cutoff yields a positive exponent exactly when $r > \rho^*(\beta)$. The impossibility statement is proved by likelihood-ratio or second-moment arguments for the mixture distribution under the null. The full proof is technical; the exponent calculation is the part that explains why a tail scan is adaptive.

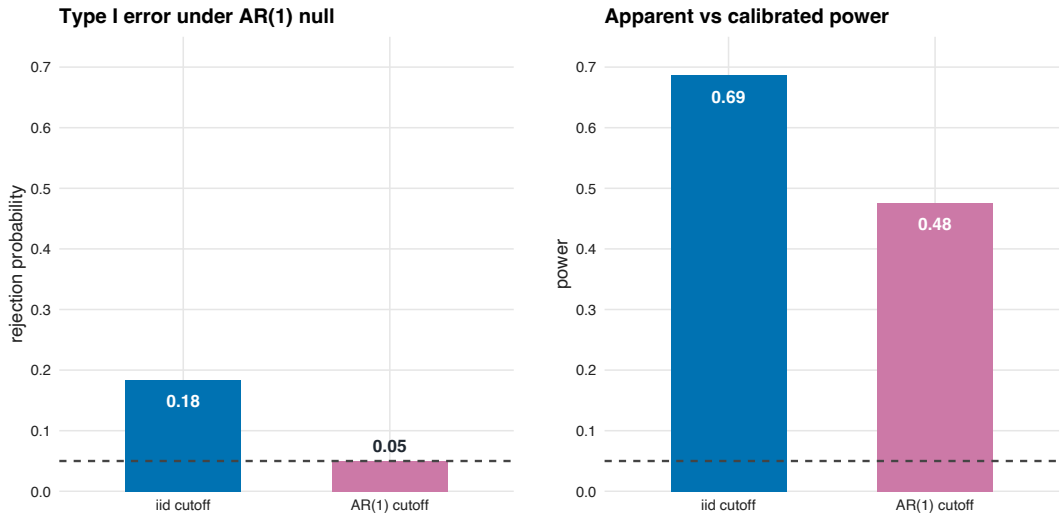

FIGURE 5. Dependence changes both calibration and apparent power. The simulation uses $m = 400$ one-sided Gaussian p-values with AR(1) null correlation $\rho = 0.85$. Cutoffs are Monte Carlo 95th-percentile null cutoffs, either under iid uniforms or under the AR(1) null. Under the AR(1) null, the iid cutoff is too low and rejects too often; under a sparse alternative with 8 randomly located signals of size 2.4, the same too-low cutoff also gives inflated apparent power. The AR(1)-calibrated cutoff restores size control and reports the honest lower power.

6. Dependence Can Hurt, but It Can Also Help

The preceding calibration assumes independent p-values. Dependence changes the fluctuations of $F_m(t)$, so an independence-calibrated HC threshold can be wrong. Positive correlation, for example, can make small p-values arrive in clusters even under the global null.

Dependence is not only a nuisance. In structured Gaussian problems it can make signals more visible if the statistic uses the covariance structure. The idea appears clearly in innovated higher criticism. If $X = \mu + Z$ with $Z \sim N(0, \Sigma)$, then an innovation transformation U with $U\Sigma U^\top = I$ turns the noise white:

$$UX = U\mu + UZ, \qquad UZ \sim N(0, I).$$

For many correlated models, such as Toeplitz or decaying correlations, the transformation can increase the effective strength of sparse signals. The iid case can therefore be the hardest case at a fixed sparsity and signal strength. The catch is that ordinary HC applied to the untransformed coordinates ignores this information, so it can suffer from both wrong calibration and lost power.

A related likelihood-ratio calculation explains why precision adjustment is a principled statistic for sparse high-dimensional mean testing. Consider n independent observations $X_1, \dots, X_n$ in $\mathbb{R}^m$ with mean μ and known covariance matrix Σ, and let $\Gamma = \Sigma^{-1}$ be the precision matrix. The sample mean $\bar{X} = n^{-1} \sum X_\ell$ has covariance Σ/n, and the precision-adjusted score vector

$$Z = \Gamma \bar{X} \qquad \text{satisfies} \qquad \mathbb{E}Z = \Gamma\mu, \quad \mathrm{Cov}(Z) = \Gamma/n.$$

(Z is the Gaussian score for the mean; its covariance is Γ/n, not a multiple of the identity, so the transformation is a precision adjustment rather than a whitening.) Each coordinate Z_j is the locally-most-powerful test statistic for the partial mean of X_j given all other coordinates.

For a fixed candidate support S, maximizing the Gaussian likelihood over mean vectors supported on S gives the likelihood-ratio statistic

$$T_S = n\, Z_S^\top \Gamma_{S,S}^{-1} Z_S,$$

and T_S has a $\chi^2_{|S|}$ distribution under $\mu = 0$. If the alternative is sparse but the support is unknown, the ideal sparse likelihood-ratio test scans over candidate sets S. For single-coordinate supports this reduces to the maximum of precision-adjusted standardized scores,

$$\max_{1\le j\le m} \frac{nZ_j^2}{\gamma_{jj}},$$

where γ_{jj} is the jth diagonal of Γ. This is the one-sparse case of the likelihood-ratio derivation in Zhang [134] and the basis of the two-sample precision-transformed maximum test of Cai et al. [29].

This likelihood-ratio view also gives the right interpretation of when dependence helps. For sparse or maximum-type tests, what matters after precision adjustment is not the raw mean μ_j, but the leading standardized transformed coordinates

$$\frac{(\Gamma\mu)_j}{\sqrt{\gamma_{jj}}}.$$

Power improves when these leading coordinates are larger than the marginal signals, or when neighboring transformed coordinates carry additional evidence for the same sparse effect. Power can be worse for a maximum-type statistic when the precision transformation spreads the signal across many coordinates or creates cancellations that reduce the largest transformed coordinate. The choice of statistic should therefore match the alternative: max-type and sparse likelihood-ratio scans are aimed at sparse transformed evidence, while quadratic or sum-of-squares tests are better suited to broad dense shifts.

The practical lesson is simple, but conditional on what can be learned from the data. If dependence is weak or well approximated by a known null model, simulate or permute from that model. If dependence is structured, scientifically meaningful, and the covariance or precision structure is known or can be estimated accurately enough, build it into the statistic through precision adjustment or score whitening rather than treating it only as a calibration problem. In high-dimensional problems this is a real modeling requirement, not a free preprocessing step: estimating Γ usually requires assumptions such as sparse precision rows or a sparse graphical model, with methods such as nodewise Lasso, CLIME, or graphical Lasso. When those assumptions are doubtful, dependence-aware null calibration is safer than a poorly estimated precision transformation.

7. Assumptions in Plain Language

Uniform p-values are the baseline for this chapter. Exact Simes calibration uses independent continuous uniforms. Empirical-process approximations use large m, and the usual Brownian-bridge limit is an independence result. Higher criticism and Berk-Jones rely on tail standardization; their finite-sample thresholds should be calibrated under the actual null dependence when possible. The rare/weak boundary is an asymptotic statement for an idealized Gaussian mixture. It is valuable because it explains which signal shapes different tests are designed to see.

8. Bibliographic Notes

The Simes global test is due to Simes [109]. Higher criticism and the rare/weak detection boundary are presented in the form of Donoho and Jin [38]. The earlier minimax testing work of Ingster [64] is important background: it shows that sparse Gaussian testing can have

likelihood-ratio limits outside the classical Gaussian regime, which is why tail-sensitive tests are not merely finite-sample heuristics. The $\sqrt{2\log\log m}$ null scale for higher criticism follows from law-of-the-iterated-logarithm arguments going back to Darling and Erdős [34]. Berk-Jones statistics go back to Berk and Jones [20]; modern comparisons of higher criticism, Berk-Jones, and related phi-divergence statistics include Jager and Wellner [67]. Innovated higher criticism and the possibility that dependence can improve sparse-signal detection are developed by Hall and Jin [56]; related heterogeneous-mixture boundaries appear in Cai et al. [27]. Precision-transformed maximum and sparse likelihood-ratio tests for high-dimensional means are developed by Cai et al. [29] and Zhang [134]. Precision-matrix estimation methods used in such settings include nodewise Lasso [91], CLIME [28], and graphical Lasso [50].

9. Exercises

Basic.

Exercise 3.4 (Simes crossing). For $m = 5$ p-values

$$0.003,\ 0.021,\ 0.040,\ 0.31,\ 0.78,$$

compute p_{Simes} and decide whether the Simes test rejects at $\alpha = 0.05$. Compare with Bonferroni.

Exercise 3.5 (Variance of the empirical CDF). Assume the p-values are independent uniforms. Derive

$$\operatorname{Var}\{F_m(t)\} = \frac{t(1-t)}{m}.$$

Use this calculation to motivate the denominator in higher criticism.

Exercise 3.6 (Brownian bridge covariance). Derive the covariance function

$$\operatorname{Cov}\{B(s), B(t)\} = \min(s,t) - st$$

for the limiting empirical process.

Intermediate.

Exercise 3.7 (Simes via spacings). The proof of Theorem 3.1 used induction and conditional order statistics. Give a second proof using spacings. Let

$$U_0 = 0, \qquad U_i = p_{(i)} - p_{(i-1)}, \quad 1 \le i \le m, \qquad U_{m+1} = 1 - p_{(m)}.$$

Use the Dirichlet$(1, \dots, 1)$ distribution of $(U_1, \dots, U_{m+1})$ to express the no-crossing event

$$p_{(i)} > \frac{\alpha i}{m}, \qquad i = 1, \dots, m,$$

as a simplex-volume problem, and verify that its probability is $1 - \alpha$.

Exercise 3.8 (Tail weighting). Compare the integrands used by Cramer-von Mises and Anderson-Darling. Explain why Anderson-Darling is more sensitive to left-tail alternatives in p-value testing.

Exercise 3.9 (Anderson-Darling ordered formula). Derive the ordered-p-value expression for A_m by integrating $\{F_m(t) - t\}^2/\{t(1-t)\}$ over the intervals $[0, p_{(1)}]$, $[p_{(i)}, p_{(i+1)}]$, and $[p_{(m)}, 1]$.

Exercise 3.10 (Berk-Jones likelihood ratio). Fix a threshold t and let $N(t) = mF_m(t)$. Under the null, $N(t) \sim \text{Bin}(m, p)$ with $p = t$. Writing $\widehat{p} = N(t)/m$ for the empirical fraction, show that the binomial log likelihood ratio for testing the null $p = t$ against the data-driven $p = \widehat{p} > t$ is

$$m\left\{\widehat{p}\log\frac{\widehat{p}}{t} + (1-\widehat{p})\log\frac{1-\widehat{p}}{1-t}\right\} = m\,h(\widehat{p}, t).$$

Conclude that the Berk-Jones statistic is, at each rank, a binomial log-likelihood ratio against the null success probability $t = p_{(i)}$.

EXERCISE 3.11 (Simulation comparison). Simulate one-sided Gaussian p-values under a sparse mixture. Compare Kolmogorov-Smirnov, Anderson-Darling, higher criticism, and Berk-Jones using Monte Carlo null calibration. Report both size and power.

Computational.

EXERCISE 3.12 (Average likelihood ratio). Implement the average likelihood ratio statistic with weights $w_i \propto \{i \log m\}^{-1}$ for $1 \le i \le m/2$. Compare its power with HC and BJ in rare/weak simulations.

EXERCISE 3.13 (Rare/weak phase diagram). For a grid of (β, r), simulate the Gaussian mixture model with $\varepsilon_m = m^{-\beta}$ and $\mu_m = \sqrt{2r \log m}$. Estimate the power of Bonferroni, HC, and BJ. Overlay the theoretical boundary $\rho^*(\beta)$ and summarize where each method is strongest.

EXERCISE 3.14 (Dependence calibration of HC). Reproduce Figure 5 for an AR(1) Gaussian null with $\rho = 0.5$ and $m \in \{200, 1000\}$. Compute the iid-calibrated and AR(1)-calibrated cutoffs by Monte Carlo and report the realized Type I error of each under the AR(1) null.

Advanced.

EXERCISE 3.15 (Bonferroni in the very sparse regime). Show that the maximum test detects the rare/weak mixture when

$$\sqrt{r} + \sqrt{1 - \beta} > 1.$$

Use this to explain why Bonferroni is rate-optimal for $3/4 \le \beta < 1$.

EXERCISE 3.16 (Dependence-aware sparse testing). In a Gaussian mean model with known covariance Σ, compare ordinary maximum testing with the precision-adjusted statistic

$$\max_{1 \le j \le m} \frac{n(\Gamma \bar{X})_j^2}{\gamma_{jj}}, \qquad \Gamma = \Sigma^{-1}.$$

Construct an example where the precision-adjusted statistic has larger power, and one where it has smaller power. What property of Γ relative to the true μ determines which regime applies?

EXERCISE 3.17 (Simes uniformity by induction). Refine the $m = 2$ calculation in the proof of Theorem 3.1 to a general inductive argument. Show that under iid uniform p-values, the Simes statistic $T_m = \min_i m p_{(i)}/i$ is exactly $\mathrm{Unif}(0, 1)$, not merely super-uniform.

CHAPTER 4

Familywise Error Rate, Closure, and Graphical Procedures

The first two chapters asked whether a collection of tests contains any signal. That is a global question. In practice, the analyst usually wants individual decisions: which endpoints, genes, models, subgroups, or validation checks can be declared significant? Once individual decisions are reported, the relevant error is no longer only the Type I error of a global test. The analyst must control the chance of making at least one false rejection among the reported discoveries.

This chapter develops familywise error rate control as an error-budgeting problem. Bonferroni spends the budget evenly. Holm spends it sequentially. Closure explains why these procedures are strongly valid. Graphical procedures then give a practical language for more structured testing plans, where alpha can flow from one hypothesis to another after rejections.

The central issue is not merely that many p-values are present. It is that the reported claims are usually read together. A paper that declares three biomarkers significant, a trial report that claims benefit on a primary and two secondary endpoints, or a model audit that flags several failure modes is making a family of statements. Familywise error control asks for a guarantee on that family as a whole: with probability at least $1 - \alpha$, every reported rejection of a true null should be absent. This is a stricter goal than controlling the expected proportion of mistakes, and it is often the right goal when a single false claim can trigger a costly follow-up experiment, regulatory conclusion, or scientific headline.

A second theme is that FWER procedures are easiest to understand when alpha is treated as a scarce resource. Bonferroni divides the resource before seeing the data. Holm recovers some efficiency by allowing earlier, smaller p-values to clear the way for larger thresholds later. Closure gives the theorem that makes such shortcuts legitimate. Graphical procedures make the same resource accounting explicit, which is why they are useful in confirmatory studies whose testing order is part of the design rather than an afterthought.

1. Why Strong Control Is Needed

Consider a platform trial with one primary endpoint and several secondary endpoints. If the primary endpoint succeeds, the study team may want to test secondary endpoints. If one secondary endpoint succeeds, the team may want to test a related subgroup. This is not an unstructured screen: the order and priorities are part of the scientific design. Still, every reported rejection can be wrong, and the sponsor or reader wants a guarantee on the whole family of claims.

Let m hypotheses be tested. For each hypothesis, there are two binary states: the null is true or false, and the procedure rejects or does not reject. The usual decision table is shown in Figure 1.

Write m_0 for the number of true null hypotheses and $m_1 = m - m_0$ for the number of nonnulls. The decision table partitions the m hypotheses into four cells according to whether the null is true or false and whether the procedure rejects or not:

$$U = \#\{i \in \mathcal{I}_0 : \text{not rejected}\}, \qquad V = \#\{i \in \mathcal{I}_0 : \text{rejected}\},$$
$$T = \#\{i \notin \mathcal{I}_0 : \text{not rejected}\}, \qquad S = \#\{i \notin \mathcal{I}_0 : \text{rejected}\},$$

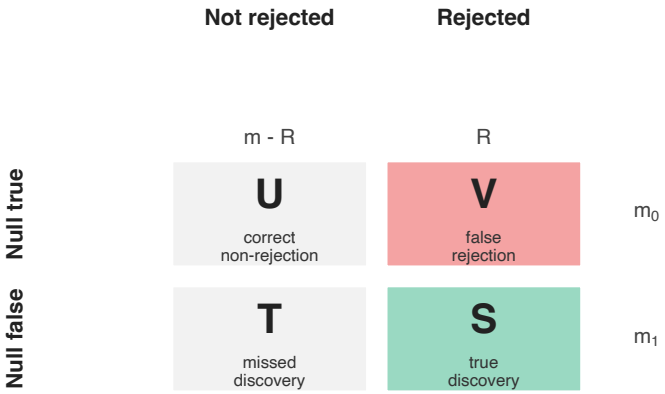

FIGURE 1. Multiple-testing decision table. Rows indicate whether the null hypothesis is true or false; columns indicate the procedure's decision. Among true nulls, U are correctly not rejected and V are falsely rejected; among false nulls, T are missed and S are correctly rejected. Row totals are m_0 and $m_1 = m - m_0$; the column total of rejections is $R = V + S$.

with marginals $U + V = m_0$, $T + S = m_1$, and $R = V + S$ the total number of rejections. In FWER analysis, V is the only random variable we directly bound, but the full table is the conceptual object that organizes every multiple-testing error rate. The familywise error rate is

$$\text{FWER} = \mathbb{P}(V \geq 1),$$

and a procedure controls FWER at level α if $\mathbb{P}(V \geq 1) \leq \alpha$ for the relevant class of data-generating distributions.

DEFINITION 4.1 (Weak and strong FWER control). A procedure has weak FWER control at level α if

$$\mathbb{P}(V \geq 1) \leq \alpha$$

under the global null, when all individual null hypotheses are true. It has strong FWER control at level α if the same guarantee holds for every configuration of true and false null hypotheses.

Strong control is the useful guarantee for individual discoveries. It says that even if some alternatives are real, the probability of falsely rejecting any true null remains at most α.

The distinction is not pedantic. Suppose a procedure first tests the global null at level α, and, if the global null is rejected, reports every individual p-value below 0.05 without adjustment. Under the global null, this has weak control if the global test is valid. But if one nonnull hypothesis makes the global test reject almost surely, then the true nulls are effectively tested at unadjusted level 0.05. With m_0 true nulls and independent p-values, the chance of at least one false rejection is

$$1 - 0.95^{m_0},$$

which can be far above α.

One can also control generalized errors such as k-FWER,

$$\mathbb{P}(V \geq k),$$

but this chapter focuses on the classical $k = 1$ case.

2. Bonferroni, Sidak, and Holm

Let $p_1, \dots, p_m$ be valid p-values for hypotheses $H_1, \dots, H_m$. Let $\mathcal{I}_0$ denote the unknown set of true null indices.

DEFINITION 4.2 (Super-uniform p-value). A p-value p_i is valid, or super-uniform, for H_i if, whenever H_i is true,

$$\mathbb{P}(p_i \le t) \le t \qquad \text{for every } 0 \le t \le 1.$$

If equality holds for every t, the p-value is exactly uniform. Most FWER arguments in this chapter require only marginal super-uniformity for each true-null p-value; dependence assumptions enter only for procedures, such as Sidak, Hochberg, and Hommel, that use sharper information than a union bound.

The Bonferroni procedure rejects

$$H_i \quad \text{if} \quad p_i \le \frac{\alpha}{m}.$$

PROPOSITION 4.3 (Bonferroni strong FWER control). *If the p-values for true nulls are super-uniform, then Bonferroni controls FWER at level α under arbitrary dependence.*

PROOF. If any false rejection occurs, then $p_i \le \alpha/m$ for at least one $i \in \mathcal{I}_0$. Therefore

$$\mathbb{P}(V \ge 1) \le \sum_{i \in \mathcal{I}_0} \mathbb{P}(p_i \le \alpha/m) \le |\mathcal{I}_0| \frac{\alpha}{m} \le \alpha.$$

□

Under independence, Sidak replaces the threshold by

$$1 - (1 - \alpha)^{1/m}.$$

This is exact for m independent true null p-values, but the exactness is not available under arbitrary dependence. Sidak is slightly less conservative than Bonferroni, but Holm's step-down procedure usually gives a more important gain.

Order the p-values as

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)},$$

with corresponding hypotheses $H_{(1)}, \dots, H_{(m)}$. Holm's procedure compares them with

$$\frac{\alpha}{m}, \quad \frac{\alpha}{m-1}, \quad \dots, \quad \alpha.$$

It starts at $p_{(1)}$. If

$$p_{(i)} \le \frac{\alpha}{m - i + 1},$$

it continues to the next rank. It stops at the first failure and rejects all hypotheses before that failure.

THEOREM 4.4 (Holm strong FWER control). *Holm's procedure controls FWER at level α under arbitrary dependence whenever the true-null p-values are super-uniform.*

PROOF. Let $m_0 = |\mathcal{I}_0|$. Fix the ordering used by Holm, including any tie-breaking, and let i^* be the first rank occupied by a true null:

$$i^* = \min\{i : H_{(i)} \text{ is true}\}.$$

Then $i^* \le m - m_0 + 1$, because only the $m - m_0$ false nulls can appear before the first true null. If Holm makes any false rejection, its step-down path must reach and reject this first true null:

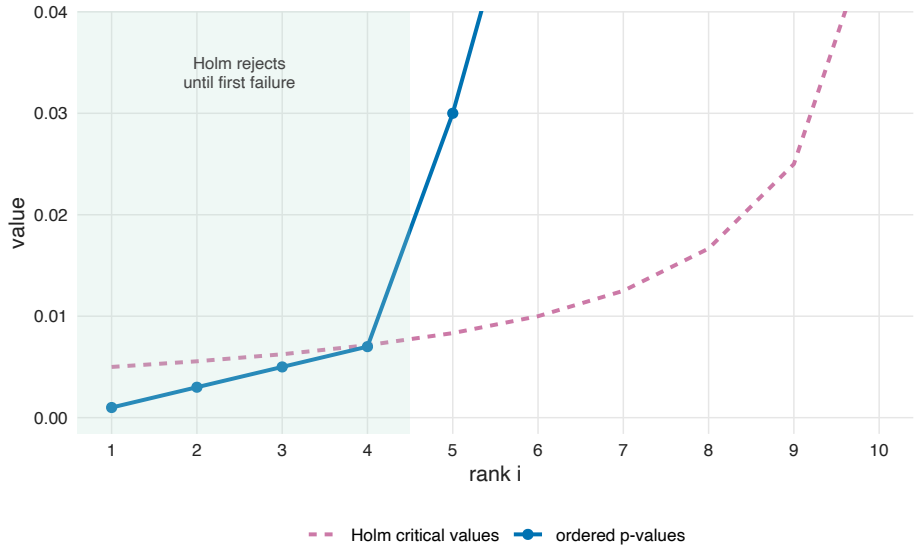

FIGURE 2. Holm's step-down rule. Ordered p-values are compared with the increasing critical values $\alpha/(m-i+1)$. The procedure rejects until the first failure.

all earlier rejected hypotheses may be false nulls, but the first false rejection can only occur when the scan first reaches a true null. Hence

$$p_{(i^*)} \le \frac{\alpha}{m-i^*+1}.$$

The rank bound gives

$$i^* \le m - m_0 + 1, \qquad \frac{\alpha}{m-i^*+1} \le \frac{\alpha}{m_0}.$$

A false rejection therefore implies

$$p_{(i^*)} \le \frac{\alpha}{m_0}.$$

Since $p_{(i^*)}$ is one of the true-null p-values, the last event is contained in

$$\left\{ \min_{j \in \mathcal{I}_0} p_j \le \frac{\alpha}{m_0} \right\}.$$

By Bonferroni applied to the m_0 true-null p-values, this event has probability at most $m_0 \cdot (\alpha/m_0) = \alpha$. □

3. Step-Up Procedures and Simes

Holm is a step-down procedure: it starts with the smallest p-value and moves toward larger ones until it fails. Hochberg's procedure uses the same critical values but scans in the opposite direction. It finds the largest i such that

$$p_{(i)} \le \frac{\alpha}{m-i+1},$$

and rejects $H_{(1)}, \dots, H_{(i)}$.

Hochberg can reject more hypotheses than Holm, but it is not an arbitrary-dependence procedure. Its validity relies on a Simes-type condition. This is the first place where Chapter

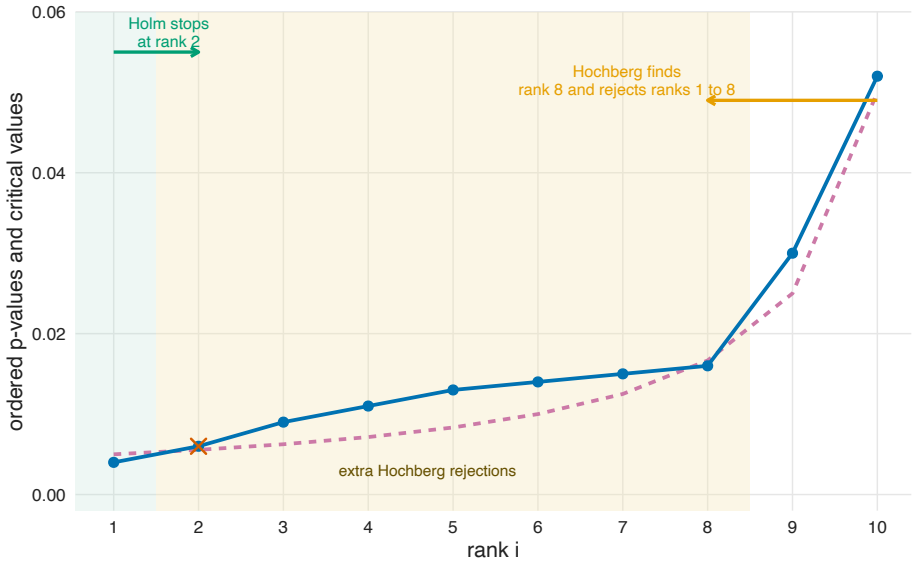

FIGURE 3. Step-down and step-up scans using the same critical curve. Holm starts from the smallest p-value and stops at the first failure, here at rank 2. Hochberg starts from the largest p-value and looks backward for the first success, here at rank 8, so it rejects ranks 1 through 8.

3 enters the multiple-testing story. Recall that the Simes global test rejects $\bigcap_{i=1}^{m} H_i$ at level α when

$$p_{(i)} \le \frac{i\alpha}{m} \quad \text{for at least one } i.$$

It is valid for independent p-values and for several forms of positive dependence, but not under arbitrary dependence. We return to the exact closed-testing interpretation after introducing the closure principle. The important point for now is that Hochberg is a step-up rule whose validity comes from Simes-type assumptions, unlike Holm's arbitrary-dependence guarantee.

There is also a useful warning. The Benjamini–Hochberg procedure, abbreviated BH, is the step-up FDR procedure that rejects $H_{(1)}, \dots, H_{(k)}$, where k is the largest rank satisfying

$$p_{(i)} \le \frac{i\alpha}{m} \quad \text{at } i = k.$$

It is studied in detail in Chapter 5. Under the global null, BH at level α rejects at least one hypothesis exactly when the Simes global test rejects. Under dependence conditions where the Simes inequality holds (independent or PRDS p-values; see Chapter 6), this observation gives *weak* FWER control: under the global null, the probability that BH rejects any hypothesis is at most α. Weak control is not strong control: when some nulls are false, BH may reject true nulls with probability larger than α. The dependence assumption is essential: under arbitrary dependence, Simes itself can fail, so this weak guarantee is not free.

4. The Closure Principle

Closure is the master theorem behind many FWER procedures. For every nonempty set $I \subseteq \{1, \dots, m\}$, define the intersection hypothesis

$$H_I = \bigcap_{i \in I} H_i.$$

A closed testing procedure specifies a level-α local test for every intersection H_I. It rejects an elementary hypothesis H_i if and only if every intersection hypothesis H_I containing i is rejected by its local test.

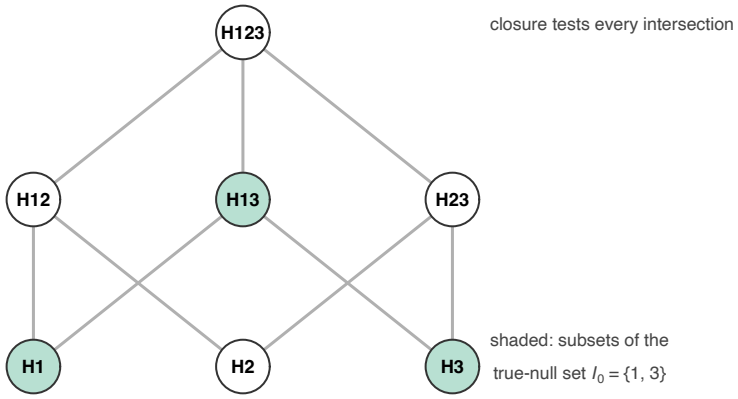

FIGURE 4. Closure lattice for three hypotheses. To reject H_1, closed testing requires rejection of every intersection containing 1: $H_1, H_{12}, H_{13}, H_{123}$. In the example shading, the true-null set is $\mathcal{I}_0 = \{1, 3\}$, so the shaded nodes are exactly the intersection hypotheses H_I with $I \subseteq \mathcal{I}_0$; the rejection probability of any of them under the data-generating distribution is at most α, and this is the lever that controls FWER.

THEOREM 4.5 (Closure principle). *If every intersection hypothesis H_I is tested by a level-α local test, then the closed testing procedure controls FWER at level α strongly.*

PROOF. Let $\mathcal{I}_0$ be the set of true null hypotheses, and note that $\mathcal{I}_0$ is among the intersection indices considered by the closed procedure (it is nonempty whenever at least one null is true; the case $\mathcal{I}_0 = \emptyset$ gives $V = 0$ deterministically). If the closed procedure makes any false rejection, then it rejects some H_i with $i \in \mathcal{I}_0$. By the rule of closed testing, it must then reject every intersection containing i, in particular the true intersection

$$H_{\mathcal{I}_0} = \bigcap_{i \in \mathcal{I}_0} H_i.$$

Under the true data-generating distribution, $H_{\mathcal{I}_0}$ is true, so its local test has Type I error at most α. Therefore

$$\mathbb{P}(V \geq 1) \leq \mathbb{P}(\text{reject } H_{\mathcal{I}_0}) \leq \alpha.$$

□

The proof is short because closure reduces strong control to one local test: the intersection of all true nulls. The cost is computational and conceptual. There are $2^m - 1$ intersections, so useful procedures are those where the closed rule has a shortcut.

Closure is not limited to classical FWER. A modern line of work uses the same idea to produce simultaneous lower bounds on the number of true discoveries, and hence upper bounds on the false discovery proportion, for post hoc selected sets of rejections [138, 139]. Recent extensions replace p-value local tests by e-value local tests, including online closed testing and a

general closure principle for FDR control [137, 140]. This chapter keeps the focus on FWER, but these extensions explain why closure remains a central organizing principle beyond familywise error.

5. Closed Bonferroni, Closed Simes, and Hommel

Closed Bonferroni tests each intersection H_I by Bonferroni within that intersection:

$$\min_{i \in I} p_i \le \frac{\alpha}{|I|}.$$

The closed procedure is exactly Holm.

PROPOSITION 4.6 (Closed Bonferroni equals Holm). *The closed testing procedure obtained by applying Bonferroni to every intersection hypothesis rejects exactly the same elementary hypotheses as Holm's step-down procedure.*

PROOF. We verify both inclusions.

Every Holm rejection is a closed-Bonferroni rejection. Suppose Holm rejects $H_{(1)}, \dots, H_{(k)}$, and fix an index (j) with $j \le k$. For any intersection I containing (j), let

$$h = \min\{r : H_{(r)} \in I\},$$

the smallest rank index appearing in I. Then $h \le j \le k$, so Holm rejected $H_{(h)}$, which means

$$p_{(h)} \le \frac{\alpha}{m - h + 1}.$$

Moreover $I \subseteq \{(h), (h+1), \dots, (m)\}$, so $|I| \le m - h + 1$. Combining these,

$$p_{(h)} \le \frac{\alpha}{m - h + 1} \le \frac{\alpha}{|I|},$$

and the Bonferroni local test for H_I rejects. Therefore closed Bonferroni rejects every $H_{(j)}$ for $j \le k$.

No Holm nonrejection is recovered by closure. Suppose Holm stops at rank k, meaning $p_{(k)} > \alpha/(m - k + 1)$. Consider the intersection

$$I^* = \{(k), (k+1), \dots, (m)\}, \qquad |I^*| = m - k + 1.$$

Every p-value in I^* satisfies $p_{(j)} \ge p_{(k)} > \alpha/(m-k+1) = \alpha/|I^*|$, so the Bonferroni local test for H_{I^*} does not reject. Since I^* contains (k), the closed procedure cannot reject $H_{(k)}$ either. The same nonrejected intersection I^* contains every $H_{(j)}$ with $j \ge k$, so none of the hypotheses after Holm's stopping rank can be rejected by closure. Thus closure adds no extra rejections beyond Holm, but it explains why Holm is strongly valid. □

Closed Simes uses the Simes test as the local test for every intersection:

$$\min_{1 \le j \le |I|} \frac{|I| p_{(j:I)}}{j} \le \alpha,$$

where $p_{(1:I)} \le \cdots \le p_{(|I|:I)}$ are the p-values inside intersection I. The resulting closed procedure is Hommel's procedure. Thus Hommel, not Hochberg, is the exact shortcut for the closure of Simes local tests. Hochberg is a simpler and more conservative step-up shortcut: it rejects only hypotheses that Hommel also rejects.

PROPOSITION 4.7 (Hochberg is contained in closed Simes). *Suppose the Simes local test is valid for every intersection hypothesis. Then every rejection made by Hochberg is also made by the closed Simes procedure that is, by Hommel's procedure. Consequently Hochberg controls FWER at level α under the same Simes-validity assumptions.*

PROOF. Let $p_{(1)} \le \cdots \le p_{(m)}$. Suppose Hochberg rejects $H_{(j)}$. Then there exists $r \ge j$ such that

$$p_{(r)} \le \frac{\alpha}{m-r+1}.$$

To show that closed Simes rejects $H_{(j)}$, fix any intersection I containing (j), and write $h = |I|$. We must show that the Simes local test rejects H_I.

Among all intersections of size h that contain (j), the hardest one to reject is the one that includes $H_{(j)}$ and the largest possible p-values. Call this set K_h. If $j \le m-h+1$, take

$$K_h = \{(j), (m), (m-1), \ldots, (m-h+2)\};$$

otherwise take $K_h = \{(m), (m-1), \ldots, (m-h+1)\}$. The ordered p-values inside any other I are coordinatewise no larger than those inside K_h, so rejection of K_h implies rejection of I.

If $r \le m-h+1$, then $p_{(j)} \le p_{(r)} \le \alpha/(m-r+1) \le \alpha/h$, so the first Simes comparison rejects K_h. If $r > m-h+1$, then K_h contains $p_{(r)}$ and all $m-r$ larger ordered p-values. Thus

$$p_{(h-m+r:K_h)} \le p_{(r)}.$$

Let $a = m-r+1$ and $b = h-m+r$. Then $a+b-1 = h$, so

$$ab - h = (a-1)(b-1) \ge 0.$$

Consequently

$$p_{(h-m+r:K_h)} \le \frac{\alpha}{m-r+1} \le \frac{(h-m+r)\alpha}{h},$$

which is exactly a Simes rejection for K_h. Therefore every intersection containing (j) is rejected by Simes, so closed Simes rejects $H_{(j)}$. Since closed Simes controls FWER by Theorem 4.5, any procedure whose rejections are a subset of closed Simes rejections also controls FWER. □

Because Simes is never less powerful than Bonferroni on the same intersection, closed Simes rejects at least everything closed Bonferroni rejects. Therefore, under the dependence conditions where Simes local tests are valid, Hommel is at least as powerful as Hochberg, and Hochberg is at least as powerful as Holm. The price is that Hommel's exact shortcut is less transparent than Holm's, and Hochberg trades away some of Hommel's power for a simpler step-up rule.

The hierarchy is therefore:

$$\text{Holm} \quad \subseteq \quad \text{Hochberg} \quad \subseteq \quad \text{Hommel},$$

with the important caveat that the latter two require Simes-type validity, not just arbitrary super-uniform p-values.

6. Graphical Error Spending

Closure is general, but it is not always the most convenient design language. In structured confirmatory studies, analysts often know the priority order: primary endpoint first, then secondary endpoints, then subgroups or supportive analyses. Graphical procedures encode this structure directly.

Each hypothesis H_i receives an initial local alpha budget α_i, with

$$\sum_{i=1}^{m} \alpha_i \le \alpha.$$

Directed edges carry transition weights $g_{ij} \ge 0$, where $g_{ii} = 0$ and

$$\sum_{j=1}^{m} g_{ij} \le 1.$$

If H_i is rejected, its current alpha budget is transferred to other hypotheses according to the outgoing weights.

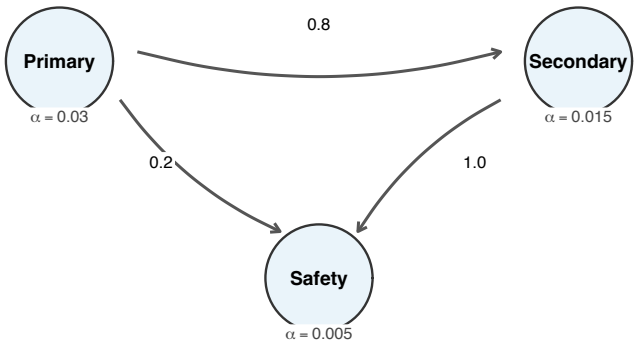

FIGURE 5. A graphical testing procedure. Initial alpha is assigned to nodes. After a rejection, the rejected node's local alpha is redistributed along outgoing edges.

The general update after rejecting H_i is:

$$\alpha_j \leftarrow \alpha_j + \alpha_i g_{ij}, \qquad j \neq i.$$

The rejected node is removed. The remaining transition weights are updated by

$$g_{jk} \leftarrow \frac{g_{jk} + g_{ji}g_{ik}}{1 - g_{ji}g_{ij}}, \qquad j \neq k, \quad j,k \neq i,$$

and $g_{jj} = 0$. The same graph is doing two jobs. First, the initial local levels define a weighted Bonferroni test for the global intersection:

$$H_{\{1,\ldots,m\}} = \bigcap_{i=1}^{m} H_i \quad \text{is rejected if} \quad p_i \leq \alpha_i \text{ for some } i.$$

Since $\sum_i \alpha_i \leq \alpha$, this local intersection test has level at most α. Second, the directed edges specify the shortcut that is used after an elementary rejection: remove the rejected node, recycle its current local level, and update the remaining graph so that it represents the same closed weighted-Bonferroni logic on the smaller family.

The random-walk view makes this update less mysterious. Normalize the current local levels by the total level, $w_i = \alpha_i/\alpha$, and think of one unit of error budget as a walker that starts at node i with probability w_i; any initially unused alpha starts outside the testing family. At a given stage, the hypotheses still under consideration are stopping states: alpha that reaches one of them stays there and becomes part of its current local level. Rejected hypotheses are transient states: alpha that reaches a rejected node follows the outgoing transition weights until it is absorbed by a remaining hypothesis. If an outgoing row sums to less than one, the missing probability can be represented by adding a lost-alpha state H_{m+1} with

$$g_{j,m+1} = 1 - \sum_{k=1}^{m} g_{jk}, \qquad g_{m+1,k} = 0 \quad (k \leq m), \qquad g_{m+1,m+1} = 1.$$

On this expanded graph, define a Markov chain by

$$\mathbb{P}(Z_0 = H_i) = \frac{\alpha_i}{\alpha}, \qquad i = 1, \dots, m,$$
$$\mathbb{P}(Z_0 = H_{m+1}) = 1 - \sum_{i=1}^{m} \frac{\alpha_i}{\alpha},$$
$$\mathbb{P}(Z_{t+1} = H_k \mid Z_t = H_j) = g_{jk}, \qquad j, k = 1, \dots, m+1.$$

For a remaining intersection $I \subseteq \{1, \dots, m\}$, let

$$\tau_I = \inf\{t \geq 0 : Z_t \in \{H_i : i \in I\} \cup \{H_{m+1}\}\}.$$

The weight assigned to H_i in the local test for H_I is then the hitting probability

$$w_i(I) = \mathbb{P}(Z_{\tau_I} = H_i), \qquad i \in I.$$

Thus the current local level of H_i is

$$\alpha w_i(I) = \alpha\, \mathbb{P}(Z_{\tau_I} = H_i).$$

The lost-alpha state is absorbing and is not tested, so $\sum_{i \in I} w_i(I) \leq 1$. Strict inequality means that some of the original alpha budget can be absorbed by H_{m+1} before reaching any hypothesis in I. The corresponding closed local test rejects H_I when $p_i \leq \alpha w_i(I)$ for at least one $i \in I$.

The denominator in the update formula is the two-node special case of this hitting-probability calculation. When the rejected node is i, a future unit of alpha leaving j can reach k directly through $j \to k$, or through $j \to i \to k$, or only after one or more round trips $j \to i \to j \to i$. The total rerouted weight is

$$(g_{jk} + g_{ji}g_{ik})\{1 + g_{ji}g_{ij} + (g_{ji}g_{ij})^2 + \cdots\} = \frac{g_{jk} + g_{ji}g_{ik}}{1 - g_{ji}g_{ij}}.$$

The formula requires $g_{ji}g_{ij} < 1$. Since both weights lie in $[0, 1]$, this fails only in the boundary case $g_{ji} = g_{ij} = 1$. That boundary describes two nodes that send all alpha to each other and never let it escape; such a two-cycle should be excluded from the design. This update formula and its validity are derived by Bretz et al. [24]: after each rejection, the updated graph again encodes the appropriate weighted Bonferroni intersection tests, with weights renormalized so that the shortcut agrees with the closed testing procedure. The next section explains the weighted Bonferroni local tests that underwrite this correspondence.

Two common special cases are fixed-sequence and fallback testing.

Example 4.8 (Fixed sequence). Assume the testing plan lists the hypotheses in a priority order

$$H_1,\ H_2,\ \dots,\ H_m,$$

before the data are examined. Fixed sequence puts the entire error budget on the first unrejected item in this list. Thus H_1 is tested at level α; only if H_1 is rejected does H_2 get tested, again at level α; and the same gatekeeping logic continues down the list. A single nonrejection closes the gate for all hypotheses that follow. As a graph, this is the chain with $\alpha_1 = \alpha$, $\alpha_i = 0$ for $i > 1$, $g_{i,i+1} = 1$ for $i < m$, and no other transitions. For instance, with $\alpha = 0.025$ and p-values in this hypothesis order

$$(0.004,\ 0.031,\ 0.006,\ 0.001),$$

only H_1 is rejected. The second p-value exceeds 0.025, so the procedure stops there; the smaller p-values later in the sequence are not eligible for confirmatory claims. The advantage is that every reached test uses the full level α; the disadvantage is the complete dependence on the preassigned order.

EXAMPLE 4.9 (Fallback). Fallback testing also uses an ordered list, but it does not put all of the initial budget on the first hypothesis. Choose nonnegative local levels $\alpha_1, \ldots, \alpha_m$ with $\sum_i \alpha_i \leq \alpha$. Hypothesis H_i starts with its own reserved amount α_i, and if H_i is rejected, its current level is transferred to H_{i+1}. The graph is again a chain, $g_{i,i+1} = 1$, with all other transition weights zero.

For example, take $\alpha = 0.05$, initial levels

$$(0.030,\ 0.015,\ 0.005),$$

and p-values

$$(0.040,\ 0.012,\ 0.018).$$

The first hypothesis is not rejected at level 0.030, so no alpha is passed forward from H_1. The second hypothesis is still tested at its reserved level 0.015, and it rejects. Its 0.015 is then added to the third hypothesis, so H_3 is tested at $0.005 + 0.015 = 0.020$ and also rejects. This is the essential distinction from fixed sequence: an early failure lowers the later testing levels, but it does not necessarily end the entire testing plan.

The graph in Figure 5 gives a concrete three-endpoint calculation. Initially the primary, secondary, and safety hypotheses have local levels

$$(0.030,\ 0.015,\ 0.005).$$

If the primary endpoint rejects, its 0.030 budget is split according to the outgoing weights: 0.8 to secondary and 0.2 to safety. The remaining local levels become

$$\alpha_{\text{secondary}} = 0.015 + 0.8(0.030) = 0.039, \qquad \alpha_{\text{safety}} = 0.005 + 0.2(0.030) = 0.011.$$

The edge weights are updated separately from the local alpha levels. Write P, S, A for primary, secondary, and safety. The nonzero initial transition weights are

$$g_{PS} = 0.8, \qquad g_{PA} = 0.2, \qquad g_{SA} = 1,$$

and the omitted reverse edges have weight zero. After rejecting P, the remaining secondary-to-safety edge is

$$g_{SA}^{\text{new}} = \frac{g_{SA} + g_{SP}g_{PA}}{1 - g_{SP}g_{PS}} = \frac{1 + 0 \cdot 0.2}{1 - 0 \cdot 0.8} = 1.$$

The reverse safety-to-secondary edge remains zero:

$$g_{AS}^{\text{new}} = \frac{g_{AS} + g_{AP}g_{PS}}{1 - g_{AP}g_{PA}} = 0.$$

Thus the updated graph has a single edge from secondary to safety with weight one. If secondary then rejects, safety receives the secondary budget, giving

$$\alpha_{\text{safety}} = 0.011 + 0.039 = 0.050.$$

Thus, for p-values $(0.020, 0.030, 0.018)$, the procedure rejects all three: primary at 0.030, secondary at 0.039, and safety after recycling at 0.050. If the primary endpoint had failed, secondary and safety would have remained at their initial 0.015 and 0.005 levels.

7. Weighted Bonferroni and Consonance

Graphical procedures are not valid because the graph is intuitive. They are valid because they correspond to closed tests based on weighted Bonferroni local tests.

DEFINITION 4.10 (Weighted Bonferroni local test). For an intersection H_I, choose weights $w_i(I) \geq 0$ for $i \in I$ with

$$\sum_{i \in I} w_i(I) \leq 1.$$

The weighted Bonferroni local test rejects H_I if

$$p_i \leq \alpha w_i(I) \quad \text{for at least one } i \in I.$$

By the union bound, this is a level-α test for H_I. Closing these local tests gives strong FWER control by Theorem 4.5.

DEFINITION 4.11 (Monotone weights). The intersection weights are monotone if, whenever $J \subseteq I$ and $i \in J$,

$$w_i(J) \geq w_i(I).$$

Intuitively, removing rejected hypotheses should not reduce the alpha available to the hypotheses that remain.

DEFINITION 4.12 (Consonance). A family of local tests is consonant if, whenever it rejects an intersection H_I, there is some $i \in I$ such that every smaller intersection H_J, $i \in J \subseteq I$, is also rejected. In the resulting closed procedure, rejection of an intersection therefore points to at least one elementary hypothesis that can be rejected. Consonance is what lets a closed procedure behave like a sequential rejection rule rather than stopping with only a rejected global intersection.

The closed weighted Bonferroni algorithm can be read as a sequential shortcut. At any stage, let I be the set of hypotheses not yet removed. Test each remaining H_i, $i \in I$, at local level $\alpha w_i(I)$. If at least one hypothesis crosses its local weighted Bonferroni threshold, reject one such hypothesis, remove it from I, update the weights for the smaller intersection, and continue. If no remaining hypothesis crosses its local weighted Bonferroni threshold, stop. Monotonicity and consonance are the properties that make this shortcut agree with the full closed-testing rule rather than merely resemble it.

Nonconsonance is easy to see with a combining test. Suppose p_1 and p_2 are independent valid p-values under their two null hypotheses. For the intersection $H_{12} = H_1 \cap H_2$, use Fisher's local test,

$$T_{12} = -2\{\log(p_1) + \log(p_2)\}, \qquad \text{reject } H_{12} \text{ if } T_{12} > \chi^2_{4,0.95}.$$

For the singleton intersections, use the ordinary level-0.05 tests $p_i \leq 0.05$. If $p_1 = p_2 = 0.07$, then

$$T_{12} = -4\log(0.07) \approx 10.64,$$

which exceeds $\chi^2_{4,0.95} \approx 9.49$. The intersection therefore rejects at level 0.05. However, both singleton tests fail because $0.07 > 0.05$. The rejected intersection does not point to any elementary hypothesis that can be rejected, which is exactly the nonconsonant behavior.

8. Simulation: Error and Power

The practical tradeoff among FWER procedures is conservative validity versus power. Figure 6 compares Bonferroni, Holm, and Hochberg in a one-sided Gaussian simulation with many near-boundary nonnull effects and strong positive equicorrelation. Holm dominates Bonferroni because it recycles unused alpha through the ordering. Hochberg can gain additional power when its dependence assumptions are appropriate; the third panel isolates this increment over Holm directly. The simulation is deliberately chosen to make the step-up advantage visible. Read the error panel first: any method above the dashed line is invalid for that setting, regardless of its power in the companion panel.

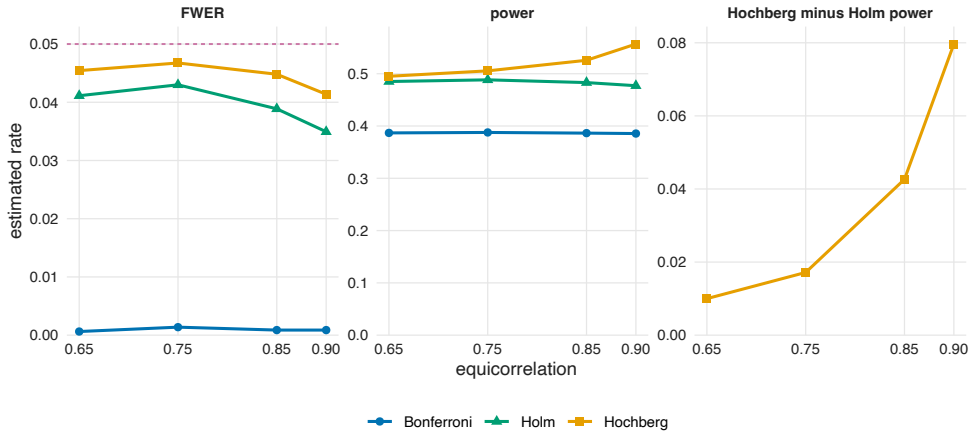

FIGURE 6. Monte Carlo FWER, average power, and Hochberg's power gain over Holm for Bonferroni, Holm, and Hochberg in a positively correlated Gaussian model with ninety-eight nonnull hypotheses among one hundred and equicorrelations from 0.65 to 0.9. The dashed line in the FWER panel marks the nominal level. The right panel isolates the Holm–Hochberg difference, which is most visible in this high-dimensional, highly correlated near-boundary configuration.

9. Assumptions in Plain Language

Bonferroni and Holm need only valid p-values for the true null hypotheses; they do not require independence. Sidak's exact threshold requires independence. Hochberg and Hommel get their extra power from Simes-type local tests, so their validity depends on conditions under which Simes is valid. Graphical procedures are valid when their alpha recycling corresponds to a closed family of weighted Bonferroni local tests. A graph is a design interface, not a proof by itself.

10. Bibliographic Notes

Bonferroni and Sidak are classical. Holm's step-down procedure is due to Holm [60]. The closure principle is due to Marcus et al. [89]. Hochberg's step-up procedure was introduced by Hochberg [59], and the closed Simes shortcut leading to Hommel's procedure is due to Hommel [61]. Simes' test, used inside closed Simes procedures, is from Simes [109]. Modern graphical testing procedures and gatekeeping formulations are described by Bretz et al. [24] and Dmitrienko et al. [37].

Closure has also become a general tool for post hoc error statements beyond FWER. The exploratory-research formulation of Goeman and Solari [138] uses closed testing to give simultaneous true-discovery guarantees for sets chosen after seeing the data. Goeman et al. [139] show that closed testing is essentially unavoidable for admissible control of false-discovery proportion tail probabilities. More recent e-value work extends the same logic to online testing [137] and to broad classes of expected-loss error rates, including FDR [140].

The regulatory side of multiple testing in clinical trials is shaped by a small set of official guidance documents. The ICH E9 statistical-principles guideline [65] and its E9(R1) addendum on estimands [66] set the language for primary and secondary endpoints and the inferential targets they certify. The EMA guideline on multiplicity issues [44] and the FDA guidance on multiple endpoints in clinical trials [124] specify how graphical and gatekeeping procedures should

be planned and reported. The practical face of closure and graphical procedures is heavily shaped by these documents.

11. Exercises

Basic.

EXERCISE 4.13 (Decision table). For a testing problem with $m = 20$, $m_0 = 15$, $R = 6$, and $V = 2$, compute the entries U, T, and S of the decision table in Figure 1, the realized FDP $= V/(R \vee 1)$, and whether a familywise error occurred.

EXERCISE 4.14 (Bonferroni strong control). Prove Bonferroni strong FWER control under arbitrary dependence using only super-uniformity of the true-null p-values.

EXERCISE 4.15 (Weak is not strong). Construct a procedure that controls FWER under the global null but fails to control FWER when one null is false. Compute the FWER in your example.

EXERCISE 4.16 (Weighted Bonferroni). Prove that a weighted Bonferroni local test for an intersection H_I has level at most α when the weights sum to at most one.

EXERCISE 4.17 (Nonconsonance numerically). Verify the Fisher-combination nonconsonance example in the text numerically: compute $T_{12} = -2\{\log(p_1) + \log(p_2)\}$ for $p_1 = p_2 = 0.07$, compare it with $\chi^2_{4,0.95}$, and compute the Fisher combined p-value. Confirm that the intersection test rejects at level 0.05, while neither singleton test $p_i \le 0.05$ rejects.

Intermediate.

EXERCISE 4.18 (Holm by direct proof). Prove Holm's strong FWER control by bounding the rank of the smallest true-null p-value among the rejected hypotheses, mirroring the rank argument sketched in Theorem 4.4. Fill in any step that the chapter compressed.

EXERCISE 4.19 (Closed Bonferroni). Show that closing Bonferroni local tests gives exactly Holm's procedure. Verify both directions: closed Bonferroni rejects every hypothesis Holm rejects, and rejects nothing Holm fails to reject.

EXERCISE 4.20 (Closure lattice). For $m = 4$ hypotheses, draw the full closure lattice (15 nonempty intersections). For p-values

$$(0.004,\ 0.012,\ 0.030,\ 0.200),$$

mark which intersections are rejected by closed Bonferroni at level 0.05, and identify the rejection set of the shortcut Holm procedure. Check that the two rejection sets agree at the elementary-hypothesis level.

EXERCISE 4.21 (Graph update). Consider three hypotheses with initial alpha budgets

$$(0.03,\ 0.015,\ 0.005)$$

and transition matrix

$$G = \begin{pmatrix} 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Trace the graphical procedure by hand for p-values $(0.02, 0.03, 0.018)$: determine which hypotheses are rejected and in what order, applying the graph-update formula after each rejection. In particular, show that after rejecting the first hypothesis, the updated transition weight from the second hypothesis to the third remains one.

EXERCISE 4.22 (Random-walk view of alpha recycling). For the same three-node graph in the previous exercise, take $\alpha = 0.05$ and write P, S, A for the primary, secondary, and safety hypotheses. Normalize the initial local levels to

$$w = (0.6,\ 0.3,\ 0.1).$$

After P is rejected, treat P as transient and S, A as absorbing. Using the hitting-time notation from the text, compute the absorption probabilities at S and A, and verify that the corresponding local levels are 0.039 and 0.011. If S is then rejected, treat both P and S as transient and A as absorbing, and verify that the absorption probability at A is one. Explain why the lost-alpha state has zero hitting probability in this example.

Computational.

EXERCISE 4.23 (Simulation). Reproduce Figure 6. Vary the effect size, correlation, and number of nonnull hypotheses, and summarize when the extra power of Hochberg is most visible. Report both FWER and average power, since a method whose FWER exceeds α is not a valid comparison even if its power is larger.

EXERCISE 4.24 (Platform-trial gatekeeping). Design a graphical procedure for a clinical trial with one primary endpoint, two co-secondary endpoints, and one prespecified subgroup endpoint. Specify the initial alpha budget and the transition weights. Simulate FWER under the global null and under at least two partial-null configurations, and simulate power under one realistic alternative.

Advanced.

EXERCISE 4.25 (Hochberg and Simes). Show that Hochberg's step-up rejections are contained in the closed Simes rejections, and state clearly the dependence condition needed for validity. Identify a small example in which Hochberg rejects a strict superset of what Holm rejects.

EXERCISE 4.26 (Holm as a graphical procedure). Show that Holm's procedure is the graphical procedure obtained by giving each hypothesis initial alpha α/m and transition weights $g_{ij} = 1/(m-1)$ for $i \neq j$. Assume rejected hypotheses are removed one at a time, with arbitrary tie-breaking if several hypotheses are eligible. Verify that the graph update after r rejections gives each remaining hypothesis local level $\alpha/(m-r)$, the next Holm threshold.

EXERCISE 4.27 (Hommel and closed Simes). For a small family, say $m \leq 8$, implement the closed Simes procedure directly by testing all intersection hypotheses. Compare its rejection set with Hommel's procedure from a standard software implementation, and verify that they agree. State the dependence conditions under which the local Simes tests are valid, and contrast Hommel's rejection set with Holm's on an example where Hommel rejects more hypotheses.

CHAPTER 5

False Discovery Rate and the BH Principle

Familywise error control is designed for settings where even one false rejection is costly. Many modern studies have a different objective. In a genome-wide screen, a model-comparison benchmark, or a feature-discovery pipeline, the analyst may be willing to tolerate some false discoveries if the overall list remains reliable. False discovery rate is the standard error rate for that regime.

The central procedure in this chapter is the Benjamini-Hochberg procedure. The right way to understand BH is not as a list of sorted p-value instructions, but as a self-consistent threshold: choose a cutoff only when the estimated fraction of false discoveries below that cutoff is small enough.

1. FDP and FDR

Use the same decision table as in Chapter 4. Let V be the number of false rejections and R the total number of rejections. The false discovery proportion is

$$\mathrm{FDP} = \frac{V}{R \vee 1}.$$

The false discovery rate is its expectation:

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

The convention $R \vee 1$ makes $\mathrm{FDP} = 0$ when no hypotheses are rejected.

FDR is weaker than FWER. Since $\mathrm{FDP} \leq \mathbf{1}\{V \geq 1\}$,

$$\mathrm{FDR} \leq \mathrm{FWER}\,.$$

Thus any FWER procedure controls FDR, but it may be unnecessarily conservative for exploratory discovery. Conversely, FDR control does not guarantee that the realized FDP in a particular study is below α. It controls the average of that random fraction across repeated studies under the assumed model. Nor does FDR control bound the probability that the realized FDP exceeds α. Procedures that target realized or post-hoc FDP statements are different objects: for example, simultaneous FDP confidence bounds give guarantees for many data-chosen rejection sets at once [53]. BH should therefore be read as an expected-error-rate guarantee, not as a certificate that every selected list has small FDP.

2. The BH Step-Up Rule

Let $p_1, \ldots, p_m$ be p-values and write

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

for the order statistics. The BH procedure at level α finds

$$\widehat{k} = \max\left\{ i : \; p_{(i)} \leq \frac{\alpha i}{m} \right\},$$

with $\widehat{k} = 0$ if the set is empty. It rejects all hypotheses with

$$p_i \le p_{(\widehat{k})}$$

when $\widehat{k} > 0$, and rejects nothing when $\widehat{k} = 0$.

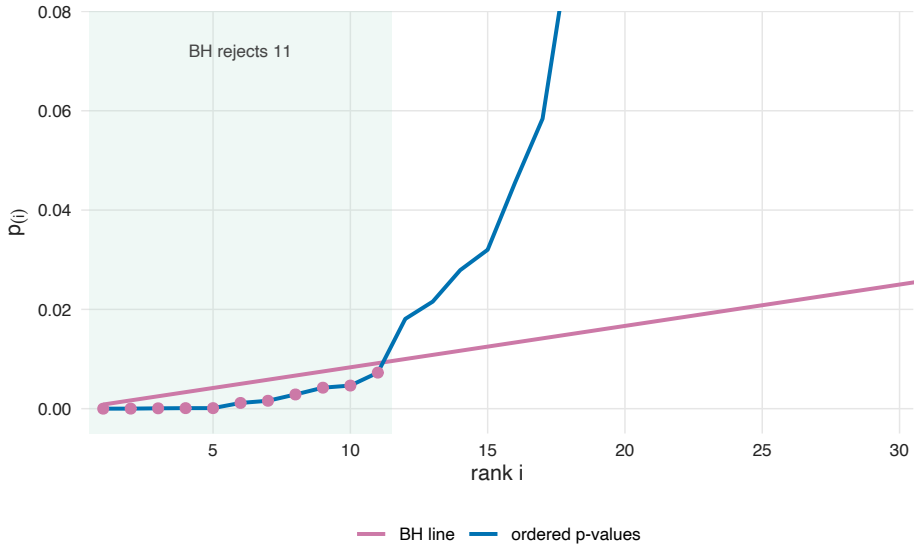

FIGURE 1. BH in ordered-p-value form. The procedure finds the largest rank at which $p_{(i)} \le \alpha i/m$ and rejects all p-values up to that rank.

For very small m, the difference between FWER and FDR procedures can be seen geometrically. Figure 2 colors the number of rejections made at each p-value configuration. Bonferroni and Holm are FWER procedures from Chapter 4; BH is the FDR step-up rule. For $m = 2$, Bonferroni rejects each hypothesis with $p_i \le \alpha/2$. Holm Thus, in the Bonferroni panel, the blue arms are one-rejection regions: exactly one of p_1, p_2 is below $\alpha/2$. Holm rejects one hypothesis if $p_{(1)} \le \alpha/2$ and $p_{(2)} > \alpha$, and rejects both if $p_{(1)} \le \alpha/2$ and $p_{(2)} \le \alpha$. BH rejects one hypothesis under the same one-rejection condition, but rejects both whenever $p_{(2)} \le \alpha$. Thus BH includes the square $\alpha/2 < p_1, p_2 \le \alpha$, where both p-values are moderately small but neither passes the Bonferroni threshold.

For $m = 3$, the figure shows the ordered slice with $p_{(3)} = 0.20 > \alpha$, so at most two hypotheses can be rejected. Bonferroni rejects $\#\{i : p_i \le \alpha/3\}$ hypotheses. Holm rejects none if $p_{(1)} > \alpha/3$, one if $p_{(1)} \le \alpha/3$ and $p_{(2)} > \alpha/2$, and two if $p_{(1)} \le \alpha/3$ and $p_{(2)} \le \alpha/2$. BH rejects two whenever $p_{(2)} \le 2\alpha/3$, and otherwise rejects one only if $p_{(1)} \le \alpha/3$. This is the geometric face of the tradeoff: FWER procedures protect against any false rejection, while BH protects the expected false discovery proportion.

The same rule can be written as a threshold rule. For $0 \le t \le 1$, define

$$\mathcal{R}(t) = \{i : p_i \le t\}, \qquad R(t) = |\mathcal{R}(t)|.$$

If all null p-values were exactly uniform, then at threshold t one would expect about $m_0 t$ false discoveries, where m_0 is the number of true nulls. Since $m_0 \le m$, mt is a conservative estimate. BH chooses

$$\widehat{T} = \frac{\alpha \widehat{k}}{m}, \qquad \widehat{k} = \max\left\{ i : \; p_{(i)} \le \frac{\alpha i}{m} \right\},$$

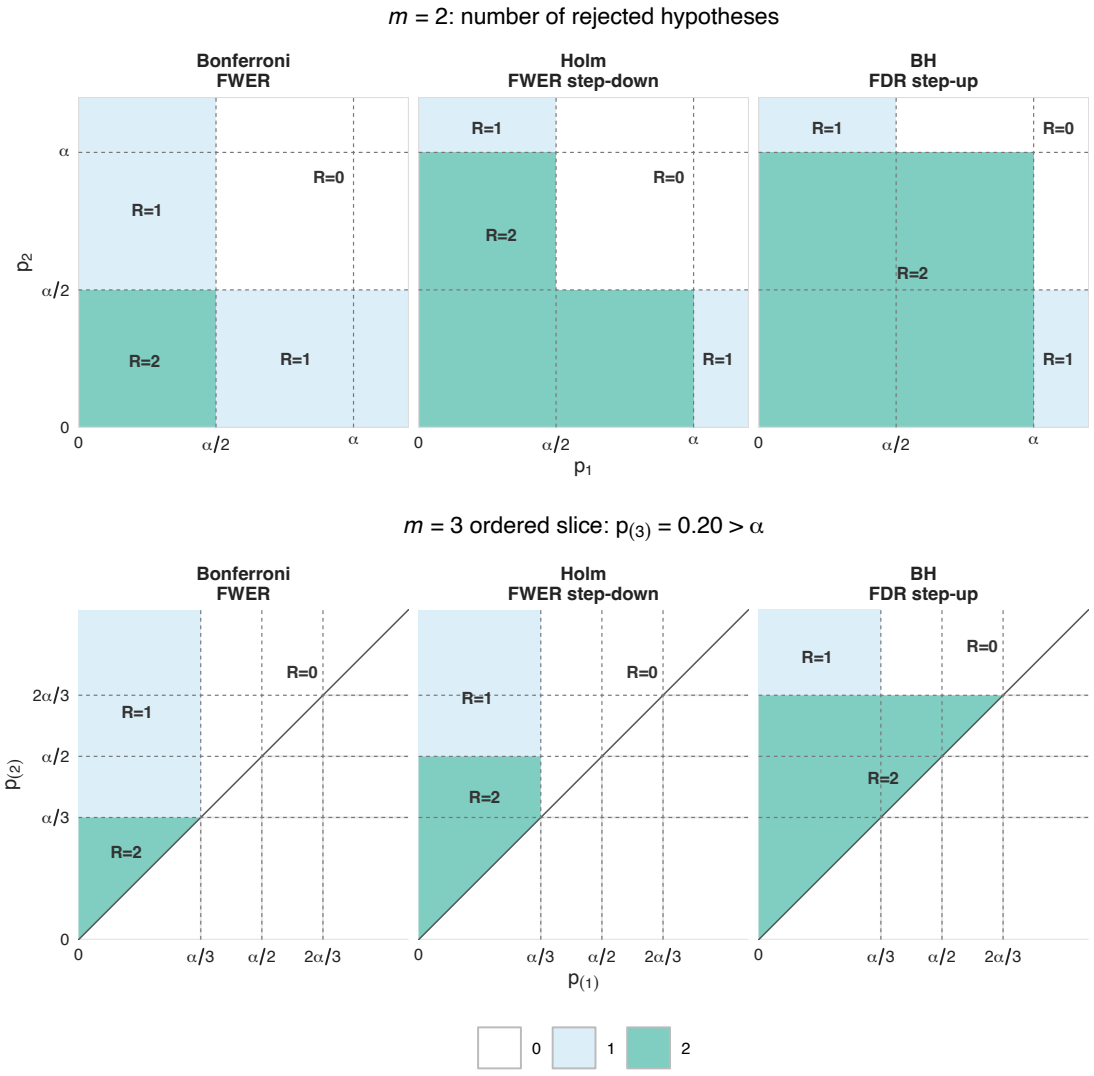

FIGURE 2. Rejection regions at $\alpha = 0.10$ for $m = 2$ and $m = 3$. Color gives the number R of rejected hypotheses: white is $R = 0$, blue is $R = 1$, and green is $R = 2$. Each panel recomputes those regions for its own method. The $m = 2$ row shows the full (p_1, p_2) square near the origin. The $m = 3$ row shows the ordered slice $(p_{(1)}, p_{(2)})$ with $p_{(3)} = 0.20 > \alpha$. Bonferroni rejects individual p-values below the first critical value. Holm requires the ordered p-values to pass the FWER step-down critical values. BH uses $\alpha i/m$, so its two-rejection region is visibly larger in the ordered $m = 3$ slice.

again with $\widehat{k} = 0$ if the set is empty. It rejects $p_i \le \widehat{T}$ when $\widehat{k} > 0$, and rejects nothing when $\widehat{k} = 0$. Equivalently, when at least one p-value is rejected, $\widehat{T}$ is the largest positive threshold satisfying

$$R(t) > 0, \qquad \frac{mt}{R(t)} \le \alpha.$$

The condition $R(t) > 0$ is important: without it the threshold set would include tiny t's below $p_{(1)}$, even in the no-rejection case. The numerical value $\widehat{T}$ is not generally an observed p-value: it sits in the open interval $(p_{(\widehat{k})}, p_{(\widehat{k}+1)})$ whenever $p_{(\widehat{k})} < \alpha\widehat{k}/m < p_{(\widehat{k}+1)}$. What matters for the rejection decision is that, for $\widehat{k} > 0$, the set $\{i : p_i \le \widehat{T}\}$ coincides with $\{i : p_i \le p_{(\widehat{k})}\}$, so the rejection set is the ordered-p-value rejection set above.

There is a third view, useful for visualizing dependence corrections. The empirical CDF of the p-values is $F_m(t) = R(t)/m$, so the BH inequality $mt/R(t) \le \alpha$ at thresholds with $R(t) > 0$ is the same as

$$F_m(t) \ge \frac{t}{\alpha},$$

i.e., the empirical CDF must lie above a line of slope $1/\alpha$ through the origin. The rejection decision is still determined by the largest ordered rank $\widehat{k}$ satisfying $p_{(\widehat{k})} \le \alpha\widehat{k}/m$. The canonical threshold $\widehat{T} = \alpha\widehat{k}/m$ is the right endpoint of the corresponding ECDF step, so it can lie to the right of the last rejected observed p-value $p_{(\widehat{k})}$. For BY the analogous endpoint is $\widehat{T}_{\mathrm{BY}} = \alpha\widehat{k}/(mH_m)$. In both cases the endpoint and the observed $p_{(\widehat{k})}$ give the same rejection set whenever no observed p-values fall between them, but they are different numerical thresholds. Replacing the line by a steeper boundary corresponds to a stricter procedure: this is how BY's harmonic correction enters in Figure 3 below.

3. Adjusted P-Values

BH-adjusted p-values report the smallest FDR level at which each hypothesis would be rejected by BH. In ordered form,

$$q_{(i)} = \min_{j \ge i} \left\{ \frac{mp_{(j)}}{j} \right\} \wedge 1.$$

The running minimum over $j \ge i$ is essential: adjusted p-values must be monotone in the ordered ranks. The value $q_{(i)}$ is then assigned back to the hypothesis that produced $p_{(i)}$.

Example 5.1. Suppose $m = 5$ and the ordered p-values are

$$0.003, \quad 0.014, \quad 0.041, \quad 0.20, \quad 0.62.$$

The raw BH ratios $mp_{(i)}/i$ are

$$0.015, \quad 0.035, \quad 0.0683, \quad 0.25, \quad 0.62.$$

These are already increasing, so they are the BH-adjusted p-values in ordered rank. At level 0.05, BH rejects the first two hypotheses.

These adjusted p-values $q_{(i)}$ are sometimes also called *q-values*. They use m as a conservative upper bound on the number of true nulls. Chapter 6 replaces m by an adaptive estimate $\widehat{\pi}_0 m$, giving a Storey-adjusted version of the same formula; that version is what is conventionally meant by "q-value" in the empirical-Bayes literature.

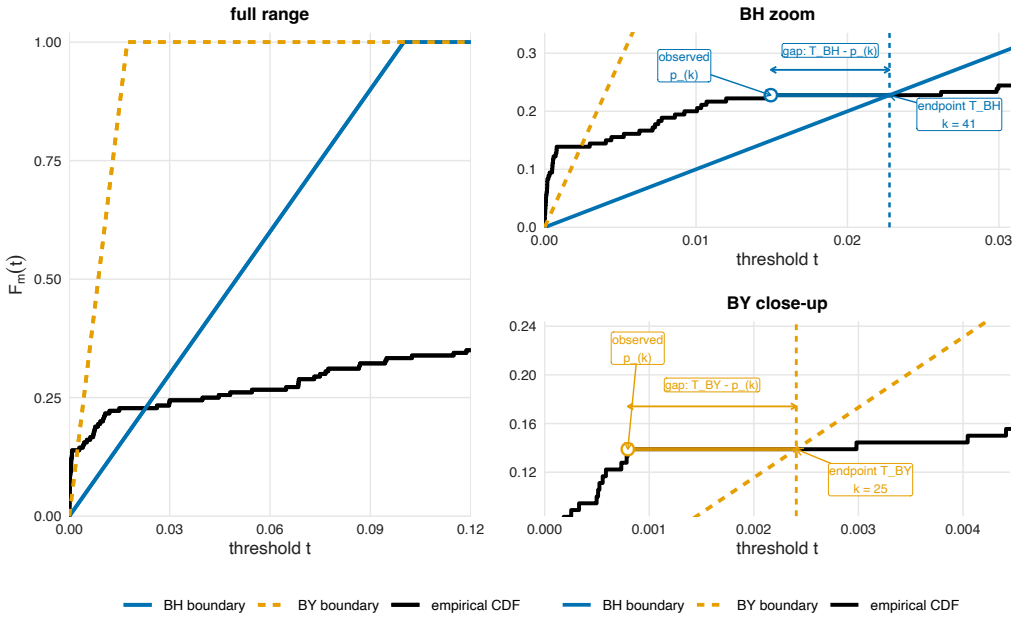

FIGURE 3. Empirical-CDF view of BH and BY. The ECDF is a step function. Open circles mark the observed last rejected p-values $p_{(\widehat{k})}$, while dashed vertical lines mark the selected endpoint thresholds $\alpha\widehat{k}/m$ and $\alpha\widehat{k}/(mH_m)$. The zoom panels explicitly mark the slack intervals $T_{\mathrm{BH}} - p_{(\widehat{k})}$ and $T_{\mathrm{BY}} - p_{(\widehat{k})}$: there are no observed p-values in such an interval, so the endpoint threshold and the last rejected observed p-value produce the same rejection set even though they are different numbers.

4. What the BH Estimate Means

BH uses

$$\widehat{\mathrm{FDP}}_{\mathrm{BH}}(t) = \frac{mt}{R(t) \vee 1}$$

as a conservative estimate of the false discovery proportion. It is not a guaranteed upper bound on the realized FDP for every threshold or every data set. Its role is more subtle: under independence, the random threshold chosen by BH interacts with null p-values in a way that controls the expectation of FDP.

5. The Independent-Null Proof

THEOREM 5.2 (BH under independent nulls). *Suppose the true-null p-values are independent uniforms and are independent of the false-null p-values. Then BH at level α satisfies*

$$\mathrm{FDR} \leq \frac{m_0}{m}\alpha \leq \alpha,$$

where m_0 is the number of true nulls.

PROOF. Let $\mathcal{I}_0$ be the set of true nulls. Write R for the number of BH rejections. Then

$$\mathrm{FDR} = \sum_{i \in \mathcal{I}_0} \mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\right].$$

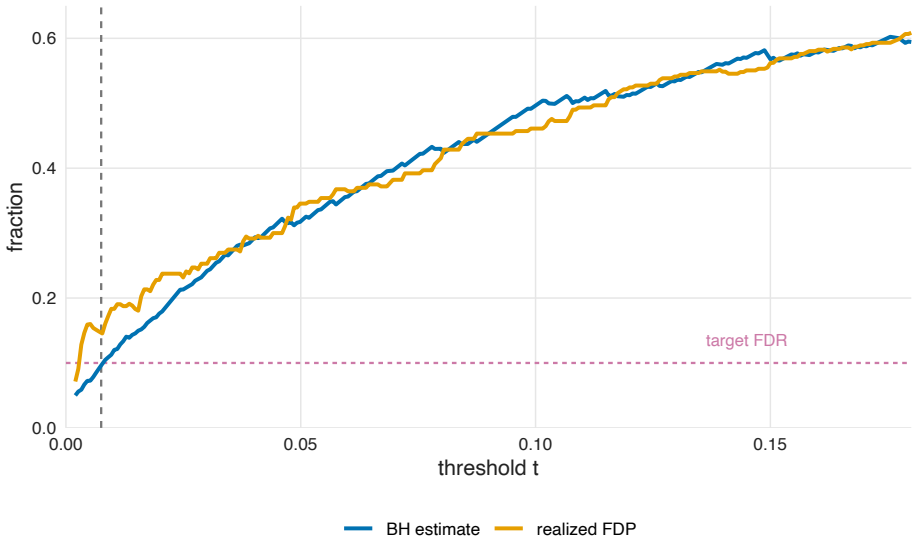

FIGURE 4. A single simulated path of the realized FDP and the BH estimator $mt/R(t)$. The estimator need not dominate the realized FDP pointwise; the BH theorem is an expectation statement at the selected threshold.

For a true null i, define R_i as the number of rejections that BH would make if p_i were set to zero while all other p-values were left fixed. Two observations make the leave-one-out argument go through. First, setting $p_i = 0$ forces the smallest p-value in the modified vector to be 0, which satisfies the smallest BH threshold $0 \le \alpha/m$; the modified BH procedure therefore rejects at least one hypothesis, giving $R_i \ge 1$ deterministically. Second, on the event that i is rejected by the original procedure, self-consistency of BH gives

$$p_i \le \frac{\alpha R_i}{m} \qquad \text{and} \qquad R = R_i.$$

The identity $R = R_i$ is the key leave-one-out fact and deserves a line. If i is rejected, then by BH self-consistency, $p_i \le \alpha R/m$; lowering p_i further to 0 only moves this already rejected p-value earlier among the first R ordered positions. The R originally rejected p-values still satisfy the rank-R BH inequality. For any rank $s > R$, the sth ordered p-value after lowering p_i is the same as the original sth ordered p-value, because only the already rejected p_i moved within the first R positions. Since R was the largest crossing, no new rank $s > R$ can cross. Hence the rejection count is unchanged and $R_i = R$. In particular, $R \vee 1 = R_i$ on this event, so

$$\frac{\mathbf{1}\{i \text{ rejected}\}}{R \vee 1} = \frac{\mathbf{1}\{p_i \le \alpha R_i/m\}}{R_i}.$$

Conditional on $\{p_j : j \ne i\}$, the quantity R_i is fixed and $p_i \sim \mathrm{Unif}(0,1)$. Therefore

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \le \alpha R_i/m\}}{R_i}\,\middle|\,\{p_j : j \ne i\}\right] = \frac{1}{R_i} \cdot \frac{\alpha R_i}{m} = \frac{\alpha}{m}.$$

Summing over the m_0 true nulls and taking expectations yields $m_0\alpha/m \le \alpha$. □

COROLLARY 5.3 (BH under independent super-uniform nulls). *Suppose the true-null p-values are independent, super-uniform, and independent of the false-null p-values. Then BH at level α satisfies*

$$\mathrm{FDR} \le \frac{m_0}{m}\alpha \le \alpha.$$

PROOF. Repeat the leave-one-out proof of Theorem 5.2. Conditional on $\{p_j : j \neq i\}$, the leave-one-out count R_i is fixed and $\alpha R_i/m \in [0,1]$. Super-uniformity gives

$$\mathbb{P}\{p_i \leq \alpha R_i/m \mid \{p_j : j \neq i\}\} \leq \frac{\alpha R_i}{m},$$

so the contribution of each true null is at most α/m instead of being exactly α/m. □

This proof is the template for much of modern FDR theory. The important features are self-consistency of the rejection threshold and a leave-one-out argument that separates one true-null p-value from the random threshold.

6. Z-Score Thresholds

Many large-scale testing problems begin with z-scores rather than p-values. For two-sided tests, let

$$R(t) = \sum_{i=1}^{m} \mathbf{1}\{|Z_i| \geq t\}.$$

Under a standard normal null, the expected number of null z-scores beyond $\pm t$ is at most $2m\{1-\Phi(t)\}$, so this is a conservative estimate of the false-discovery count. The clean way to state the procedure is to form the two-sided p-values

$$p_i = 2\{1-\Phi(|Z_i|)\}$$

and apply ordinary BH. Thus

$$\widehat{k} = \max\{k : p_{(k)} \leq \alpha k/m\},$$

with no rejections when the set is empty. If $\widehat{k} > 0$, set

$$\widehat{s} = \alpha\widehat{k}/m, \qquad \widehat{T} = \Phi^{-1}(1-\widehat{s}/2),$$

and reject $|Z_i| \geq \widehat{T}$. If $\widehat{k} = 0$, no z-score threshold is reported and the rejection set is empty.

Equivalently, when at least one BH rejection occurs, $\widehat{T}$ is the smallest observed-relevant t at which the z-scale FDP estimate is controlled:

$$\widehat{T} = \inf\left\{t \geq 0 : \frac{2m\{1-\Phi(t)\}}{R(t) \vee 1} \leq \alpha\right\},$$

but the no-rejection case is not an empty search over all $t \geq 0$: as $t \to \infty$, the numerator tends to zero. The no-rejection case is simply $\widehat{k} = 0$ in the p-value BH rule, equivalently no observed candidate threshold satisfies the BH inequality. Theorem 5.2 therefore gives FDR $\leq \alpha$ under independent null z-scores, with no additional calculation needed.

7. Weighted BH

External information can improve power. Some hypotheses may be more promising, more scientifically important, or measured with higher precision. A fixed weighted BH procedure assigns weights $w_i \geq 0$ satisfying

$$\frac{1}{m}\sum_{i=1}^{m} w_i = 1.$$

It applies BH to the transformed p-values

$$\frac{p_i}{w_i},$$

with the convention that hypotheses with $w_i = 0$ cannot be rejected. Larger weights make a hypothesis easier to reject; smaller weights make it harder. When $w_i > 1$, the transformed

quantity p_i/w_i is not itself a valid p-value under the null; the validity comes from the normalized weight budget. Section 9 proves this by the same leave-one-out argument as ordinary BH: a true null with weight w_i contributes at most $\alpha w_i/m$, and these charges sum to at most α because $\sum_i w_i = m$.

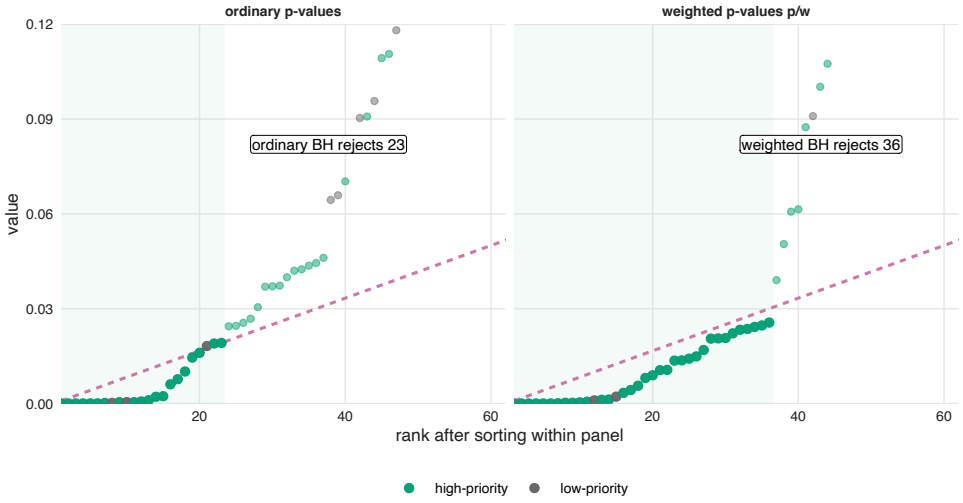

FIGURE 5. Weighted BH as BH on transformed p-values p_i/w_i. High-priority hypotheses receive larger weights, so their transformed p-values move downward relative to low-priority hypotheses; in this simulated example the weighted rule rejects visibly more hypotheses than ordinary BH.

Weights must be fixed independently of the null p-values they are used to test, or learned in a way justified by additional theory such as sample splitting, external covariates, or structural modeling. Reusing the same noise twice can destroy FDR control.

8. Learning Weights from Covariates: IHW

Modern multiple-testing problems often come with auxiliary information for each hypothesis: gene length, technical replicates, baseline expression, prior significance, distance to a known feature, or a side-data summary derived from an entirely different experiment. Let $X_i \in \mathcal{X}$ denote such side information for hypothesis i. The ideal covariate is informative about power or the prior chance of being nonnull, but under a true null it does not use the focal testing noise that produced p_i.

Weighted BH explains how to use fixed side information. Independent hypothesis weighting (IHW) asks the next question: can the weights themselves be learned from the collection of covariates and p-values? The answer is yes when the learning is arranged so that the weight assigned to a true null is not learned from that null p-value.

Ignatiadis et al. [63] implement this idea with fold splitting. The clean cross-fitted version below is a sufficient recipe for the leave-one-out proof. Under mutual independence, a weight for hypothesis i may depend on other p-values, even other p-values in the same fold, because those p-values are independent of a true-null p_i. What the finite-sample proof cannot allow is for the final weight assigned to i, or a normalizing constant that changes it, to depend on p_i itself. A simplified cross-fitted version that enforces this condition is:

(1) Partition the hypotheses into K folds without using the focal p-values; in applications the folds are usually balanced across covariate bins.
(2) For each fold k, use hypotheses outside fold k to learn a nonnegative weight function $\widehat{w}_{-k}(x)$ of the covariate. The training folds may use both their covariates and p-values to choose this function. IHW commonly learns piecewise-constant weights over covariate bins, with regularization to avoid unstable choices.
(3) For each held-out hypothesis $i \in F_k$, set a raw weight $\widetilde{w}_i = \widehat{w}_{-k}(X_i)$. This evaluates the learned function on the held-out covariate X_i, not on the held-out p-value p_i; p_i is saved for the final weighted BH step. After all folds have received raw weights, normalize globally so that
$$\sum_{i=1}^{m} w_i = m.$$
This normalization is part of the weight-construction step; the final normalized weights must still satisfy the leave-one-out exogeneity condition stated below.
(4) Run one weighted BH procedure on the transformed p-values p_i/w_i, with $w_i = 0$ interpreted as making hypothesis i unrejectable.

The optimization used to choose the bin weights is not the main point here. The main point is the exogeneity condition. Power considerations naturally push us to use as much p-value information as possible when learning weights. With mutually independent p-values, the weight for i can depend on $\{p_j : j \neq i\}$, including p-values from the same fold, without depending on p_i. Thus a more data-efficient implementation can be valid if it is effectively leave-one-out for every focal null. By contrast, fitting one fold-level weight rule using all p-values in F_k and assigning it back to every member of F_k usually makes w_i depend on p_i, and then this simple finite-sample proof no longer applies without an additional masking, monotonicity, or stability argument.

PROPOSITION 5.4 (Leave-one-out weighting validity). *Suppose the p-values are mutually independent, each true-null p-value is uniform, and the final nonnegative weight vector $W = (w_1, \ldots, w_m)$ satisfies $\sum_i w_i = m$. Assume also that, for every true null i, the final weight vector used by the BH step can be conditioned on without using p_i; equivalently, after conditioning on the information that produced W and on the other p-values, p_i remains uniform and W is fixed. Then weighted BH applied to p_i/w_i controls FDR at level α.*

PROOF. Condition on the learned weights and on all p-values except the focal true-null p-value p_i. Under the assumption above, w_i is fixed in this conditional calculation and p_i remains uniform. Let R_i be the rejection count after setting $p_i = 0$, using the same weights. On the event that i is rejected, the weighted BH leave-one-out identity gives $R = R_i$ and
$$p_i \leq \frac{\alpha w_i R_i}{m}.$$
Therefore
$$\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\middle| \text{ other data}\right] \leq \mathbb{E}\left[\frac{\mathbf{1}\{p_i \leq \alpha w_i R_i/m\}}{R_i}\middle| \text{ other data}\right] \leq \frac{\alpha w_i}{m}.$$
If $\alpha w_i R_i/m > 1$, the same bound is trivial because then $1/R_i \leq \alpha w_i/m$. Summing over true nulls gives
$$\text{FDR} \leq \sum_{i \in \mathcal{I}_0} \frac{\alpha w_i}{m} \leq \alpha,$$
since $\sum_i w_i = m$. Averaging over the data used to learn the weights completes the argument. □

This proposition is deliberately cleaner than the full IHW theorem. It captures the reason fold splitting is needed, while the actual IHW analysis also verifies that the fold construction,

regularization, and normalization preserve the required leave-one-out structure. If the same p-values are used both to choose their own weights and to test themselves, small null p-values can receive large weights precisely because they are small, and the weighted-BH charge calculation no longer applies.

More adaptive covariate-threshold methods require additional ideas. In particular, procedures such as AdaPT use local-FDR-style models to propose covariate-dependent threshold surfaces, and masking or mirror symmetry to keep the adaptive search valid. Those ingredients are introduced after local FDR in Chapter 6.

9. A Generalized BH Template

A broad class of BH-type rules can be put in a single framework. Let $\psi_i : [0,1] \to [0,\infty)$ be a coordinate-specific strictly increasing transformation with $\psi_i(0) = 0$, so that the rejection event $\psi_i(p_i) \le t$ is equivalent to

$$p_i \le \psi_i^{-1}(t) \wedge 1,$$

where the cap at 1 handles inputs t for which the inverse exceeds the unit interval (in which case the inequality holds trivially because $p_i \le 1$). When we write the rejection threshold below we always interpret $\psi_i^{-1}(t)$ capped at 1 without further comment. Let $g(t)$ be an increasing boundary function with $g(0) = 0$, and define

$$R(t) = \sum_{i=1}^{m} \mathbf{1}\{p_i \le \psi_i^{-1}(t)\}.$$

To avoid threshold-attainment distractions, state the template on a deterministic finite grid $\mathcal{T} \subset (0,1]$. Continuous-threshold versions require the same proof plus one-sided continuity or a largest-attained-threshold convention. The generalized BH threshold is

$$\widehat{T} = \max\left(\{0\} \cup \left\{t \in \mathcal{T} : \frac{mg(t)}{R(t) \vee 1} \le \alpha\right\}\right).$$

The procedure rejects

$$p_i \le \psi_i^{-1}(\widehat{T}).$$

Proposition 5.5 (Generalized BH bound). *Assume the true-null p-values are independent uniforms and are independent of the false-null p-values. Define*

$$C = \frac{1}{m} \sum_{i \in \mathcal{I}_0} \sup_{t \in \mathcal{T}} \frac{\psi_i^{-1}(t)}{g(t)}.$$

Assume additionally the following leave-one-out stability property: for every true null i, if i is rejected by the original procedure and $(R_i, \widehat{T}_i)$ are the rejection count and threshold after setting $p_i = 0$, then $R_i = R$ and $p_i \le \psi_i^{-1}(\widehat{T}_i)$. Ordinary BH and fixed weighted BH have this property because they are ordinary BH procedures applied to transformed p-values. Then the generalized BH procedure satisfies

$$\mathrm{FDR} \le C\alpha.$$

Proof. The proof is the same leave-one-out argument as for BH. For a true null i, set $p_i = 0$ and let R_i and $\widehat{T}_i$ be the resulting rejection count and threshold. On the event that the original procedure rejects i, leave-one-out stability ensures that the attained grid threshold $\widehat{T}_i$ is a threshold at which the leave-one-out procedure rejects, so the self-consistency relation reads

$$\frac{mg(\widehat{T}_i)}{R_i} \le \alpha.$$

Conditioning on all other p-values makes R_i and $\widehat{T}_i$ fixed, while p_i is uniform. By leave-one-out stability, the contribution of hypothesis i to the FDR is bounded by

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \le \psi_i^{-1}(\widehat{T}_i)\}}{R_i}\,\middle|\,\{p_j : j \ne i\}\right] = \frac{\psi_i^{-1}(\widehat{T}_i)}{R_i} \le \frac{\alpha}{m}\cdot\frac{\psi_i^{-1}(\widehat{T}_i)}{g(\widehat{T}_i)} \le \frac{\alpha}{m}\sup_{t\in\mathcal{T}}\frac{\psi_i^{-1}(t)}{g(t)}.$$

Summing over $i \in \mathcal{I}_0$ and using the definition of C gives the result. □

For ordinary BH, $\psi_i^{-1}(t) = t$ and $g(t) = t$, so

$$C = \frac{m_0}{m}.$$

For weighted BH with $\psi_i(p) = p/w_i$ and weights normalized so that $\sum_i w_i = m$, one has $\psi_i^{-1}(t) = w_i t$; taking $g(t) = t$ gives

$$C = \frac{1}{m}\sum_{i\in\mathcal{I}_0} w_i \le \frac{1}{m}\sum_{i=1}^{m} w_i = 1.$$

Thus fixed weighted BH controls FDR at level α. The inequality is strict when larger weights are assigned mostly to nonnull hypotheses, which is why genuinely external prior information can improve power.

The value of this template is conceptual. It shows that BH is one member of a larger family of self-consistent thresholding rules. Later chapters use the same logic for dependence, adaptive estimation, and e-value procedures.

10. Arbitrary Dependence: The BY Correction

BH is valid under independence and, as Chapter 6 will explain, under certain positive-dependence conditions. Under arbitrary dependence the BH procedure can fail. Where does the failure come from? The leave-one-out proof of Theorem 5.2 used the unconditional uniform distribution of a true-null p_i to compute the expectation $\mathbb{E}[\mathbf{1}\{p_i \le \alpha R_i/m\}/R_i] = \alpha/m$; under arbitrary dependence, the random threshold R_i is correlated with p_i, and the same expectation can be larger. A careful arbitrary- dependence calculation by Benjamini and Yekutieli [16] shows that FDR $\le \alpha H_m$ under arbitrary dependence, where

$$H_m = \sum_{j=1}^{m}\frac{1}{j}$$

is the harmonic number. Dividing the nominal level by H_m restores the α bound. The BY procedure rejects according to

$$p_{(i)} \le \frac{\alpha i}{m H_m},$$

i.e., BY is BH run at level α/H_m. Since $H_m = \log m + O(1)$, the power cost can be substantial, but the benefit is robustness: no dependence structure among the p-values is required.

THEOREM 5.6 (BY correction). *Assume the true-null p-values are marginally continuous uniforms, with arbitrary dependence allowed among all p-values. The BY procedure at level α, equivalently BH run at level α/H_m, satisfies*

$$\mathrm{FDR} \le \frac{m_0}{m}\alpha \le \alpha.$$

PROOF. The displayed proof uses exact continuous uniformity of the true-null p-values. For merely super-uniform p-values, one should use the standard reshaping formulation or state the

corresponding conservative interval-probability inequalities explicitly. It is useful to prove a slightly more general bound for a step-up rule with nondecreasing critical values

$$0 = a_0 \le a_1 \le \cdots \le a_m.$$

Let R be the number of rejections. Step-up self-consistency says that if hypothesis i is rejected and $R = r$, then $p_i \le a_r$. Hence, for a true null i,

$$\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1} \le \frac{\mathbf{1}\{p_i \le a_R\}}{R \vee 1}.$$

On the event $p_i \in (a_{s-1}, a_s]$, the inequality $p_i \le a_R$ can hold only if $R \ge s$. Therefore, path by path,

$$\frac{\mathbf{1}\{p_i \le a_R\}}{R \vee 1} \le \sum_{s=1}^{m} \frac{1}{s} \mathbf{1}\{a_{s-1} < p_i \le a_s\}.$$

Taking expectations and using the marginal uniformity of the true-null p-value gives

$$\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\right] \le \sum_{s=1}^{m} \frac{a_s - a_{s-1}}{s}.$$

For BH run at nominal level q, $a_s = qs/m$, so

$$\sum_{s=1}^{m} \frac{a_s - a_{s-1}}{s} = \frac{q}{m} \sum_{s=1}^{m} \frac{1}{s} = \frac{qH_m}{m}.$$

Summing over the m_0 true nulls shows that BH at level q has

$$\mathrm{FDR} \le \frac{m_0}{m} q H_m.$$

Setting $q = \alpha / H_m$ proves the BY bound. □

Remark 5.7 (Reshaping the step-up boundary). The proof shows more than the harmonic correction. For any step-up rule with critical values $a_s = \alpha\beta(s)/m$, where $\beta(0) = 0 \le \beta(1) \le \cdots \le \beta(m)$, the same argument gives

$$\mathrm{FDR} \le \frac{m_0 \alpha}{m} \sum_{s=1}^{m} \frac{\beta(s) - \beta(s-1)}{s}.$$

Thus arbitrary-dependence control follows whenever the displayed sum is at most one. Ordinary BH uses $\beta(s) = s$, giving the factor H_m; BY uses $\beta(s) = s/H_m$. This is the simplest instance of the reshaping viewpoint for self-consistent FDR procedures: the rejection boundary is bent downward just enough that the worst-case dependence calculation still sums to one. More general reshaped and self-consistent procedures are developed in Blanchard and Roquain [21].

11. Mirror Estimation

Another way to estimate false discoveries uses symmetry. Suppose null p-values are uniform, while alternatives mostly produce small p-values. Then the upper tail near one can act as a mirror for the lower null tail. For $0 < t \le 1/2$, define

$$\widehat{\mathrm{FDP}}_{\mathrm{mir}}(t) = \frac{1 + \#\{i : p_i \ge 1 - t\}}{R(t) \vee 1}.$$

The plus one stabilizes the estimate. A mirror procedure chooses

$$\widehat{T}_{\mathrm{mir}} = \sup\left\{0 < t \le \frac{1}{2} : \widehat{\mathrm{FDP}}_{\mathrm{mir}}(t) \le \alpha\right\},$$

with $\widehat{T}_{\mathrm{mir}} = 0$ if the set is empty, and rejects $p_i \le \widehat{T}_{\mathrm{mir}}$.

This rule is attractive because it estimates the number of false discoveries from the data rather than using the conservative upper bound mt. Under suitable joint sign-symmetry and stopping-time conditions – for example, conditionally independent null signs given magnitudes and nonnull information, with the threshold chosen by a valid sign-revealing filtration – the procedure satisfies

$$\mathrm{FDR} \le \alpha.$$

The argument is a martingale stopping-time argument on the sign-flip variables $\mathbf{1}\{p_i \le 1/2\} - \mathbf{1}\{p_i > 1/2\}$; the same sign-symmetry logic appears in the knockoff proofs of Chapter 9, so we defer the detailed argument to Barber and Candès [6]. Three caveats matter in practice: the orientation of the mirror count is essential (counting $p_i \ge 1 - t$ rather than $p_i \le 1 - t$), the plus-one stabilizer is what gives finite-sample control, and dependence among null p-values is not free — the sign-flip argument needs symmetry under the joint null, not merely marginal symmetry.

Figure 6 uses nominal level $\alpha = 0.10$. BH rejects where the conservative estimate $mt/R(t)$ first reaches this level; in the example this gives 53 rejections. The mirror rule uses the observed upper tail instead. Because the upper-tail mirror count is small at the selected threshold, $(1 + \#\{p_i \ge 1 - t\})/(R(t) \vee 1)$ stays near 0.10 until a larger threshold and the rule rejects 84 hypotheses. This is a data-adaptive power gain in this symmetric example, not a guarantee that mirror always dominates BH.

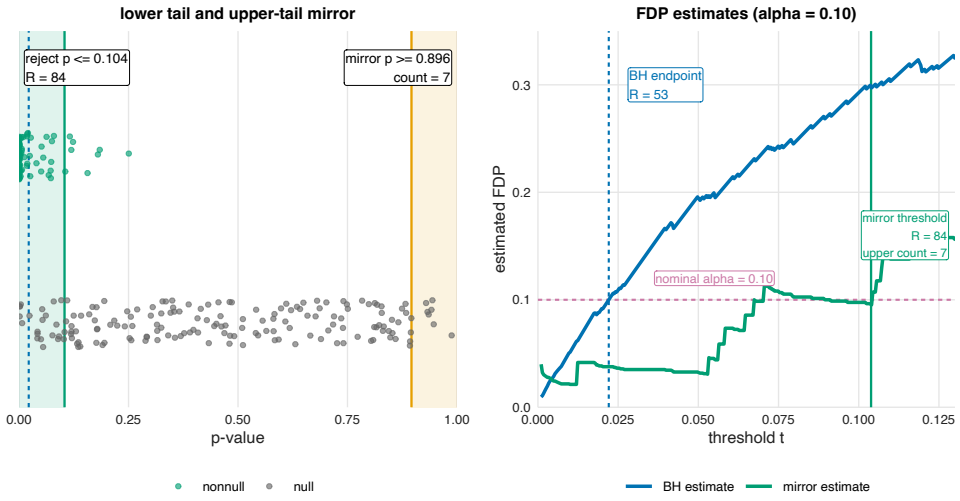

FIGURE 6. Mirror estimation in one simulated example at nominal level $\alpha = 0.10$. The lower tail supplies candidate discoveries, while the upper tail $p_i \ge 1 - t$ estimates the null contribution in the lower tail. Compared with BH's conservative estimate $mt/R(t)$, the mirror estimate can allow a larger threshold when the upper tail is sparse; here BH rejects 53 hypotheses and the mirror rule rejects 84.

12. Assumptions in Plain Language

BH needs valid null p-values and a dependence condition that lets the random threshold be analyzed. Independence is the cleanest case; PRDS is treated in Chapter 6. BY sacrifices power to avoid dependence assumptions. Weighted BH is valid when the weights are external to the

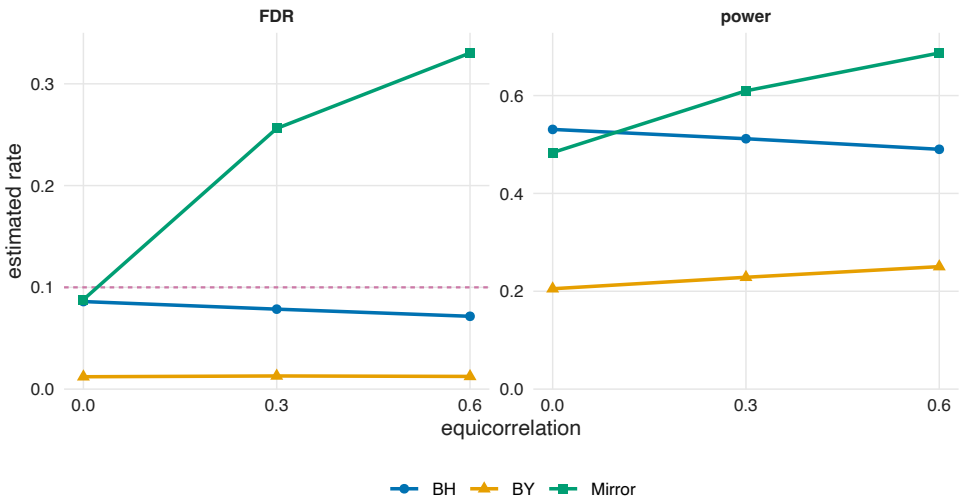

FIGURE 7. Simulation comparing BH, BY, and a mirror estimator. BY is robust but conservative. The mirror rule can gain power, but its validity is more sensitive to dependence and symmetry assumptions.

tested null noise. Mirror estimators use symmetry and require more care than the formula alone suggests.

13. Bibliographic Notes

The BH procedure was introduced by Benjamini and Hochberg [15]. The BY arbitrary-dependence correction is due to Benjamini and Yekutieli [16]. Generalized and reshaped FDR procedures are discussed by Blanchard and Roquain [21]. The mirror estimator and its sign-flip finite-sample argument are due to Barber and Candès [6], where the related knockoff filter is introduced. The empirical-process, adaptive, and empirical-Bayes extensions are deferred to Chapter 6.

Independent hypothesis weighting is introduced by Ignatiadis et al. [63]; the procedure is widely used in genomics, where the side information is typically a measure of statistical precision such as gene length or baseline mean expression. Masking-based and local-FDR-guided covariate thresholding, including AdaPT, is deferred to Chapter 6, after local FDR has been introduced. Group and multilayer extensions to grouped or tree-structured families are from Barber and Ramdas [7] and Yekutieli [132]; these are developed in detail in Chapter 7. A frontier topic not pursued in this chapter is *differentially private* BH, which uses Laplace-noised log p-values to obtain FDR control under (ε, δ)-differential privacy with a multiplicative inflation factor; see Dwork et al. [41] for a treatment in the privacy-utility literature.

14. Exercises

Basic.

EXERCISE 5.8 (FDP versus FWER). For a testing problem with $m = 100$, $m_0 = 70$, $V = 3$, and $R = 8$, compute the FDP and decide whether a familywise error occurred. Then, keeping $m = 100$ and $m_0 = 70$ fixed, give the smallest value of R for which a procedure rejecting $V = 3$ true nulls would still achieve FDP ≤ 0.10, and verify that this R is feasible given how many nonnull hypotheses are available.

EXERCISE 5.9 (BH by hand). For p-values

$$0.004,\ 0.011,\ 0.018,\ 0.042,\ 0.067,\ 0.21,\ 0.33,\ 0.46,\ 0.71,\ 0.88,$$

compute the BH rejection set at $\alpha = 0.10$.

EXERCISE 5.10 (Adjusted p-values). For the same p-values, compute the BH-adjusted p-values. Be careful to apply the running minimum from the largest rank back to the smallest rank.

EXERCISE 5.11 (Threshold equivalence). Prove that the ordered-p-value definition of BH is equivalent to the threshold definition

$$\widehat{k} = \max\{i : p_{(i)} \le \alpha i/m\}, \qquad \widehat{T} = \alpha\widehat{k}/m,$$

with $\widehat{k} = 0$ if the set is empty. Show also that when $\widehat{k} > 0$, this $\widehat{T}$ is the largest threshold with $R(t) > 0$ and $mt/R(t) \le \alpha$, while $\widehat{k} = 0$ gives no rejections.

Intermediate.

EXERCISE 5.12 (Leave-one-out proof). Fill in the details of Theorem 5.2. In particular, prove the self-consistency relation between rejection of a true null and the leave-one-out rejection count R_i.

EXERCISE 5.13 (Z-score BH). Starting from two-sided p-values $p_i = 2\{1 - \Phi(|Z_i|)\}$, derive the z-score threshold used when BH makes at least one rejection:

$$\widehat{T} = \inf\left\{t \ge 0 : \frac{2m\{1-\Phi(t)\}}{R(t) \vee 1} \le \alpha\right\}.$$

Explain why the infimum is the right operator (rather than supremum) when larger $|Z_i|$ is stronger evidence. Also explain why the no-rejection case is $\widehat{k} = 0$ in the p-value BH rule, not an empty search over all $t \ge 0$.

EXERCISE 5.14 (Weighted BH). Show that fixed weighted BH is ordinary BH applied to p_i/w_i. State the normalization condition on the weights and explain why data-learned weights are not automatically valid.

EXERCISE 5.15 (IHW validity through folds). Suppose an IHW-style procedure splits the hypotheses into K folds. For each fold k, a weight function $\widehat{w}_{-k}(x)$ is learned using only $\{(p_j, X_j) : j \notin F_k\}$, assigned to held-out hypotheses by $w_i = \widehat{w}_{-k}(X_i)$ for $i \in F_k$, and normalized so that the final weights satisfy $\sum_i w_i = m$. Assume the normalization is leave-one-out-safe: when analyzing a focal true null i, the final weight vector used by the BH step is fixed after conditioning on data that exclude p_i. Condition on that information and on all p-values except p_i. Explain why p_i remains uniform and the weighted-BH leave-one-out proof applies. Under mutual independence, why would it also be safe for the weight of i to use p-values in F_k other than p_i? What can go wrong if p_i itself is used to tune the weight assigned to i, or if a global normalizing constant reintroduces dependence on p_i?

EXERCISE 5.16 (BY correction). Compute H_m for $m = 10, 100, 1000$. Compare the BH and BY critical values and discuss the power cost of arbitrary-dependence robustness.

Computational.

EXERCISE 5.17 (Pathwise FDP). Reproduce Figure 4. Explain why the BH estimator need not dominate the realized FDP for every threshold.

Exercise 5.18 (Simulation study). Simulate BH, BY, and the mirror procedure under independent and equicorrelated Gaussian z-scores. Report FDR, power, and the distribution of realized FDP. As a sanity check, under the all-null independent setting the empirical FDR should be close to the nominal level for BH, while BY should be visibly more conservative. The figure-generation function `fig04_bh_by_bc_sim()` in `scripts/make_figures.R` gives one reference implementation.

Advanced.

Exercise 5.19 (Mirror orientation). For the mirror estimator, explain why the numerator counts $\#\{p_i \geq 1-t\}$ rather than $\#\{p_i \leq 1-t\}$. What is the role of the plus-one term? Construct a small example with two independent null p-values showing that omitting the plus-one breaks the finite-sample bound.

Exercise 5.20 (Generalized BH). Specialize the generalized threshold template to weighted BH and verify the constant C under independent true-null p-values. Then show that ordinary BH is the case $\psi_i(p) = p$, $g(t) = t$, and that the BY procedure corresponds to a different choice of g (which one?).

CHAPTER 6

Dependence, Adaptive FDR, and Empirical Bayes

Chapter 5 presented BH under independent null p-values and then introduced BY as a robust but conservative arbitrary-dependence correction. In large-scale inference, neither independence nor worst-case dependence is a satisfactory default. Gene-expression measurements are correlated through pathways. Imaging statistics are spatially dependent. Benchmark scores share training data, prompts, and evaluation pipelines. Dependence is part of the problem.

This chapter has two goals. First, it explains when dependence is benign for BH. Positive dependence, formalized through PRDS, often preserves FDR control. Second, it explains how the null fraction and posterior null probabilities can be estimated. Storey's estimator, q-values, and local FDR all refine the basic BH idea, but they answer different questions and rely on different assumptions.

1. Positive Dependence and PRDS

The BH proof in Theorem 5.2 used a leave-one-out argument: a true-null p-value was separated from the random threshold. Under dependence, that separation is no longer automatic. PRDS is a condition that keeps the dependence in a favorable direction.

Because p-values are small when evidence is strong, one must be careful with monotonicity. For two p-value vectors $p = (p_1, \dots, p_m)$ and $q = (q_1, \dots, q_m)$, write $q \geq p$ when $q_j \geq p_j$ for every coordinate. In this chapter an event $A \subseteq [0,1]^m$ is called increasing if

$$p \in A, \quad q \geq p \quad \Longrightarrow \quad q \in A.$$

Thus an increasing event is more likely when p-values are made larger.

Definition 6.1 (PRDS for p-values). A p-value vector $P = (P_1, \dots, P_m)$ is positively regression dependent on a set of true nulls $\mathcal{I}_0$ if, for every increasing event $A \subseteq [0,1]^m$ and every $i \in \mathcal{I}_0$,

$$t \mapsto \mathbb{P}(P \in A \mid P_i = t)$$

is nondecreasing in t, using a regular conditional distribution.

The definition says that conditioning a true-null p-value to be less stringent should not make increasing events less likely. In the BH proof, the key increasing events are events such as "the procedure makes at most k rejections." Indeed, if $p \leq q$ coordinatewise, every ordered p-value of q is at least the corresponding ordered p-value of p. A BH crossing that disappears after increasing p-values cannot be replaced by a new crossing, so the rejection count can only decrease. Therefore $\{R \leq k\}$ is increasing in the p-value coordinates.

Theorem 6.2 (Gaussian sufficient condition for PRDS). *Let $Z = (Z_1, \dots, Z_m)$ be multivariate normal with means μ_i and covariance Σ, and let*

$$P_i = 1 - \Phi(Z_i)$$

be the right-tail p-value for testing $H_i : \mu_i = 0$ against positive alternatives, so $\mu_i = 0$ for true nulls. If

$$\mathrm{Cov}(Z_i, Z_j) \geq 0 \qquad \text{for every } i \in \mathcal{I}_0,\ j \in \{1, \dots, m\},$$

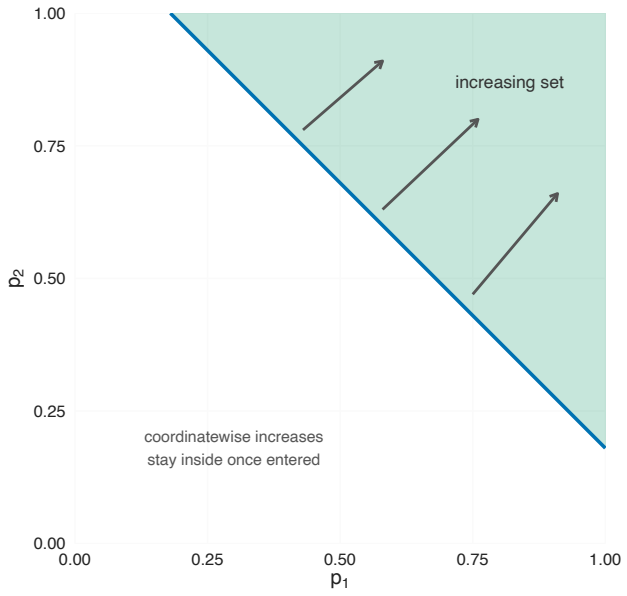

FIGURE 1. A schematic increasing set in p-value coordinates. Once a point is inside the shaded set, making coordinates larger keeps it inside. PRDS is a positive-dependence condition stated in terms of these increasing events.

then the p-value vector $P = (P_1, \dots, P_m)$ is PRDS on $\mathcal{I}_0$ in the sense of Definition above.

PROOF. Fix $i \in \mathcal{I}_0$ and let A be an increasing event in p-value coordinates. The corresponding event in Z-coordinates is *decreasing* in Z: increasing the Z_j's decreases the P_j's, so an event that gets more likely as p-values grow gets less likely as Z-scores grow. Write this Z-event as

$$B = \{z : \ (1 - \Phi(z_1), \dots, 1 - \Phi(z_m)) \in A\}.$$

If $z' \le z$ coordinatewise and $z \in B$, then the corresponding p-value vector at z' is at least the one at z, so z' also lies in B. Equivalently, B is coordinatewise decreasing.

The conditional distribution of Z_{-i} given $Z_i = z$ is Gaussian with covariance not depending on z and mean

$$\mathbb{E}[Z_{-i} \mid Z_i = z] = \mu_{-i} + \Sigma_{-i,i}\Sigma_{i,i}^{-1}(z - \mu_i).$$

Under the covariance assumption, $\Sigma_{-i,i}$ is entrywise nonnegative, so this conditional mean is coordinatewise nondecreasing in z. To see the monotonicity explicitly, take $z_1 < z_2$ and represent the two conditional vectors as

$$Z_{-i}^{(r)} = m(z_r) + \varepsilon, \qquad r = 1, 2,$$

using the same centered Gaussian noise ε, where $m(z) = \mu_{-i} + \Sigma_{-i,i}\Sigma_{i,i}^{-1}(z - \mu_i)$. Then $m(z_1) \le m(z_2)$ coordinatewise. Since B is decreasing,

$$\mathbf{1}\{(z_2, m(z_2) + \varepsilon) \in B\} \le \mathbf{1}\{(z_1, m(z_1) + \varepsilon) \in B\}$$

after inserting the fixed ith coordinate into the same position in both vectors. This holds for every value of the shared noise, and hence $\mathbb{P}(P \in A \mid Z_i = z)$ is nonincreasing in z.

Translating back to p-values reverses the direction once more. The map $P_i = 1 - \Phi(Z_i)$ is strictly decreasing, so conditioning on $P_i = t$ is the same as conditioning on $Z_i = \Phi^{-1}(1 - t)$, and this value of Z_i decreases as t increases. Therefore a conditional probability that is nonincreasing

in Z_i becomes nondecreasing in t:

$$t \mapsto \mathbb{P}(P \in A \mid P_i = t)$$

is nondecreasing, which is the PRDS condition for the p-value vector. □

The sign manipulation above is the source of many mistakes in informal descriptions of PRDS. The two coordinate systems are equivalent, but monotone-decreasing transformations such as $1 - \Phi(\cdot)$ reverse the relevant partial order, and the proof must state the direction unambiguously.

Remark 6.3 (MTP2, PRDS, and Gaussian signs). Multivariate total positivity of order two (MTP2) is a stronger positive- dependence condition. A density f on a product lattice is MTP2 if

$$f(x \vee y)f(x \wedge y) \geq f(x)f(y) \qquad \text{for all } x, y,$$

where $\vee$ and $\wedge$ are coordinatewise maximum and minimum. MTP2 implies positive association and, in the appropriate coordinate order, PRDS [70, 103]. For a nonsingular Gaussian vector, MTP2 is equivalent to the precision matrix Σ^{-1} having nonpositive off-diagonal entries. This is stronger than pairwise nonnegative covariance in dimensions larger than two. Theorem 6.2 used a direct one-sided Gaussian sufficient condition; it should not be read as an MTP2 characterization.

Lemma 6.4 (PRDS conditioning inequality). *If P is PRDS on the true-null set $\mathcal{I}_0$, then for every increasing event $A \subseteq [0,1]^m$ and every $i \in \mathcal{I}_0$,*

$$t \mapsto \mathbb{P}(P \in A \mid P_i \leq t)$$

is nondecreasing.

Proof. Let $h(s) = \mathbb{P}(P \in A \mid P_i = s)$, which is nondecreasing by PRDS, and let F_i denote the marginal CDF of P_i. Then

$$\mathbb{P}(P \in A \mid P_i \leq t) = \frac{1}{F_i(t)} \int_0^t h(s)\, dF_i(s).$$

For $t' > t$,

$$\mathbb{P}(P \in A \mid P_i \leq t') = \frac{F_i(t)}{F_i(t')}\mathbb{P}(P \in A \mid P_i \leq t) + \frac{F_i(t') - F_i(t)}{F_i(t')}\mathbb{P}(P \in A \mid t < P_i \leq t').$$

The right-hand side is a convex combination of $\mathbb{P}(P \in A \mid P_i \leq t)$ and an average of h over $(t, t']$. Since h is nondecreasing, the second average is at least the first, so the convex combination is at least $\mathbb{P}(P \in A \mid P_i \leq t)$. Therefore $\mathbb{P}(P \in A \mid P_i \leq t)$ is nondecreasing in t. □

PRDS is fragile under coordinatewise transformations. If T_i is a strictly increasing transformation of P_i, then T is PRDS on $\mathcal{I}_0$ in the same direction; but if T_i is strictly decreasing, the partial order reverses, and the PRDS property must be restated for the new order. This is exactly the point that required care in the Gaussian sufficient condition above: the right-tail p-value is a decreasing function of Z_i, so the sign of the covariance $\Sigma_{-i,i}$ interacts with the direction of the PRDS statement.

Theorem 6.5 (BH under PRDS). *If the p-value vector is PRDS on the set of true nulls and true-null p-values are continuous uniform, then the BH procedure at level α satisfies*

$$\mathrm{FDR} \leq \frac{m_0}{m}\alpha \leq \alpha.$$

Proof. Let R be the number of BH rejections and fix a true null i. It is enough to show

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \leq \alpha R/m\}}{R \vee 1}\right] \leq \frac{\alpha}{m}.$$

Decompose by the value of R:

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le \alpha R/m\}}{R \vee 1}\right] = \sum_{k=1}^{m} \frac{1}{k}\mathbb{P}\left(P_i \le \frac{\alpha k}{m}, R = k\right)$$
$$= \frac{\alpha}{m}\sum_{k=1}^{m} \mathbb{P}\left(R = k \,\middle|\, P_i \le \frac{\alpha k}{m}\right),$$

where uniformity of the true-null p-value gives $\mathbb{P}(P_i \le \alpha k/m) = \alpha k/m$ and converts the unconditional probability to a conditional one. The event $\{R \le k\}$ is increasing in the p-value coordinates: BH compares $p_{(j)}$ with $\alpha j/m$, so making p-values larger only weakens crossings, hence R can only decrease. Therefore $\{R \le k\}$ is preserved under coordinatewise increase, and Lemma 6.4 applies. Writing each term as a difference of $\{R \le k\}$-probabilities,

$$\sum_{k=1}^{m} \mathbb{P}\left(R = k \,\middle|\, P_i \le \frac{\alpha k}{m}\right)$$
$$= \mathbb{P}(R \le m \mid P_i \le \alpha) - \mathbb{P}(R \le 0 \mid P_i \le \alpha/m)$$
$$+ \sum_{k=1}^{m-1}\left[\mathbb{P}\left(R \le k \,\middle|\, P_i \le \frac{\alpha k}{m}\right) - \mathbb{P}\left(R \le k \,\middle|\, P_i \le \frac{\alpha(k+1)}{m}\right)\right].$$

The first term is at most 1. The second is zero: conditional on $P_i \le \alpha/m$, the p-value P_i itself satisfies the smallest BH threshold $p_{(1)} \le \alpha/m$, so the BH procedure rejects at least hypothesis i, giving $R \ge 1$ deterministically. Each bracket in the final sum is nonpositive: by Lemma 6.4, the conditional probability $\mathbb{P}(R \le k \mid P_i \le t)$ is nondecreasing in t, so increasing t from $\alpha k/m$ to $\alpha(k+1)/m$ only increases the right-hand side. Hence the total is at most 1, and summing the bound α/m over the m_0 true nulls proves the claim. □

PRDS is not arbitrary dependence. Negative or adversarial dependence can break the proof. This is why BY remains useful as a safe fallback and why dependence modeling deserves attention.

Remark 6.6 (Two-sided Gaussian p-values). For two-sided p-values

$$P_i = 2\{1 - \Phi(|Z_i|)\}$$

from a Gaussian vector with arbitrary covariance, the one-sided PRDS argument above no longer applies directly: the map $Z_i \mapsto |Z_i|$ is not coordinatewise monotone. Under the global null, $V = R$, so FDR control at level α is equivalent to the two-sided Gaussian Simes bound

$$\mathbb{P}\left(\exists k : \; P_{(k)} \le \frac{k\alpha}{m}\right) \le \alpha.$$

The case $m = 2$ is known, and several positive-dependence or MTP2 subclasses are covered by existing comparison arguments. In dimensions $m \ge 3$, an arbitrary Gaussian correlation matrix can have negative-dependence patterns that fall outside PRDS, MTP2, and the usual Gaussian correlation inequality machinery. The full arbitrary-covariance statement remains a conjecture, so the book treats it as an open problem.

2. Empirical Processes and Reverse Martingales

Threshold procedures can be studied as empirical-process stopping rules. For a p-value threshold t, write

$$V(t) = \#\{i \in \mathcal{I}_0 : P_i \le t\}, \qquad R(t) = \#\{i : P_i \le t\}.$$

BH chooses the largest t for which

$$\frac{mt}{R(t) \vee 1} \leq \alpha.$$

Under independent true nulls, the process $V(t)$ is a binomial counting process with mean $m_0 t$. The ratio $V(t)/t$ has a reverse-martingale structure when time runs from 1 down to 0. One convenient reverse filtration is

$$\mathcal{F}_t = \sigma(\{\mathbf{1}(P_i \leq u) : i \in \mathcal{I}_0,\ u \geq t\}),$$

which records the configuration of true-null p-values in the already revealed upper part $[t, 1]$, together with the number $V(t)$ that remain below t. With respect to this decreasing-time filtration,

$$\mathbb{E}\left[\frac{V(s)}{s} \,\middle|\, \mathcal{F}_t\right] = \frac{V(t)}{t}, \qquad 0 < s < t \leq 1,$$

because the true-null p-values not yet revealed, those lying in $[0, t]$, are conditionally exchangeable and uniformly scattered in that interval [114].

For a first reading, it is enough to check the same calculation with only $V(t)$ conditioned on. If $V(t) = v$, then the v true-null p-values that fell in $[0, t]$ are conditionally iid uniform on $[0, t]$. Therefore, for $s < t$,

$$V(s) \mid V(t) = v \sim \text{Bin}\left(v, \frac{s}{t}\right), \qquad \mathbb{E}\left[\frac{V(s)}{s} \,\middle|\, V(t) = v\right] = \frac{v}{t}.$$

The full filtration version says that this identity remains true after also revealing which true-null p-values lie above t, and after conditioning on nonnull p-values when they are independent of the nulls. This is the technical reason Storey-style thresholds can be stopped before λ without losing the null-count calibration.

This viewpoint unifies several FDR proofs. The selected BH threshold is a stopping time for the reverse filtration, and optional stopping controls the expected value of the null count divided by its null expectation. Mirror estimators, Storey estimators, and other empirical-process procedures use the same idea with different estimates of the null process.

3. Estimating the Null Fraction

BH uses m as a conservative upper bound on the number of true nulls. If many hypotheses are nonnull, this can be conservative. Storey's procedure estimates the null fraction $\pi_0 = m_0/m$ from the right side of the p-value histogram. For a tuning parameter $\lambda \in [0, 1)$, define

$$\widehat{\pi}_0^{\lambda} = \frac{1 + \#\{i : P_i > \lambda\}}{(1 - \lambda)m}.$$

The plus one is a finite-sample stabilizer. The idea is that alternatives tend to produce small p-values, so p-values above λ are enriched for nulls.

Given $\widehat{\pi}_0^{\lambda}$, an adaptive BH threshold replaces m by $m\widehat{\pi}_0^{\lambda}$:

$$\widehat{T} = \sup\left\{t \leq \lambda : \frac{m\widehat{\pi}_0^{\lambda} t}{R(t) \vee 1} \leq \alpha\right\}.$$

If the set is empty, set $\widehat{T} = 0$ and make no rejections. The restriction $t \leq \lambda$ is important in finite-sample proofs. The right-tail estimate $\widehat{\pi}_0^{\lambda}$ is formed from p-values above λ, and the threshold search then moves downward through $[0, \lambda]$. Thus the estimate is fixed before the reverse-time martingale is stopped. Under independence, the proof uses the binomial identity for the number of true-null p-values above λ:

$$V_0(\lambda) = \#\{i \in \mathcal{I}_0 : P_i > \lambda\} \sim \text{Bin}(m_0, 1 - \lambda).$$

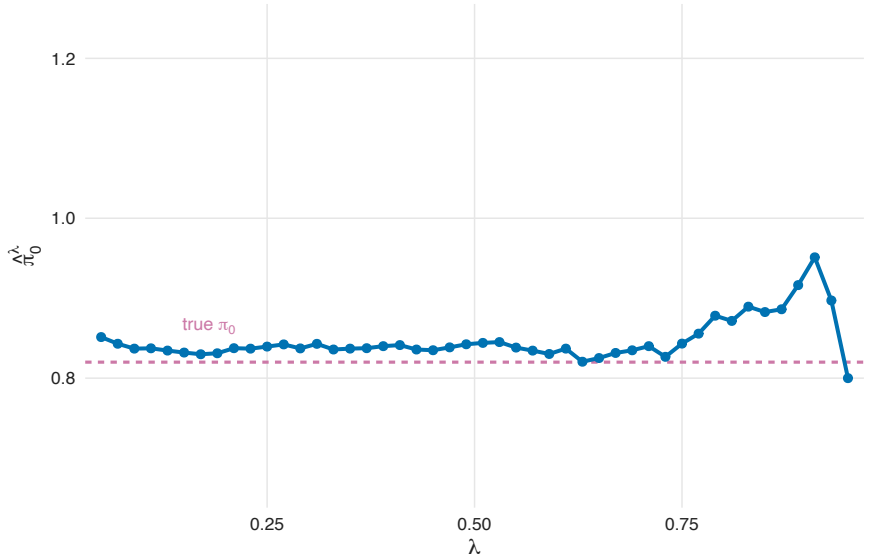

FIGURE 2. Storey's estimate $\widehat{\pi}_0^\lambda$ as λ varies in one simulated mixture. Small λ can be biased by alternatives; large λ can be noisy because few p-values remain in the tail.

The reciprocal moments of $1 + V_0(\lambda)$ are what make the plus-one version conservative.

THEOREM 6.7 (Storey's finite-sample bound, simplified). *Assume the true-null p-values are independent uniforms and are independent of the nonnull p-values. For fixed $\lambda \in (0,1)$, Storey's procedure above satisfies*

$$\mathrm{FDR} \le \alpha.$$

PROOF. Let $V(t) = \#\{i \in \mathcal{I}_0 : P_i \le t\}$. The value λ is fixed in advance, and the threshold search is restricted to $0 < t \le \lambda$. Conditional on the nonnull p-values and on which true-null p-values lie above λ, the remaining true-null p-values in $[0,\lambda]$ are iid uniforms on $[0,\lambda]$. Call this conditioning the *tail information at* λ. After this information is fixed, $\widehat{\pi}_0^\lambda$ is fixed, and the only remaining randomness relevant to $V(t)$ comes from the unobserved null p-values below λ.

In the reverse filtration that reveals the lower tail as t decreases, $V(t)/t$ is therefore a reverse martingale. The event $\{\widehat{T} \ge t\}$ can be decided from the p-values at thresholds $u \in [t,\lambda]$, together with the fixed tail information at λ, so $\widehat{T}$ is a reverse stopping time bounded by λ. Optional stopping gives

$$\mathbb{E}\left[\mathbf{1}\{\widehat{T} > 0\}\frac{V(\widehat{T})}{\widehat{T}}\,\middle|\,\text{tail information at } \lambda\right] = \frac{V(\lambda)}{\lambda}.$$

Thus

$$\begin{aligned}
\mathrm{FDR} &= \mathbb{E}\left[\frac{V(\widehat{T})}{R(\widehat{T}) \vee 1}\right] \\
&\le \frac{\alpha}{m}\mathbb{E}\left[\mathbf{1}\{\widehat{T} > 0\}\frac{V(\widehat{T})}{\widehat{\pi}_0^\lambda \widehat{T}}\right] \\
&= \frac{\alpha}{m}\mathbb{E}\left[\frac{m(1-\lambda)}{1 + \#\{j : P_j > \lambda\}}\frac{V(\lambda)}{\lambda}\right].
\end{aligned}$$

Since the denominator includes all p-values above λ,

$$1 + \#\{j : P_j > \lambda\} \geq 1 + m_0 - V(\lambda).$$

Here $m_0 - V(\lambda)$ is exactly the number of true-null p-values above λ. With $X = V(\lambda) \sim \mathrm{Bin}(m_0, \lambda)$, a direct calculation gives

$$\mathbb{E}\left[\frac{X}{1 + m_0 - X}\right] = \frac{\lambda(1 - \lambda^{m_0})}{1 - \lambda}.$$

Indeed, if $X \sim \mathrm{Bin}(n, p)$ and $Y = n - X \sim \mathrm{Bin}(n, 1 - p)$, then

$$\frac{X}{1 + n - X} = \frac{n - Y}{1 + Y} = \frac{n + 1}{1 + Y} - 1,$$

and

$$\mathbb{E}\left[\frac{1}{1 + Y}\right] = \sum_{y=0}^{n} \frac{1}{1 + y} \binom{n}{y} (1 - p)^y p^{n-y} = \frac{1 - p^{n+1}}{(n + 1)(1 - p)}.$$

Substituting $n = m_0$, $p = \lambda$, and $X = V(\lambda)$ gives the displayed reciprocal moment. Substitution yields

$$\mathrm{FDR} \leq \alpha(1 - \lambda^{m_0}) \leq \alpha.$$

□

Storey's estimate is not automatically valid under arbitrary dependence or arbitrary data-dependent choices of λ. The PRDS theorem for ordinary BH does not automatically transfer to Storey's adaptive denominator: dependence can make the right-tail estimate too small exactly on samples where the lower tail contains many null p-values. Positive dependence can therefore lead to FDR inflation for adaptive Storey procedures unless the dependence and adaptivity assumptions are controlled. The more adaptive the procedure, the more carefully its randomness must be separated from the null p-values it will test.

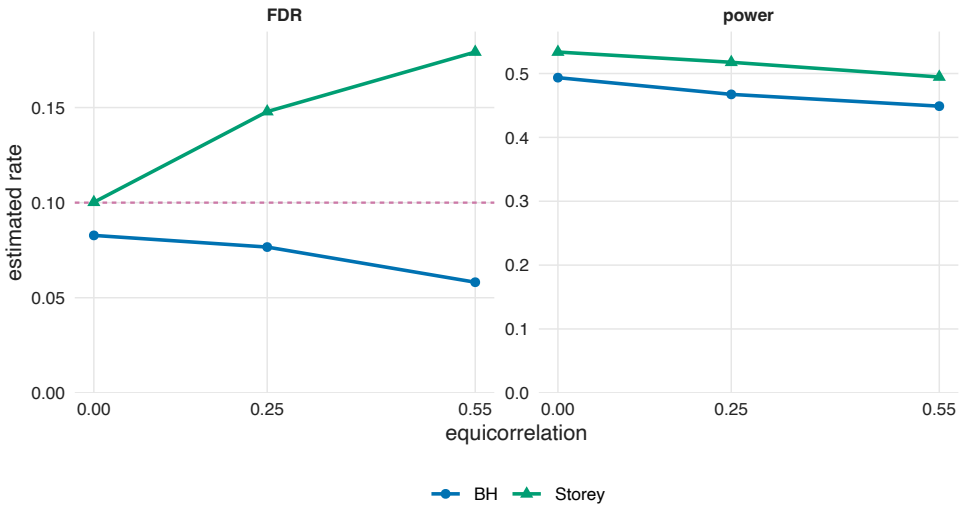

FIGURE 3. BH and a simple Storey adaptive procedure under equicorrelated one-sided Gaussian tests. The adaptive procedure gains power when the estimator is reliable, but this simulation also shows why dependence and adaptivity require explicit assumptions.

4. Q-Values

The q-value of a hypothesis is the smallest FDR level at which that hypothesis would be rejected. In practice, q-values are often computed by applying a Storey-type estimate of π_0 to BH-style adjusted p-values. If $\widehat{\pi}_0$ is fixed, the ordered q-values are

$$q_{(i)} = \min_{j \geq i} \left\{ \frac{\widehat{\pi}_0 m p_{(j)}}{j} \right\} \wedge 1.$$

As with BH-adjusted p-values, the running minimum ensures monotonicity across ordered ranks.

Q-values are tail-area quantities: they summarize the error rate for a rejection set containing all hypotheses at least as significant as the current one. For a one-sided statistic, the q-value attached to z asks about a region such as $\{Z \geq z\}$, not about the single point $Z = z$. This is why q-values are closer to pFDR for nested rejection regions than to local posterior probabilities for individual hypotheses. The distinction becomes clear in the two-groups model.

5. The Two-Groups Model and Local FDR

The empirical-Bayes view starts by putting each hypothesis into a latent state. Here Z_i denotes the scalar evidence statistic for hypothesis i; in a one-sided Gaussian testing problem, larger Z_i is stronger evidence against the null. Let $\theta_i = 0$ denote a null and $\theta_i = 1$ a nonnull. The two-groups model assumes

$$\mathbb{P}(\theta_i = 0) = \pi_0, \qquad Z_i \mid \theta_i = 0 \sim f_0, \qquad Z_i \mid \theta_i = 1 \sim f_1.$$

The marginal density is

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z).$$

The local false discovery rate is the posterior null probability

$$\operatorname{lfdr}(z) = \mathbb{P}(\theta = 0 \mid Z = z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1 - \pi_0) f_1(z)}.$$

Local FDR is a pointwise quantity. A q-value is a tail-area quantity. In a one-sided problem, the tail-area FDR for rejecting $Z \geq z$ is

$$\mathbb{P}(\theta = 0 \mid Z \geq z).$$

This averages local FDR values over the whole rejection tail:

$$\mathbb{P}(\theta = 0 \mid Z \geq z) = \mathbb{E}\{\operatorname{lfdr}(Z) \mid Z \geq z\}.$$

Here is the calculation. Let $I = \mathbf{1}\{\theta = 0\}$. Since $\operatorname{lfdr}(Z) = \mathbb{P}(\theta = 0 \mid Z) = \mathbb{E}[I \mid Z]$, conditioning further on the tail event and applying iterated expectation gives

$$\begin{aligned} \mathbb{P}(\theta = 0 \mid Z \geq z) &= \mathbb{E}[I \mid Z \geq z] \\ &= \mathbb{E}\{\mathbb{E}[I \mid Z] \mid Z \geq z\} \\ &= \mathbb{E}\{\operatorname{lfdr}(Z) \mid Z \geq z\}. \end{aligned}$$

Equivalently, the numerator of the tail probability is $\int_z^\infty \pi_0 f_0(u)\, du$, while averaging local FDR over the same tail gives

$$\frac{\int_z^\infty \operatorname{lfdr}(u) f(u)\, du}{\int_z^\infty f(u)\, du} = \frac{\int_z^\infty \pi_0 f_0(u)\, du}{\int_z^\infty f(u)\, du}.$$

Thus $\operatorname{lfdr}(z)$ is the posterior null probability for a case observed near z, while the tail-area quantity is the average posterior null probability among all cases at least as extreme as z. A q-value reports the smallest tail error level at which an observation would enter a nested rejection rule. The two numbers can differ substantially, especially when the mixture density changes rapidly.

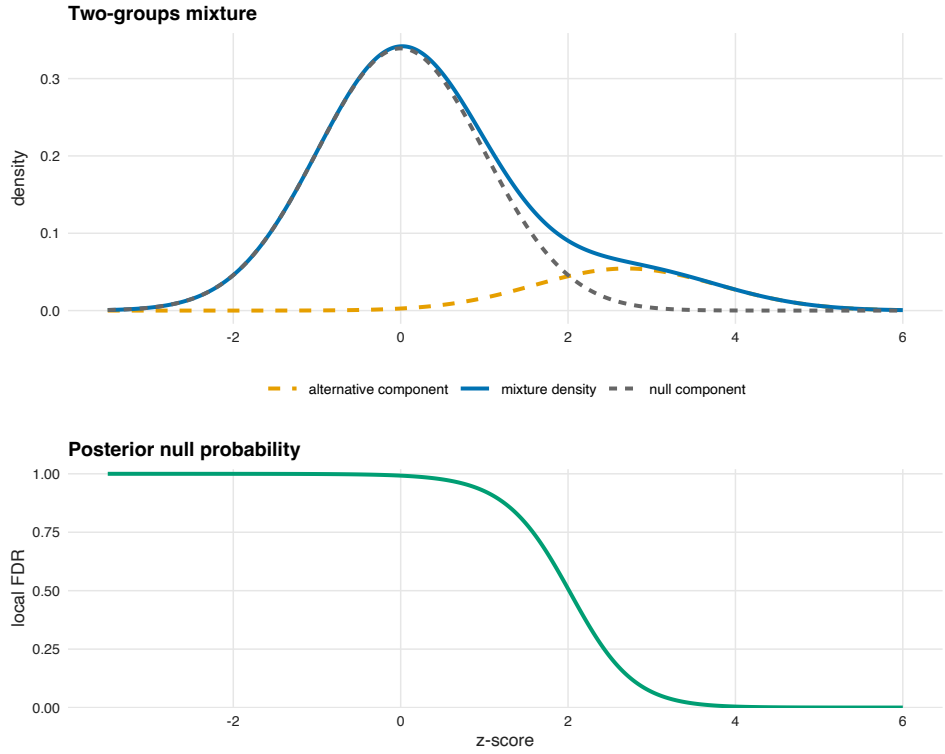

FIGURE 4. Two-groups model. The top panel shows null, alternative, and mixture densities. The bottom panel shows local FDR, the posterior probability that a case with z-score z is null.

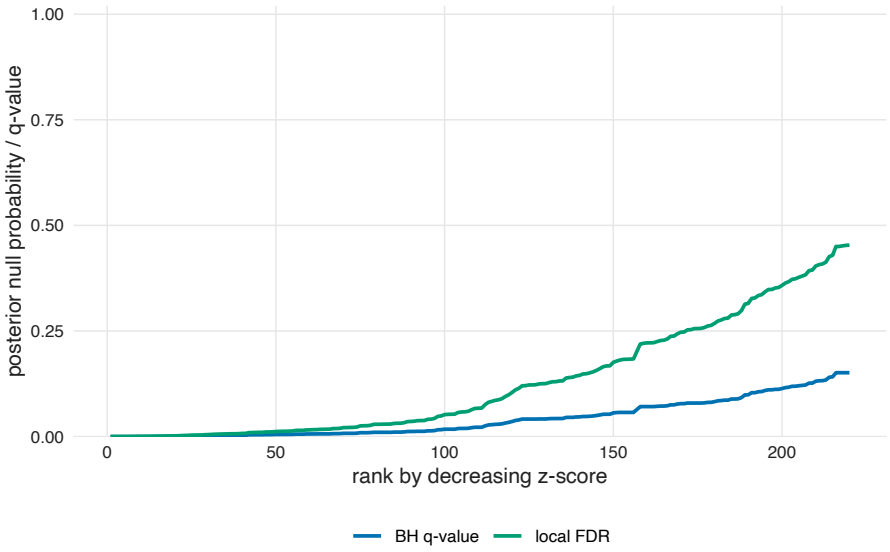

FIGURE 5. Local FDR and BH q-values for the most significant simulated z-scores. Local FDR ranks individual cases by posterior null probability, while q-values summarize the tail set that would be reported at each rank.

6. Optimal Thresholding by Local FDR

The two-groups model also gives an optimization principle. This section treats the no-side-information setting: all hypotheses share the same null fraction π_0, null density f_0, and alternative density f_1. Suppose a rule rejects observations in a region Γ. For m independent draws from the two-groups model, the expected numbers of false and true discoveries are

$$\mathrm{EFP}(\Gamma) = m\pi_0 \int_\Gamma f_0(z)\,dz, \qquad \mathrm{ETP}(\Gamma) = m(1-\pi_0)\int_\Gamma f_1(z)\,dz.$$

The corresponding marginal false discovery rate is

$$\mathrm{mFDR}(\Gamma) = \frac{\mathrm{EFP}(\Gamma)}{\mathrm{EFP}(\Gamma)+\mathrm{ETP}(\Gamma)} = \frac{\mathbb{E}[V]}{\mathbb{E}[R]},$$

when $\mathbb{E}[R] > 0$. This ratio is not the same object as the frequentist FDR $\mathbb{E}[V/(R \vee 1)]$, but it is often the natural oracle objective in empirical-Bayes decision theory.

This objective can be written directly in terms of local FDR. Let

$$f(z) = \pi_0 f_0(z) + (1-\pi_0) f_1(z), \qquad \ell(z) = \mathrm{lfdr}(z) = \frac{\pi_0 f_0(z)}{f(z)}.$$

Then

$$\mathrm{EFP}(\Gamma) = m\int_\Gamma \ell(z) f(z)\,dz, \qquad \mathbb{E}[R(\Gamma)] = m\int_\Gamma f(z)\,dz,$$

so $\mathrm{mFDR}(\Gamma)$ is the average value of $\ell(z)$ over the rejected region, with respect to the mixture density f. Thus a region containing mostly small local-FDR values spends less false-discovery budget per rejection.

The following oracle statement is the continuous version of this sorting principle. To include boundary randomization, let a procedure be a measurable function $\delta(z) \in [0,1]$, where $\delta(z)$ is the probability of rejecting an observation with statistic z. A deterministic rejection region Γ is the special case $\delta(z) = \mathbf{1}\{z \in \Gamma\}$. For such a rule,

$$\mathrm{mFDR}(\delta) = \frac{\int \delta(z)\ell(z)f(z)\,dz}{\int \delta(z) f(z)\,dz},$$

when the denominator is positive.

Proposition 6.8 (Oracle local-FDR level sets). *Among all oracle procedures δ satisfying $\mathrm{mFDR}(\delta) \le \alpha$, a procedure maximizing expected true discoveries can be chosen to reject a lower level set of local FDR:*

$$\delta_c(z) = \mathbf{1}\{\ell(z) < c\} + \rho(z)\mathbf{1}\{\ell(z) = c\},$$

for some c and some boundary randomization $0 \le \rho(z) \le 1$. If $\ell(Z)$ has no atom at the boundary, the rule is simply $\Gamma_c = \{z : \ell(z) \le c\}$, with c chosen as the largest threshold for which

$$\frac{\int_{\{\ell(z)\le c\}} \ell(z) f(z)\,dz}{\int_{\{\ell(z)\le c\}} f(z)\,dz} \le \alpha.$$

Thus the optimal oracle rejection boundary is a level surface of local FDR.

Proof. Write integrals with respect to the mixture measure $d\nu(z) = f(z)\,dz$. For a rule δ, define its rejection mass and false-discovery mass by

$$r(\delta) = \int \delta(z)\,d\nu(z), \qquad q(\delta) = \int \delta(z)\ell(z)\,d\nu(z).$$

Then $\mathrm{mFDR}(\delta) = q(\delta)/r(\delta)$ when $r(\delta) > 0$. If no positive-mass rule is feasible, the no-rejection rule is optimal and the claim is trivial. Hence assume below that the optimizer has positive rejection mass. The mFDR constraint is therefore

$$q(\delta) - \alpha r(\delta) \le 0,$$

and the expected true-discovery mass is

$$\int \delta(z)\{1 - \ell(z)\}\, d\nu(z) = r(\delta) - q(\delta).$$

The Lagrange calculation makes the shape of the optimum transparent. Attach a multiplier $\eta \ge 0$ to the constraint and consider the penalized objective

$$\begin{aligned}\mathcal{L}_\eta(\delta) &= \{r(\delta) - q(\delta)\} - \eta\{q(\delta) - \alpha r(\delta)\} \\ &= \int \delta(z)\left[(1 - \ell(z)) - \eta\{\ell(z) - \alpha\}\right] d\nu(z) \\ &= \int \delta(z)\left[1 + \eta\alpha - (1+\eta)\ell(z)\right] d\nu(z).\end{aligned}$$

For a fixed η, this integral is maximized pointwise: set $\delta(z) = 1$ where the coefficient of $\delta(z)$ is positive, set $\delta(z) = 0$ where it is negative, and allow randomization where it is zero. Thus any rule selected by this constrained optimization has the form

$$\delta(z) = 1\{\ell(z) < c_\eta\} + \rho(z)1\{\ell(z) = c_\eta\}, \qquad c_\eta = \frac{1 + \eta\alpha}{1 + \eta}.$$

All nontrivial boundary points therefore have the same local-FDR value. This is the no-side-information version of the equal-boundary condition: the rejection boundary is a level surface of $\ell(z)$.

The exchange argument below is the self-contained justification that the Lagrange form loses no generality. Fix the rejection mass $r(\delta) = r > 0$. Among rules with this same mass, maximizing expected true discoveries is the same as minimizing $q(\delta)$. If a rule puts positive rejection probability on a higher-local-FDR set B while not fully rejecting a lower-local-FDR set A, take smaller subsets if necessary so the available ν-masses match, and move a small amount of rejection probability from B to A. The rejection mass r is unchanged, but q decreases because the moved mass is charged a smaller value of ℓ. Hence $r - q$ increases. Repeating this exchange forces the rule, up to ties, to fill the smallest local-FDR values first, with randomization only on a boundary set $\{\ell = c\}$.

Now take any mFDR-feasible rule and replace it by this lower-level-set rule with the same rejection mass. The replacement cannot increase q/r and cannot decrease $r - q$, so an optimizer may be chosen from the nested family of local-FDR level sets. Within that family, adding points in increasing order of ℓ increases expected true discoveries until adding more points would violate the mFDR constraint, or until all points have been included. If $\ell(Z)$ has no boundary atom, this gives the displayed largest feasible threshold; if there is a boundary atom, ρ randomizes on that atom to obtain the same boundary level. □

The finite-sample posterior version of the same argument treats local FDR as the cost of rejecting one observed case. Conditional on the observed Z_i's, define

$$\ell_i = \mathrm{lfdr}(Z_i) = \mathbb{P}(\theta_i = 0 \mid Z_i).$$

For a rejection set A, the posterior expected number of false discoveries is

$$\mathbb{E}[V(A) \mid Z_1, \ldots, Z_m] = \sum_{i \in A} \ell_i,$$

and the posterior plug-in mFDR of A is

$$\widehat{\mathrm{mFDR}}(A) = \frac{\sum_{i\in A}\ell_i}{|A| \vee 1}.$$

Thus a discovery with smaller ℓ_i has smaller posterior false-discovery cost. The optimal oracle rule should spend its false-discovery budget on the smallest local-FDR values first.

PROPOSITION 6.9 (Oracle local-FDR step-up rule). *Suppose the oracle local FDR values $\ell_i = \mathrm{lfdr}(Z_i)$ are known, and order them as $\ell_{(1)} \le \ell_{(2)} \le \cdots \le \ell_{(m)}$. Among rejection sets of size k that minimize the posterior expected number of false discoveries, the size-k set rejects the k smallest ℓ_i's; among all oracle prefix rules satisfying $\widehat{\mathrm{mFDR}} \le \alpha$, the one with the most rejections is*

$$\widehat{k} = \max\left\{1 \le k \le m : \frac{1}{k}\sum_{j=1}^{k}\ell_{(j)} \le \alpha\right\},$$

with the convention $\widehat{k} = 0$ if the set is empty. It rejects the $\widehat{k}$ hypotheses with the smallest local-FDR values and has posterior plug-in mFDR at most α.

PROOF. Given the observed Z_i's, ℓ_i is the posterior probability that hypothesis i is null. Therefore the posterior expected number of false discoveries in a rejection set A is $\sum_{i\in A}\ell_i$, while the number of rejections is $|A|$. For any fixed size k, suppose A contains an index a with $\ell_a > \ell_b$ while omitting an index b. Swapping a out and b in changes the posterior expected false count by $\ell_b - \ell_a < 0$ and leaves the size equal to k. Repeating this exchange proves that the unique minimal set, up to ties, consists of the k smallest ℓ_i's.

Among these nested optimal size-k sets, the posterior plug-in mFDR is the prefix average $k^{-1}\sum_{j=1}^{k}\ell_{(j)}$. The displayed rule chooses the largest prefix whose average is at most α, so no other oracle prefix rule can make more rejections while satisfying the same plug-in mFDR constraint. □

In practice the ℓ_i are replaced by estimates $\widehat{\ell}_i$ from a fitted mixture model. The proposition then describes the *oracle* optimum; the realized procedure inherits the model errors of the mixture fit.

This explains why empirical-Bayes ranking can differ from p-value ranking. If the alternative density is asymmetric, heteroscedastic, or multimodal, the most promising observations are those with small posterior null probability, not necessarily those with the smallest classical p-values.

In practice, π_0, f_0, and f_1 must be estimated. Common approaches include maximum likelihood, EM algorithms, central matching for the empirical null, and flexible mixture modeling. These methods are powerful but model-dependent. Their output should be read as posterior-style evidence under the fitted model, not as a distribution-free finite-sample FDR guarantee.

7. Covariate-Assisted Local FDR

The previous section assumed that all hypotheses share the same mixture weights. Side information lets those weights, or the alternative distribution, depend on observed covariates. Let $X_i \in \mathcal{X}$ be a covariate observed before the testing decision, and let P_i denote the random p-value with observed realization p_i. A covariate-assisted p-value mixture model has the form

$$\theta_i \mid X_i = x \sim \mathrm{Bernoulli}(1 - \pi_0(x)), \qquad P_i \mid X_i = x, \theta_i = 0 \sim \mathrm{Unif}[0,1],$$

and

$$P_i \mid X_i = x, \theta_i = 1 \sim f_1(\cdot \mid x).$$

The covariate is useful when either the conditional null fraction $\pi_0(x)$ or the conditional alternative density $f_1(\cdot \mid x)$ varies with x. For example, in an omics-wide screen, genes or CpG sites with

higher read depth, stronger baseline expression, known functional annotation, or corroborating evidence from a previous GWAS may have a different prior chance of being nonnull or a different power curve. Bayes' rule gives the covariate local FDR

$$\mathrm{lfdr}(p, x) = \mathbb{P}(\theta_i = 0 \mid P_i = p, X_i = x) = \frac{\pi_0(x)}{\pi_0(x) + (1 - \pi_0(x)) f_1(p \mid x)}.$$

Thus the same p-value can carry different posterior evidence at different covariate values: a moderate p-value may be more compelling in a covariate stratum with high power or low null fraction.

PROPOSITION 6.10 (Oracle covariate local-FDR thresholding). *Suppose the oracle local-FDR values at the observed data, $\ell_i = \mathrm{lfdr}(p_i, X_i)$, are known, and order them as $\ell_{(1)} \le \cdots \le \ell_{(m)}$. Define*

$$\widehat{k} = \max\left\{ 1 \le k \le m : \frac{1}{k} \sum_{j=1}^{k} \ell_{(j)} \le \alpha \right\},$$

with $\widehat{k} = 0$ if the set is empty, and reject the $\widehat{k}$ hypotheses with smallest ℓ_i's. Conditional on the observed data $(p_i, X_i)_{i=1}^m$, the posterior expected FDP of this rejection set is at most α. Among rejection sets of any fixed size, the set with the smallest posterior expected number of false discoveries rejects the smallest local-FDR values. Among these nested oracle sets, the displayed rule makes the most rejections subject to the posterior average-local-FDR constraint.

PROOF. Given the observed data (p_i, X_i), ℓ_i is the posterior probability that hypothesis i is null. If a rejection set A is chosen after seeing the data, its rejection count is fixed at $R(A) = |A|$, while

$$V(A) = \sum_{i \in A} \mathbf{1}\{\theta_i = 0\}.$$

Taking posterior expectation conditional on the observed data gives

$$\mathbb{E}[V(A) \mid (p_i, X_i)_{i=1}^m] = \sum_{i \in A} \mathbb{P}(\theta_i = 0 \mid P_i = p_i, X_i) = \sum_{i \in A} \ell_i.$$

Therefore the posterior expected FDP, with the usual convention that the FDP is zero when A is empty, is

$$\mathbb{E}\left[\frac{V(A)}{|A| \vee 1} \,\middle|\, (p_i, X_i)_{i=1}^m \right] = \frac{1}{|A| \vee 1} \sum_{i \in A} \ell_i.$$

For fixed $|A| = k$, the denominator is fixed, so minimizing the posterior expected FDP is the same as minimizing $\sum_{i \in A} \ell_i$. If A contains an index a with $\ell_a > \ell_b$ and omits b, replacing a by b lowers the sum by $\ell_a - \ell_b$. Repeating this exchange yields the set of the k smallest local-FDR values, up to ties.

For that optimal size-k set, the posterior expected FDP is $k^{-1} \sum_{j=1}^{k} \ell_{(j)}$. The step-up rule scans these nested optimal sets and chooses the largest prefix whose average is at most α. Thus its posterior expected FDP is at most α, and no other oracle prefix rule satisfying the same average-local-FDR constraint can make more rejections. □

This is a model-based oracle statement. If $\pi_0(x)$ and $f_1(p \mid x)$ are correctly specified and known, it describes the best posterior ranking rule. In practice they are estimated, for example by smoothing $\pi_0(x)$ over the covariate axis and fitting flexible conditional alternative densities. Then the guarantee is no longer automatic. The usual justification is asymptotic: as the number of hypotheses grows, the fitted local-FDR scores must be accurate enough that the empirical average false-discovery cost is close to the oracle one. This is different from a distribution-free,

finite-sample FDR proof. When finite-sample validity is needed, the model should be paired with a separate device that prevents overfitting the null p-values.

8. AdaPT: Local-FDR-Guided Masking

AdaPT adds such a validity device. It combines two ideas: a model, often local-FDR-like, proposes powerful covariate-dependent threshold surfaces; a masking and mirror argument prevents the adaptive search from using the null signs it is trying to test. In this section P_i denotes a random p-value, and p_i denotes its observed value.

For each observed p-value define the masked value and sign

$$\widetilde{p}_i = \min(p_i, 1 - p_i), \qquad s_i = \mathbf{1}\{p_i \le 1/2\}.$$

The masked value tells us how close p_i is to either tail, but it hides which tail contains the point. Under a true null with $P_i \mid X_i \sim \mathrm{Unif}[0,1]$, the random sign $\mathbf{1}\{P_i \le 1/2\}$ is a fair coin conditional on $(\min(P_i, 1-P_i), X_i)$. AdaPT lets the analyst or algorithm update a threshold surface $s_t : \mathcal{X} \to [0, 1/2]$ using the covariates, masked p-values, and signs that have already been revealed, but not the hidden signs of hypotheses still eligible for rejection.

For a fixed threshold surface s, the lower-tail rejection count and upper-tail mirror count are

$$R(s) = \#\{i : p_i \le s(X_i)\}, \qquad B(s) = \#\{i : p_i \ge 1 - s(X_i)\}.$$

The two sets use the same covariate-dependent width $s(X_i)$. The lower tail contains candidate discoveries. The upper tail contains mirror cases that are not rejected but, under a null, are as likely as lower-tail cases after conditioning on the masked value. AdaPT uses the mirror FDP estimate

$$\widehat{\mathrm{FDP}}_{\mathrm{AdaPT}}(s) = \frac{1 + B(s)}{R(s) \vee 1}.$$

The $+1$ is the same finite-sample stabilizer as in mirror procedures. If $s(x) \equiv t$ is constant, this reduces to the no-covariate mirror estimate from Chapter 5. AdaPT is more than the no-covariate mirror rule, however: the threshold surface can change with x. A local-FDR model can suggest larger thresholds in covariate regions where discoveries look more credible, while the masking rule controls which sign information the adaptive search is allowed to see.

Proposition 6.11 (Conditional symmetry under masking). *For a true-null hypothesis with $P \mid X \sim \mathrm{Unif}(0,1)$, the sign $\mathbf{1}\{P \le 1/2\}$ is Bernoulli$(1/2)$ and conditionally independent of $(\min(P, 1-P), X)$.*

Proof. Conditional on $X = x$, the null p-value is uniform. The map

$$P \mapsto (\min(P, 1-P),\ \mathbf{1}\{P \le 1/2\})$$

sends the uniform distribution on $[0,1]$ to the product of a uniform distribution on $[0,1/2]$ and a Bernoulli$(1/2)$ sign. Since this holds for every x, the sign is also conditionally independent of X. □

Theorem 6.12 (AdaPT FDR control, informal simplified statement). *Assume the true-null p-values are conditionally independent uniforms given their covariates, and that the threshold sequence s_t is adapted to the filtration generated by the covariates, masked p-values, and previously revealed signs. In particular, while a hypothesis is still eligible for rejection, the update rule has not seen whether its p-value lies in the lower or upper tail. If AdaPT stops at a threshold surface $\widehat{s}$ satisfying*

$$\frac{1 + \#\{i : p_i \ge 1 - \widehat{s}(X_i)\}}{\#\{i : p_i \le \widehat{s}(X_i)\} \vee 1} \le \alpha,$$

then, under the regularity conditions of Lei and Fithian [79], the rejection set $\{i : p_i \le \widehat{s}(X_i)\}$ *controls FDR at level* α.

Proof idea. Among null hypotheses whose masked values place them inside the current two-sided band $\widetilde{p}_i \le s_t(X_i)$, the hidden signs are fair coins. The lower side contributes false discoveries, while the upper side contributes mirror decoys. The key point is not that the threshold surface is fixed in advance; it may be chosen adaptively. The key point is that, for still-hidden hypotheses, the choice of the next surface is based only on information that is independent of their null signs. Conditional on the visible information, the unrevealed null signs remain independent fair coins. This makes the upper-tail mirror count a conservative proxy for the number of lower-tail nulls, with the $+1$ term providing finite-sample stabilization. The formal proof packages this comparison into a supermartingale and applies optional stopping when the mirror estimate falls below α. The local-FDR model is used for power, by proposing promising surfaces $s_t(x)$; the masking and mirror symmetry are what supply the finite-sample validity argument.

This separation is important. A poor local-FDR model can make AdaPT weak or inefficient, but the masking proof can still protect FDR when its assumptions hold. Conversely, if the adaptive model is allowed to peek at the hidden signs of candidate discoveries, it can steer the threshold toward lower-tail nulls without paying for the corresponding upper-tail mirrors, and the mirror comparison is no longer valid.

9. Positive FDR

The positive false discovery rate is

$$\mathrm{pFDR} = \mathbb{E}\left[\frac{V}{R}\,\middle|\, R > 0\right].$$

Since $\mathrm{FDP} = 0$ when $R = 0$,

$$\mathrm{FDR} = \mathrm{pFDR}\ \mathbb{P}(R > 0).$$

Under the two-groups model, if a rejection region is Γ, then pFDR has the posterior interpretation

$$\mathrm{pFDR}(\Gamma) = \mathbb{P}(\theta = 0 \mid Z \in \Gamma).$$

This identity is one reason q-values and empirical-Bayes FDR methods are often described as posterior error probabilities for rejection regions. The phrase "for rejection regions" is essential: pFDR and q-values average over all cases inside a selected set, whereas local FDR is the posterior null probability for a single observed case.

Proof. Condition on the event $R(\Gamma) = k > 0$. By exchangeability in the two-groups model, the expected number of nulls among the k selected observations is

$$k\,\mathbb{P}(\theta = 0 \mid Z \in \Gamma).$$

Thus

$$\mathbb{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\,\middle|\, R(\Gamma) = k\right] = \mathbb{P}(\theta = 0 \mid Z \in \Gamma),$$

and averaging over $k \ge 1$ gives the result. □

For a nested family of rejection regions $\{\Gamma_u : 0 < u < 1\}$, the q-value of an observed statistic z can be defined as

$$q(z) = \inf_{\Gamma_u : z \in \Gamma_u} \mathrm{pFDR}(\Gamma_u).$$

This is the smallest positive-FDR level at which z would enter the rejection region. In monotone one-sided models it coincides with the tail-area interpretation used earlier in this chapter.

10. Conditional Calibration

The final idea is that known dependence can sometimes be exploited rather than ignored. Conditional calibration and dependence-adjusted BH methods construct p-values or rejection thresholds using conditional distributions given dependence-relevant statistics. The goal is to get closer to BH power while retaining valid FDR control under a structured dependence model.

The basic move is the following. Suppose the joint null distribution of the p-value vector $P = (P_1, \ldots, P_m)$ has a sufficient or approximately sufficient statistic S that captures the dependence. Write p_i for the observed value of P_i, and write s_{obs} for the observed value of S. Conditional on $S = s_{\text{obs}}$, the true-null p-values may become exchangeable or may have a tractable conditional null distribution. Rather than running BH on the unconditional observed p-values p_i, one runs it on conditional null p-values

$$\tilde{p}_i = \mathbb{P}_0\Big(P_i^{\text{null}} \le p_i \,\Big|\, S = s_{\text{obs}}\Big),$$

where P_i^{null} denotes a draw from the conditional null law for hypothesis i. Because $\tilde{p}_i$ is the conditional null CDF evaluated at the observation, it is uniform in the continuous exact case and super-uniform with conservative randomization or discreteness. The price is computational: S must be modeled and the conditional distribution must be tractable. The dependence-adjusted BH procedure of Fithian and Lei [47] formalizes this idea for a broad class of dependence structures and shows that BH applied to the recalibrated p-values regains FDR control.

This topic is more advanced than the core PRDS and Storey material. The important message for now is that dependence creates a spectrum of choices: use BH when benign dependence conditions are justified, use BY when one wants robustness without structure, or model the dependence explicitly when the structure is scientifically meaningful and reliable.

11. Assumptions in Plain Language

PRDS is a positive-dependence condition, not a synonym for correlation. Its direction depends on whether the coordinates are evidence scores or p-values. BH remains valid under PRDS, while BY remains valid under arbitrary dependence at a power cost. Storey's estimator needs the right tail of the p-value distribution to behave like mostly null p-values and requires care under dependence or data-adaptive tuning. Local FDR and q-values are different: local FDR is pointwise and model-based; q-values summarize tail-area error for sets of discoveries.

12. Bibliographic Notes

The PRDS framework and BH validity under positive dependence are due to Benjamini and Yekutieli [16]; related Simes-type positive-dependence results include Sarkar [103]. Storey's estimator and q-values are developed in Storey [112], Storey [113], and Storey et al. [114]. The empirical-Bayes local-FDR perspective originates with Efron [42] and is treated comprehensively in the book Efron [43]. The compound-decision formulation and oracle/adaptive local-FDR thresholding are developed by Sun and Cai [118]. Covariate-assisted local-FDR rules and their oracle optimality theory are developed by Zhang and Chen [135] and Cao et al. [31]. Related structured empirical-Bayes approaches, including local-index ideas for dependent and spatial settings, are represented by Sun et al. [119]. AdaPT is due to Lei and Fithian [79], and the broader masking/interactive FDR framework is developed by Lei et al. [80]. General FDR reshaping and dependence ideas are discussed by Blanchard and Roquain [21]. Conditional calibration and dependence-adjusted BH are represented by Fithian and Lei [47]. A frontier direction with growing practical relevance under regulatory constraints such as GDPR and the EU AI Act is *differentially private* FDR control: the analyst adds calibrated Laplace or Gaussian noise to the log-p-values before selection, trading a small multiplicative inflation of the FDR

bound for (ε, δ)-differential privacy of the released rejection set. Dwork et al. [41] develops the theoretical framework.

13. Exercises

Basic.

Exercise 6.13 (Increasing sets). Draw two increasing and two non-increasing subsets of $[0,1]^2$ in p-value coordinates. Explain which direction is relevant for the PRDS definition used in this chapter.

Exercise 6.14 (Local FDR). Derive

$$\mathrm{lfdr}(z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1-\pi_0) f_1(z)}$$

from Bayes' rule.

Exercise 6.15 (pFDR identity). Show that

$$\mathrm{FDR} = \mathrm{pFDR}\ \mathbb{P}(R > 0).$$

Then, under the two-groups model, show that for a fixed rejection region Γ, $\mathrm{pFDR}(\Gamma) = \mathbb{P}(\theta = 0 \mid Z \in \Gamma)$.

Intermediate.

Exercise 6.16 (Gaussian PRDS direction). Let Z be a Gaussian vector with nonnegative covariances and suppose larger Z_i is stronger evidence. Explain why monotone transformations to right-tail p-values reverse the coordinate order. State the corresponding PRDS condition in p-value coordinates.

Exercise 6.17 (BH under PRDS). In the BH proof under PRDS, show that the event $\{R(t) \le k\}$ is increasing in the p-value vector. Explain how this replaces the independence step in the leave-one-out proof.

Exercise 6.18 (Storey's estimator). Derive Storey's estimator

$$\widehat{\pi}_0^{\lambda} = \frac{1 + \#\{P_i > \lambda\}}{(1-\lambda)m}.$$

Discuss the bias-variance tradeoff as λ changes.

Exercise 6.19 (BH adjusted p-values as q-values). The BH adjusted p-values from Chapter 5 and the Storey-adapted q-values

$$q_{(i)} = \min_{j \ge i} \widehat{\pi}_0 m p_{(j)}/j \wedge 1$$

agree when $\widehat{\pi}_0 = 1$. Show that for any fixed $\widehat{\pi}_0 < 1$, the Storey-adapted q-values are uniformly no larger than the BH-adjusted ones. Discuss the FDR price of using a poor estimate of π_0.

Exercise 6.20 (Oracle local-FDR level sets). Let $d\nu(z) = f(z)\,dz$ be the mixture measure and let $\ell(z) = \mathrm{lfdr}(z)$. For a randomized oracle rule $\delta(z) \in [0,1]$, define

$$r(\delta) = \int \delta(z)\, d\nu(z), \qquad q(\delta) = \int \delta(z)\ell(z)\, d\nu(z),$$

and verify that $\mathrm{mFDR}(\delta) = q(\delta)/r(\delta)$ when $r(\delta) > 0$. Form the Lagrangian for maximizing expected true discoveries subject to $\mathrm{mFDR}(\delta) \le \alpha$, and show that the pointwise optimizer rejects where $\ell(z) < c_\eta$. Explain why all nontrivial boundary points have the same local-FDR value, and how boundary randomization handles atoms of $\ell(Z)$.

EXERCISE 6.21 (Optimal covariate local-FDR thresholding). Let p_i be the observed value of P_i, and set $\ell_i = \mathrm{lfdr}(p_i, X_i)$. For a rejection set A, verify that the posterior expected number of false discoveries is $\sum_{i\in A} \ell_i$ and that the posterior expected FDP is

$$\frac{1}{|A| \vee 1} \sum_{i\in A} \ell_i.$$

Use an exchange argument to show that, among rejection sets of fixed size k, the set with smallest posterior expected number of false discoveries rejects the k smallest local-FDR values. Derive the oracle step-up rule that chooses the largest k for which

$$\frac{1}{k} \sum_{j=1}^{k} \ell_{(j)} \le \alpha,$$

with $k = 0$ if no nonempty prefix satisfies the inequality. Explain why this is a model-based posterior optimality statement rather than a finite-sample frequentist FDR proof.

Computational.

EXERCISE 6.22 (Storey under dependence). Reproduce Figure 3. Vary λ, the signal strength, and the correlation. Identify settings where the adaptive estimator is too aggressive.

EXERCISE 6.23 (Q-values and local FDR). Simulate $m = 10{,}000$ z-scores from the two-groups model

$$Z \sim \pi_0 N(0,1) + (1-\pi_0) f_1, \qquad \pi_0 = 0.9,$$

where

$$f_1(z) = 0.7\,\varphi(z - 2.5) + 0.3\,\varphi\{(z + 1.5)/2\}/2$$

is an asymmetric heavy-tailed alternative density. Use one-sided p-values $p = 1 - \Phi(Z)$. Compute BH q-values and the oracle local FDR

$$\mathrm{lfdr}(z) = \frac{\pi_0 \varphi(z)}{\pi_0 \varphi(z) + (1-\pi_0) f_1(z)}.$$

Identify specific ranks where the two rankings disagree. Explain why q-values summarize tail rejection sets, while local FDR ranks individual cases by posterior null probability.

EXERCISE 6.24 (Storey's π_0 reproducibility). Reproduce Figure 2. Then add a second curve showing the true π_0 and a third curve showing the resulting adaptive BH rejection count as a function of λ. Discuss a sensible default choice of λ for a screen with $m = 10{,}000$ and a presumed signal fraction of 2%.

EXERCISE 6.25 (Two-stage BKY). Let $q \in (0,1)$ and set $q_1 = q/(1+q)$. The two-stage procedure of Benjamini et al. [18] first runs BH at level q_1, obtaining R_1 rejections. If $R_1 = 0$ or $R_1 = m$, it stops. Otherwise it runs BH again at level

$$q_2 = q_1 \frac{m}{m - R_1},$$

equivalently using critical values $iq_1/(m - R_1)$. Implement this procedure and compare it with BH and Storey's procedure under independent p-values with $\pi_0 \in \{1, 0.9, 0.7\}$. Then repeat under positively correlated Gaussian z-scores. Report empirical FDR and power, and explain why this is an adaptive method for structured settings, with independence as the basic theorem, rather than an arbitrary-dependence replacement for BY.

Advanced.

EXERCISE 6.26 (Why masking suffices for AdaPT). Consider a single null hypothesis with $P \sim \mathrm{Unif}[0,1]$, masked value $\widetilde{P} = \min(P, 1-P)$, and sign $S = \mathbf{1}\{P \le 1/2\}$. Show that conditional on $\widetilde{P}$, the sign S is Bernoulli(1/2) and independent of any function of $\widetilde{P}$. For a covariate-dependent threshold $s(x)$, identify the lower-tail event $P \le s(X)$ and the upper-tail mirror event $P \ge 1 - s(X)$. Explain why an AdaPT update based only on covariates, masked p-values, and already revealed signs cannot peek at the hidden null signs. Then explain why fitting the next threshold using unmasked signs of still-eligible hypotheses would invalidate the upper-tail mirror comparison.

EXERCISE 6.27 (Reverse martingale). For independent uniform true-null p-values, prove the reverse-martingale property of $V(t)/t$ with respect to a decreasing threshold filtration.

EXERCISE 6.28 (Binomial reciprocal identity). Let $X \sim \mathrm{Bin}(n,p)$. Prove that

$$\mathbb{E}\left[\frac{X}{1+n-X}\right] = \frac{p(1-p^n)}{1-p}.$$

Hint. Substitute $Y = n - X \sim \mathrm{Bin}(n, 1-p)$, write $(n-Y)/(1+Y) = (n+1)/(1+Y) - 1$, and use the auxiliary identity $\mathbb{E}[1/(1+W)] = (1-(1-r)^{N+1})/((N+1)r)$ for $W \sim \mathrm{Bin}(N,r)$. Apply the resulting reciprocal moment to Storey's proof with $n = m_0$, $p = \lambda$.

EXERCISE 6.29 (Why two-sided Gaussian p-values need care). Let $Z \sim N(0,\Sigma)$ and define two-sided p-values

$$P_i = 2\{1 - \Phi(|Z_i|)\}.$$

Show that the rejection event $\{P_i \le t\}$ is not monotone in the original coordinate Z_i. Explain why this prevents the one-sided Gaussian PRDS argument in Theorem 6.2 from being applied directly to two-sided p-values, even though the two-sided Gaussian FDR conjecture is believed to hold in broad settings.

CHAPTER 7

Structured and Hierarchical Multiple Testing

The BH and BY procedures of Chapters 5 and 6 treat the hypotheses as a flat list. They sort the p-values, compare them with a common boundary, and rely on broad assumptions about dependence among those p-values. They do not use the scientific map that says which hypotheses belong together. Real families are rarely flat. Genomic studies test variants inside genes and genes inside pathways. Clinical trials separate primary endpoints from secondary and tertiary endpoints. Benchmark suites for AI models pool prompts inside tasks, tasks inside capability domains, and domains inside an overall evaluation. When this structure is part of the question, ignoring it can waste power and can report discoveries at the wrong resolution.

This chapter develops procedures that use such structure without losing FDR control. We treat four cases: (i) hypotheses grouped into disjoint families with one layer of aggregation (group BH), (ii) several partition layers tested simultaneously (p-filter), (iii) hypotheses arranged on a rooted tree with per-level FDR targets (TreeBH), and (iv) replicability across studies, where a discovery requires evidence in enough separate experiments (partial conjunction). A knockoff analogue also exists for structured feature selection; it is discussed after ordinary knockoffs are introduced in Chapter 9.

1. Why Flat BH Is Not Enough

Suppose the m hypotheses are partitioned into G groups $\mathcal{G}_1, \dots, \mathcal{G}_G$, where group g contains m_g hypotheses. Call group g a *null group* if every hypothesis in $\mathcal{G}_g$ is a true null, and a *rejected group* if at least one hypothesis in $\mathcal{G}_g$ is rejected. Two natural error rates are

$$\mathrm{FDR}_{\mathrm{flat}} = \mathbb{E}\left[\frac{V}{R \vee 1}\right],$$

where V is the number of false rejections across all hypotheses, and

$$\mathrm{FDR}_{\mathrm{group}} = \mathbb{E}\left[\frac{V^{\mathrm{group}}}{R^{\mathrm{group}} \vee 1}\right],$$

where V^{group} is the number of *rejected null groups* (i.e., null groups in which the procedure rejected at least one hypothesis) and R^{group} is the total number of rejected groups. Under this definition, a group that mixes true signal with one falsely rejected null is *not* counted as a false group discovery: the group's null hypothesis "all members are null" is genuinely false, so flagging the group is a true discovery at the group resolution.

These rates can differ dramatically. Consider a stylized genomic example: $G = 50$ genes, each with $m_g = 200$ tightly correlated single-nucleotide variants, so that the variants within a gene effectively measure the same underlying signal. Suppose five genes carry a real signal and BH at the variant level rejects 800 variants concentrated in those five real genes plus 40 variants spread across five null genes (where correlated noise produced small p-values). Then the realized flat FDP is $40/840 \approx 4.8\%$, which is within target. But the realized group-level FDP is $V^{\mathrm{group}}/R^{\mathrm{group}} = 5/10 = 50\%$: half the genes flagged are false positives at the gene resolution. A geneticist who is going to follow up at gene resolution needs the group-level guarantee, not the variant-level one.

REMARK 7.1. The flat and group rates are not directly comparable. A procedure controls *both* only if it accounts for the group structure explicitly. The methods of this chapter let the analyst declare the resolutions of interest in advance and obtain calibrated FDR at each one simultaneously.

A second issue is heterogeneity of group sizes and group signal density. A flat BH procedure pools all p-values into a single ordering, so groups with many small p-values dominate the rejection set even when those p-values reflect correlated noise within a single null group. Procedures aware of the group structure can balance rejections across groups or apply group-level filtering before within-group testing. Either approach restores the connection between "what the procedure rejects" and "what the analyst will act on."

2. Group BH via Simes Aggregation

The simplest group-level procedure aggregates each group into a single super-uniform p-value and applies BH to the group representatives. Given within-group p-values $p_{g,1}, \dots, p_{g,m_g}$, the Simes representative for group g is

$$p_g^{\text{Simes}} = \min_{1 \le k \le m_g} \frac{m_g \, p_{g,(k)}}{k},$$

where $p_{g,(k)}$ is the kth-smallest p-value in group g. Group BH then applies the ordinary BH procedure at level α to $p_1^{\text{Simes}}, \dots, p_G^{\text{Simes}}$ and rejects the corresponding groups. The plain-language idea is simple: a null group should not look significant unless at least one of its ordered p-values crosses the Simes line. Under independence, and under the usual positive-dependence conditions for Simes, that crossing probability is no larger than the nominal level. Thus the whole group can be represented by one valid group-level p-value.

THEOREM 7.2 (Simes super-uniformity inside a group). *Suppose group g is null, meaning every constituent $H_{g,i}$ is true. Assume either that $p_{g,1}, \dots, p_{g,m_g}$ are mutually independent and super-uniform, or more generally that their joint null law satisfies the Simes inequality*

$$\mathbb{P}\left(\min_{1 \le k \le m_g} \frac{m_g p_{g,(k)}}{k} \le t \right) \le t, \qquad 0 \le t \le 1.$$

For example, the latter condition holds for exact-valid PRDS p-values under the usual Simes assumptions. Then $\mathbb{P}(p_g^{\text{Simes}} \le t) \le t$ for every $t \in [0,1]$.

PROOF. First take all m_g p-values to be independent and exactly uniform. The classical Simes calculation, proved in Theorem 3.1, states that for independent uniform $U_1, \dots, U_{m_g}$ with order statistics $U_{(1)} \le \cdots \le U_{(m_g)}$,

$$\mathbb{P}\left(\bigcup_{k=1}^{m_g} \{ U_{(k)} \le tk/m_g \} \right) = t \qquad \text{for every } t \in [0,1],$$

and hence $\mathbb{P}(p_g^{\text{Simes}} \le t) = t$. For completeness, the induction proof is the same as in Theorem 3.1: split on the largest order statistic $U_{(m_g)}$, condition on $U_{(m_g)} = s$, and rescale the remaining $m_g - 1$ uniforms to $[0,1]$. The integral gives $t(1 - t^{m_g - 1}) + t^{m_g} = t$.

Now relax to independent super-uniform nulls. Since each $p_{g,i}$ stochastically dominates a uniform random variable, Strassen's monotone coupling theorem gives a pair $(p'_{g,i}, U_i)$ with $p'_{g,i} \overset{d}{=} p_{g,i}$, $U_i \sim \text{Unif}(0,1)$, and $p'_{g,i} \ge U_i$ almost surely. Applying this coupling independently for $i = 1, \dots, m_g$ preserves the product joint law of the p-values and makes $U_1, \dots, U_{m_g}$ independent uniforms. Coordinate-wise domination implies order-statistic domination, $p'_{g,(k)} \ge U_{(k)}$ for every k path-wise, and since the Simes statistic is non-decreasing in each input, $\min_k m_g p'_{g,(k)}/k \ge$

$\min_k m_g U_{(k)}/k$ path-wise. Because $(p'_{g,1}, \dots, p'_{g,m_g})$ has the same joint distribution as the original independent p-values, this gives $\mathbb{P}(p_g^{\text{Simes}} \le t) \le \mathbb{P}(\min_k m_g U_{(k)}/k \le t) = t$, as claimed. In the general case where the displayed Simes inequality is assumed directly, the conclusion is exactly that inequality rewritten in terms of p_g^{Simes}. The PRDS example is the positive-dependence Simes theorem of Sarkar [103]; see Chapter 6. □

The across-group condition in the next theorem is a condition on the group-level representatives, not merely on the original leaf-level p-values. Two safe situations are worth keeping in mind. If the groups are disjoint and the p-values in different groups are independent, then the Simes representatives are independent. More generally, if the group representatives themselves are one-sided Gaussian p-values whose underlying Gaussian vector has the positive-dependence structure of Theorem 6.2, then they are PRDS. What is *not* automatic is that a PRDS statement for all leaf-level p-values survives the nonlinear operation that maps each group to its Simes minimum.

THEOREM 7.3 (Group BH FDR control). *Suppose that*

(1) *within each true-null group $\mathcal{G}_g$, the p-values satisfy the assumptions of Theorem 7.2, and*
(2) *the Simes representatives $p_1^{\text{Simes}}, \dots, p_G^{\text{Simes}}$ are mutually independent (or PRDS) across groups.*

Then group BH at level α applied to the Simes representatives $p_1^{\text{Simes}}, \dots, p_G^{\text{Simes}}$ controls group FDR at level α:

$$\mathbb{E}\left[\frac{V^{group}}{R^{group} \vee 1}\right] \le \alpha.$$

PROOF. By Theorem 7.2, each p_g^{Simes} is super-uniform under its group null. Across-group independence (or PRDS) of the representatives is assumed. Applying Corollary 5.3 (or the PRDS variant from Chapter 6) to the family $\{p_g^{\text{Simes}}\}_{g=1}^G$ yields $\text{FDR}^{\text{group}} \le \alpha$. □

Interpretively, group BH differs from flat BH by changing the reported unit. A single strong signal in a group of size m_g can produce many small correlated within-group p-values, inflating the variant-level discovery count even when the scientific signal lives at the gene level. Group BH collapses these into one calibrated representative, so the output is a group-level list rather than a long list of correlated within-group hits. If within-group correlations are strong enough to violate the Simes assumptions, one needs a Bonferroni-type aggregation (or a BY-reshaped Simes) inside the group before invoking Theorem 7.3.

EXAMPLE 7.4 (Two genes by hand). Let $G = 2$ groups have within-group p-values

$$(p_{1,1}, p_{1,2}, p_{1,3}) = (0.028, 0.028, 0.028), \qquad (p_{2,1}, p_{2,2}, p_{2,3}) = (0.031, 0.40, 0.50).$$

The Simes representatives are

$$p_1^{\text{Simes}} = \min\left(\frac{3 \cdot 0.028}{1}, \frac{3 \cdot 0.028}{2}, \frac{3 \cdot 0.028}{3}\right) = 0.028,$$
$$p_2^{\text{Simes}} = \min\left(\frac{3 \cdot 0.031}{1}, \frac{3 \cdot 0.40}{2}, \frac{3 \cdot 0.50}{3}\right) = 0.093.$$

Group BH at $\alpha = 0.05$ compares the sorted representatives $(0.028, 0.093)$ with the thresholds $\alpha k/G = 0.025k$ for $k = 1, 2$, namely $(0.025, 0.05)$. Because $0.028 > 0.025$ and $0.093 > 0.05$, group BH rejects no group.

Compare with flat BH on the six pooled p-values, sorted as

$$0.028,\ 0.028,\ 0.028,\ 0.031,\ 0.40,\ 0.50,$$

against thresholds $\alpha k/6 = 0.00833k$ for $k = 1, \ldots, 6$,

$$(0.0083,\ 0.0167,\ 0.025,\ 0.0333,\ 0.0417,\ 0.05).$$

The step-up rule of BH rejects the largest k for which $p_{(k)} \le \alpha k/m$. Here $p_{(4)} = 0.031 \le 0.0333$, so flat BH rejects all four of the smallest p-values: the three from gene 1 *and* the first p-value of gene 2. Group BH does not: it has refused both groups, including the one for which flat BH would have rejected a (likely null) variant. The price of group-level interpretability is this conservatism: group BH refuses to rebrand a single gene's correlated evidence as multiple within-gene discoveries, and refuses to leak a variant-level rejection into a gene that fails its group-level test.

3. The p-Filter

Group BH controls FDR at a single layer at a time. Multi-resolution problems require simultaneous FDR control at several layers. The *p-filter* of Barber and Ramdas [7] solves this.

Setup. Let the hypotheses $\{1, \ldots, m\}$ carry L partition layers; layer ℓ is described by a many-to-one map $\pi_\ell : \{1, \ldots, m\} \to \{1, \ldots, G_\ell\}$ that sends each hypothesis to its group at layer ℓ. Layer $\ell = 1$ is typically the finest (often $G_1 = m$ with one hypothesis per "group"); deeper layers are coarser. The layers need not be nested; the p-filter applies to multiple partitions of the hypotheses, with nested hierarchies as an important special case. Call a layer-ℓ group g a *null group* if all of its constituent hypotheses are true nulls. For a rejection set $R \subseteq \{1, \ldots, m\}$, define

$$R^{(\ell)} = |\pi_\ell(R)|, \qquad V^{(\ell)} = |\{g \in \pi_\ell(R) : g \text{ is a null group}\}|,$$

that is, the number of layer-ℓ groups containing at least one rejection, and the number of those that are null groups. The p-filter takes layer levels $\alpha_1, \ldots, \alpha_L$ and returns a rejection set $\widehat{R}$ such that $\mathrm{FDR}^{(\ell)} = \mathbb{E}[V^{(\ell)}/(R^{(\ell)} \vee 1)] \le \alpha_\ell$ simultaneously for every ℓ.

The iterative algorithm. For each layer ℓ and each group g at that layer, let $p_g^{(\ell)}$ denote a layer-ℓ group p-value – in practice, a super-uniform aggregator such as Simes applied to the within-group p-values (at $\ell = 1$ with $G_1 = m$, one has $p_h^{(1)} = p_h$ for each hypothesis h). Given a threshold vector $\boldsymbol{t} = (t_1, \ldots, t_L)$, the joint hypothesis-level rejection set and its layer-ℓ image are

$$\widehat{R}(\boldsymbol{t}) = \{i : p^{(\ell)}_{\pi_\ell(i)} \le t_\ell \text{ for every } \ell\}, \qquad \widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}) = \pi_\ell(\widehat{R}(\boldsymbol{t})). \tag{2}$$

The p-filter of Barber and Ramdas [7] can be described through count-space coordinates. For a proposed vector $\boldsymbol{k} = (k_1, \ldots, k_L)$, set

$$t_\ell(k_\ell) = \frac{\alpha_\ell k_\ell}{G_\ell}, \qquad k_\ell \in \{0, 1, \ldots, G_\ell\}.$$

The algorithm chooses the coordinate-wise largest feasible $\boldsymbol{k}$ such that

$$\#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}(\boldsymbol{k})) \ge k_\ell \qquad \text{for every layer } \ell.$$

Equivalently, at the returned threshold vector $\boldsymbol{t}$, the BH-style boundary

$$\frac{G_\ell\, t_\ell}{\#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}) \vee 1} \ \le\ \alpha_\ell \tag{3}$$

holds at every layer ℓ, and $\boldsymbol{t}$ is the largest such vector on the finite BH ladders. A coordinate-wise reduction implementation is:

(1) Initialize $k_\ell = G_\ell$, equivalently $t_\ell = \alpha_\ell$, for every layer.
(2) Given the current $\boldsymbol{t}$, compute $\widehat{R}(\boldsymbol{t})$ and the layer counts $r_\ell = \#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t})$.
(3) Replace k_ℓ by $\min(k_\ell, r_\ell)$ and t_ℓ by $\alpha_\ell k_\ell / G_\ell$ for every layer, then repeat until $\boldsymbol{k}$ stabilizes.

The iteration is coordinate-wise monotone: each k_ℓ can only decrease, and each coordinate takes values in the finite set $\{0, 1, \dots, G_\ell\}$. Hence the reduction iteration terminates in a finite number of steps at a self-consistent fixed point; convergence and uniqueness of the maximal feasible point are established in Barber and Ramdas [7].

Validity. The theorem below is quoted as a result from Barber and Ramdas [7]. A simple sufficient case is mutual independence of all base p-values, with true-null base p-values super-uniform, and group p-values formed by Simes or weighted Simes as in that paper. More general positive-dependence cases are possible, but they are assumptions on the full collection of null group p-values across layers, not automatic consequences of having valid p-values within each group.

THEOREM 7.5 (p-filter FDR control, cited result [7]). *Under the assumptions of Barber and Ramdas [7] (in particular, the layer-ℓ group p-values $\{p_g^{(\ell)}\}$ are valid super-uniform combinations of the within-group p-values, and the dependence structure of the null group p-values satisfies their joint independence-or-PRDS condition), the p-filter output at levels $\alpha_1, \dots, \alpha_L$ satisfies $\mathrm{FDR}^{(\ell)} \le \alpha_\ell$ for every ℓ.*

Proof idea. At a single layer ($L = 1$), the boundary (3) is exactly the BH self-consistency characterization: the largest t for which $G\,t/\#\widehat{\mathcal{R}} \le \alpha$ and $\widehat{\mathcal{R}} = \{g : p_g \le t\}$ is the BH threshold. The BH leave-one-out proof of Chapter 5 then applies verbatim and gives $\mathrm{FDR} \le \alpha$. For multiple layers, Barber and Ramdas [7] show, via a per-layer leave-one-out argument that mirrors the single-layer case but accounts for the cross-layer coupling through the shared hypothesis-level rejection set, that the aggregate contribution of null groups at layer ℓ to $\mathrm{FDR}^{(\ell)}$ is bounded by α_ℓ; their detailed bookkeeping under within-layer PRDS is the main technical content of that paper, and we refer the reader to it for the full argument.

In the special case $L = 1$ with $G_1 = m$ (the trivial layering at the hypothesis level), the boundary (3) reduces to BH at level α_1 applied to $p_1, \dots, p_m$. In the special case $L = 1$ with coarser groups and Simes aggregation, the layer-1 p-values are $p_g^{(1)} = p_g^{\mathrm{Simes}}$ and the p-filter reduces to the group BH of Section 2. The new content of the p-filter is that the same iteration controls FDR at multiple resolutions simultaneously without inflating the per-layer levels.

4. Trees of Hypotheses and TreeBH

A different and more permissive structure is a rooted tree $\mathcal{T}$. Let the leaves of $\mathcal{T}$ be the testable hypotheses $\{H_1, \dots, H_m\}$ and let each internal node v carry the intersection hypothesis

$$H_v = \bigcap_{i:\, i \preceq v} H_i,$$

where $i \preceq v$ means leaf i is a descendant of v. A rejection set R is *ancestor-closed* (or "prefixed") if, whenever a leaf is rejected, all of its ancestors up to the root are also rejected. Ancestor-closure is the natural constraint for hierarchical procedures because it makes scientific sense: one cannot claim a gene-level discovery without also claiming the containing pathway as a discovery, since the pathway null "no leaf in the pathway is non-null" is implied to be false by the very rejection of one of its descendant genes.

Algorithm. The TreeBH procedure of Bogomolov et al. [23] implements a top-down rejection rule. Let $C(v)$ denote the children of node v, and write ρ for the formal root. We index depth from $d = 1$ at the topmost *tested* layer (the children of ρ) down to $d = D$ at the leaves; ρ itself is only a container, not a tested family, so it does not consume a target FDR level. This is the

opposite direction from the p-filter notation above: here depth 1 is the coarsest tested layer and depth D is the leaf layer. Each tested depth carries its own target FDR level α_d for $d = 1, \dots, D$.

TreeBH uses valid p-values throughout the tree. Leaves carry their original p-values. For an internal node w, define its p-value by a valid combination rule applied to the p-values of its immediate children, typically Simes; recursively, this supplies valid p-values for all internal nodes under the TreeBH dependence assumptions. Using Simes over all descendant leaves is a valid simplified aggregation in some settings, but the formal TreeBH algorithm is naturally stated through immediate-child p-values.

For a selected parent v, write $\widehat{S}(v) \subseteq C(v)$ for the children rejected by BH within that child family. The formal root ρ is selected by convention, so its children form the first tested family.

(1) At depth 1, apply BH to the p-values of the children of ρ at level α_1; call the rejected children $\widehat{S}(\rho)$.
(2) At depth $d+1$, for every selected node v at depth d, apply BH to the p-values of $C(v)$ at the path-specific level

$$q_v^{(d+1)} = \alpha_{d+1} \prod_{r=1}^{d} \frac{|\widehat{S}(v_{r-1})|}{|C(v_{r-1})|},$$

where $v_0 = \rho, v_1, \dots, v_d = v$ is the ancestor chain of v. Reject the children selected by that BH run and call the set $\widehat{S}(v)$.
(3) Recurse on rejected children until no further rejections occur or the leaves are reached.

The product of selected proportions along the ancestor chain is the key design choice. It concentrates the available α_{d+1} budget into branches that survived earlier screens. For a two-level hierarchy this product reduces to the familiar selected-family factor $R^{(1)}/G_1$; in deeper trees it is path-specific, not a single global ratio at the previous depth.

Validity. For a two-level hierarchy, the selected-family target is the average FDP among the selected parent families. For deeper trees, TreeBH controls the recursive selected-family FDR, denoted $s\,\mathrm{FDR}^{\ell}$, rather than a flat average over all selected parents at the immediately previous depth.

Fix a target level ℓ. For a selected node u at depth ℓ, let

$$A_u^{(\ell)} = \mathbf{1}\{H_u \text{ is a true null}\}.$$

For a selected node v at depth $k < \ell$, define recursively

$$A_v^{(\ell)} = \frac{1}{|\widehat{S}(v)| \vee 1} \sum_{u \in \widehat{S}(v)} A_u^{(\ell)}.$$

Finally set

$$s\,\mathrm{FDR}^{\ell} \;=\; \mathbb{E}[A_\rho^{(\ell)}]. \tag{4}$$

The quantity $A_v^{(\ell)}$ is the false-discovery fraction passed upward from the selected descendants of v at target depth ℓ. At the formal root ρ, it is the recursively averaged FDP for the whole selected tree at depth ℓ. At $\ell = 2$, this reduces to the usual average of within-family FDPs over selected depth-1 families. At $\ell = 3$, it first averages over selected children within each selected depth-1 family, then averages those quantities over selected depth-1 families. This recursive weighting differs from pooling all selected depth-2 parents together when selected branches have different numbers of selected children.

The displayed algorithm is a pedagogical summary of the procedure in Bogomolov et al. [23]. The exact path-specific levels, valid tree p-values, the paper's dependence assumptions, and the simple-selection-rule condition are all part of the cited theorem. The simple-selection-rule

condition is substantive: selected child families must come from the prescribed BH-on-valid-child-p-values step so that the selected-family adjustment applies.

THEOREM 7.6 (TreeBH recursive selected-family FDR control, cited result). *Apply the TreeBH procedure of Bogomolov et al. [23] with the exact path-specific levels above, valid tree p-values, the paper's dependence assumptions, and the required simple selection rule at each tested family. Then the TreeBH output satisfies*

$$s\,\mathrm{FDR}^{\ell} \le \alpha_{\ell}, \qquad \ell = 1, \dots, D.$$

Why the rescaling has this form. Two levels suffice to see the idea: a root whose children are groups, whose own children are leaves. At depth $d = 1$, BH applied to the group-level Simes representatives at level α_1 gives $\mathrm{FDR}^{(1)} \le \alpha_1$ by Theorems 7.2–7.3.

For depth $d + 1 = 2$, the standard mistake is to condition on "group g was selected at depth 1": the selection event $\{p_g^{\mathrm{Simes}} \le \alpha_1 R^{(1)}/G_1\}$ is a function of the very leaf p-values inside g, so within-g conditional super-uniformity is not available. The α_1 in this event is correct: it is the level used to select top-level groups. The later BH run inside each selected group uses the depth-2 level $\alpha_2 R^{(1)}/G_1$. Instead we work unconditionally, noting that the within-g FDP contributes to $s\,\mathrm{FDR}^2$ only when g is selected:

$$s\,\mathrm{FDR}^2 \;=\; \mathbb{E}\left[\frac{1}{R^{(1)} \vee 1} \sum_g \mathbf{1}\left\{g \in \widehat{\mathcal{R}}^{(1)}\right\} \cdot \frac{V_g^{(2)}}{R_g^{(2)} \vee 1}\right].$$

The selected-family theorem of Bogomolov et al. [23] – of which TreeBH is the prototypical instance – shows that testing each selected family at the rescaled level $\alpha_2 R^{(1)}/G_1$ controls the selected-family average in (4) at α_2. The factor $R^{(1)}/G_1$ is what compensates for choosing the family by looking at its own data: only a fraction $R^{(1)}/G_1$ of top-level families are selected, so the within-family level is reduced by that selected fraction. In deeper trees the same adjustment is applied recursively along the realized ancestor chain. A selected gene inside a selected pathway therefore inherits the pathway-level selected fraction and the selected-gene fraction inside that pathway; replacing this product by a single global previous-depth ratio is not the TreeBH rule.

REMARK 7.7. TreeBH's per-level target (4) is a recursive selected-family FDR, not the unconditional "overall" FDR $\mathbb{E}[V^{(d+1)}/(R^{(d+1)} \vee 1)]$. Bounding the overall FDR at depth $d + 1$ directly would require simultaneous control across all rejected ancestors at depth d, and a naive per-rejected-parent BH does not deliver this without additional adjustment. "BH inside each rejected family" is the right tool for the recursive selected-family target, which is the natural multi-level error rate in genomics, neuroimaging, and benchmark hierarchies.

EXAMPLE 7.8 (Three-level genomic hierarchy). Suppose 20,000 single-variant p-values are nested inside 800 gene-level hypotheses, which are nested inside 50 pathway-level hypotheses, with target levels $(\alpha_1, \alpha_2, \alpha_3) = (0.05, 0.10, 0.10)$ for pathway, gene, and variant FDR. At depth 1, BH is applied to the 50 pathway-level Simes representatives at level $\alpha_1 = 0.05$; suppose this rejects $R^{(1)} = 6$ pathways. At depth 2, within each rejected pathway, BH is applied to the gene-level Simes representatives at the local rescaled BH level $q_2 = \alpha_2\, R^{(1)}/G_1 = 0.10 \cdot 6/50 = 0.012$ (here q_d denotes the per-family rescaled level used inside the BH test of the depth-d children of a rejected parent). Suppose this rejects genes inside those pathways. If a selected pathway P has $R_{\mathrm{genes}}(P)$ rejected genes among $G_{\mathrm{genes}}(P)$ tested genes, then at depth 3, within each rejected gene in pathway P, BH is applied to the variant p-values at the path-specific level

$$q_{3,P} = \alpha_3 \cdot \frac{6}{50} \cdot \frac{R_{\mathrm{genes}}(P)}{G_{\mathrm{genes}}(P)}.$$

The variant threshold is therefore pathway-specific. The global ratio $0.10 \cdot R_{\text{genes,total}}/800$, for example $0.10 \cdot 80/800$ if 80 genes are rejected in total across all pathways, is not the TreeBH formula because it omits the top-level selected-pathway factor and replaces the selected-parent-specific gene ratio by a pooled count. TreeBH controls the recursive selected-family FDR $s\,\text{FDR}^{\ell}$ at each depth, not the ordinary depth-level FDR $\mathbb{E}[V^{(\ell)}/(R^{(\ell)} \vee 1)]$.

The trade-off with flat BH is power for interpretability. TreeBH may gain or lose leaf-level power relative to flat BH, depending on how signal is distributed across branches and how strongly the parent levels screen the search. Its main advantage is interpretability: each rejected leaf has a structurally certified parent at every level, so a rejected variant lies in a rejected gene in a rejected pathway, and the chain of rejections is itself the report.

Structured versions of the knockoff filter address a related multi-resolution feature-selection problem using knockoff statistics rather than p-values. We defer that topic to Chapter 9, Section 10, after the ordinary knockoff construction has been introduced.

5. Replicability and Partial Conjunction

The discussion so far concerns multi-resolution error rates within a single study. A different structural problem appears when the analyst wants to report only findings that replicate across two or more separate studies.

Definition 7.9 (Replicability). Let $H_{i,k}$, $k = 1, \dots, K$, be the null hypothesis for feature i in study k. Feature i is *r-out-of-K replicated* if at least r of the per-study nulls $H_{i,1}, \dots, H_{i,K}$ are false.

The *partial conjunction null* negates replicability. Writing $\mathcal{N}_i$ for the set of true-null per-study nulls of feature i,

$$H_i^{(r/K)} : \quad |\mathcal{N}_i| \geq K - r + 1,$$

that is, at most $r - 1$ of the per-study nulls $H_{i,1}, \dots, H_{i,K}$ are false. A finding is replicated if and only if this null is false.

The naive "r-th smallest p-value" statistic is not super-uniform. For instance, in the any-of-two case $r = 1$, $K = 2$ (one wants at least one of the two studies to detect a signal), the minimum of two independent uniforms has CDF $2t - t^2 > t$, so $p_{(1)}$ cannot serve as a global-null p-value for $H^{(1/2)}$ without correction. This is the global-null / any-of-two regime, not the two-study full-replication regime more commonly meant by "replication," which corresponds to $r = K = 2$ and is addressed separately below. The construction that handles general r is due to Bogomolov and Heller [22] (with foundational antecedents in Benjamini and Heller's earlier work on partial conjunction testing) and uses a Bonferroni-type correction applied to the r-th smallest p-value.

Definition 7.10 (Partial conjunction p-value). Let $p_{i,1}, \dots, p_{i,K}$ be per-study p-values for feature i and let $p_{i,(1)} \leq \cdots \leq p_{i,(K)}$ be their order statistics. The r-out-of-K Bonferroni-style partial conjunction p-value for feature i is

$$p_i^{(r/K)} = (K - r + 1)\, p_{i,(r)} \wedge 1.$$

The intuition is that under $H_i^{(r/K)}$ at least $K - r + 1$ of the p-values are super-uniform nulls. When the procedure looks at the r-th smallest p-value, it is looking at the smallest among at least $K - r + 1$ null p-values (in the worst case where the $r - 1$ non-null p-values occupy the $r - 1$ smallest slots). Applying Bonferroni to the $K - r + 1$ nulls and using the smallest of them gives a multiplier of $K - r + 1$. Two boundary cases anchor the formula:

- $r = 1$ (any-of-K): $p_i^{(1/K)} = K p_{i,(1)}$, the Bonferroni global-null p-value applied to all K studies. The PC null is "all studies are null," and the p-value is super-uniform by the union bound.
- $r = K$ (all-of-K): $p_i^{(K/K)} = p_{i,(K)}$, the largest p-value. The PC null is "at least one study is null," and the maximum p-value is bounded below by the smallest null p-value, hence super-uniform.

THEOREM 7.11 (Super-uniformity of the Bonferroni PC p-value). *Suppose that, under the partial conjunction null $H_i^{(r/K)}$, each true-null per-study p-value is super-uniform. No independence assumption on the joint distribution of the per-study p-values is required. Then*

$$\mathbb{P}\Big(p_i^{(r/K)} \le t\Big) \le t, \qquad t \in [0,1].$$

PROOF. The cleanest case is $K = 2$, $r = 2$: $H_i^{(2/2)}$ says at least one of the two nulls is true, and the PC p-value is $p_i^{(2/2)} = p_{i,(2)} = \max(p_{i,1}, p_{i,2})$. If $p_{i,1}$ is the super-uniform null, then $p_{i,(2)} \ge p_{i,1}$, so $\mathbb{P}(p_{i,(2)} \le t) \le \mathbb{P}(p_{i,1} \le t) \le t$. This inheritance is unconditional on the joint distribution: the inequality $\max(p_1, p_2) \ge p_1$ holds path-wise.

The general case has the same flavor. Under $H_i^{(r/K)}$, at least $K - r + 1$ of the per-study p-values are super-uniform nulls. Choose a subset $\mathcal{N}_i^\star \subseteq \mathcal{N}_i$ of true-null study indices with size $K - r + 1$. For any subset S of size $K - r + 1$, the path-wise inequality

$$p_{i,(r)} \ge \min_{k \in S} p_{i,k}$$

holds: the smallest element of S can be beaten in rank by at most the $r - 1$ elements outside S, so its rank among all K p-values is at most r. Applying this to $S = \mathcal{N}_i^\star$, the union bound over those $K - r + 1$ super-uniform nulls gives $\mathbb{P}(\min_{k \in \mathcal{N}_i^\star} p_{i,k} \le s) \le (K - r + 1)\, s$ without any independence assumption, because the union bound is dependence-free. Hence $\mathbb{P}(p_i^{(r/K)} \le t) \le \mathbb{P}(\min_{k \in \mathcal{N}_i^\star} p_{i,k} \le t/(K - r + 1)) \le t$, establishing the claim. □

REMARK 7.12 (Bonferroni PC vs. Simes/Fisher/Stouffer PC). The proof above gives the Bonferroni-style PC p-value of Definition 7.10 and is dependence-robust: it remains valid under *arbitrary* joint dependence among the per-study p-values $p_{i,1}, \dots, p_{i,K}$ for a given feature i, because the union bound does not require independence. Several other PC p-values trade dependence-robustness for sharper rejection thresholds:

- The *Simes-style PC p-value*

$$p_i^{(r/K),\text{Simes}} \;=\; \left[\min_{r \le j \le K} \frac{(K - r + 1)\, p_{i,(j)}}{j - r + 1}\right] \wedge 1$$

  sharpens Bonferroni under conditions ensuring the Simes inequality for the relevant true-null per-study p-values, such as independence or suitable PRDS [103]; it is well-defined and super-uniform under those conditions, and uniformly no larger than the Bonferroni PC p-value at every realization. The cap at 1 is harmless and kept for parallelism with the Bonferroni formula, although the minimum is already at most one because the $j = K$ term equals $p_{i,(K)}$. Outside the PRDS class, Simes can be anti-conservative, depending on the specific dependence structure.
- Combination-based global-null p-values (Fisher's $-2\sum_k \log p_k$, Stouffer's combined z-score) are typically used in the *global-null* regime $r = 1$, where they combine all K per-study p-values into a single p-value for the global null "all studies are null." Their usual calibration relies on independence (or on specific, justified dependence-aware variance/covariance calibration); positive dependence can inflate the variance of the

combined statistic relative to the independent-study calibration and can make Fisher's and Stouffer's p-values anti-conservative. For example, with two perfectly dependent global-null p-values $p_1 = p_2 = U$, the Stouffer statistic uses $\sqrt{2}\,\Phi^{-1}(1-U)$, so its nominal 0.05 rejection probability is about 0.122. Fisher's independent χ^2_4 calibration similarly rejects with probability $e^{-9.488/4} \approx 0.093$ under $p_1 = p_2 = U$. The advanced exercises work through these calculations.

This is the relevant distinction for replication across studies that share samples, share calibration sets, or otherwise induce dependence among the per-study p-values; see Heller and Bogomolov [58] for a unified treatment.

REMARK 7.13. For $r = K = 2$, Definition 7.10 gives $p_i^{(2/2)} = p_{i,(2)}$, the larger of the two per-study p-values. This recovers the elementary "max p-value" rule for two-study full replication, a simple and common operational choice.

THEOREM 7.14 (FDR control on replicated findings). *Let $\mathcal{R}^{(r/K)}$ be the set of features for which $H_i^{(r/K)}$ is false (i.e., the r-out-of-K replicated features). Assume that for each feature i with $H_i^{(r/K)}$ true, each true-null per-study p-value $p_{i,k}$ is super-uniform – no independence across studies is required for the Bonferroni PC p-value – and that the PC p-values $\{p_i^{(r/K)}\}_{i=1}^m$ are independent (or PRDS) across features. Then applying BH at level α to $\{p_i^{(r/K)}\}_{i=1}^m$ controls FDR on the discovery set with respect to the target $\mathcal{R}^{(r/K)}$:*

$$\mathbb{E}\left[\frac{|\widehat{R} \setminus \mathcal{R}^{(r/K)}|}{|\widehat{R}| \vee 1}\right] \leq \alpha.$$

PROOF. By Theorem 7.11, each PC p-value $p_i^{(r/K)}$ is super-uniform under its PC null, with no cross-study independence needed. Across-feature independence (or PRDS) is assumed. Corollary 5.3 (or the PRDS variant from Chapter 6) now applies directly to the family $\{p_i^{(r/K)}\}_{i=1}^m$. □

EXAMPLE 7.15 (Two studies by hand). Two studies report p-values for four features:

$$(p_{1,1}, \ldots, p_{4,1}) = (0.001,\ 0.002,\ 0.03,\ 0.45), \qquad (p_{1,2}, \ldots, p_{4,2}) = (0.0005,\ 0.5,\ 0.02,\ 0.03).$$

The 2-out-of-2 PC p-values are the maxima per feature:

$$(p_1^{(2/2)}, p_2^{(2/2)}, p_3^{(2/2)}, p_4^{(2/2)}) = (0.001,\ 0.5,\ 0.03,\ 0.45).$$

Sorted, the PC p-values are $0.001, 0.03, 0.45, 0.5$, to be compared with the BH thresholds $0.05 \cdot k/4$ for $k = 1, 2, 3, 4$, namely $0.0125, 0.025, 0.0375, 0.05$. Only $0.001 \leq 0.0125$ crosses (since $0.03 > 0.025$, $0.45 > 0.0375$, and $0.5 > 0.05$), so feature 1 is the unique replicated discovery.

For contrast, BH applied separately within each study at $\alpha = 0.05$ rejects features $\{1, 2, 3\}$ in study 1 (since the sorted study-1 p-values $0.001, 0.002, 0.03, 0.45$ compared with thresholds $0.0125, 0.025, 0.0375, 0.05$ yield rejections at $k = 1, 2, 3$) and features $\{1, 3, 4\}$ in study 2 (the sorted study-2 p-values $0.0005, 0.02, 0.03, 0.5$ compared with the same thresholds yield rejections at $k = 1, 2, 3$, which correspond to features $1, 3, 4$ by original index). The naive intersection of single-study rejections is $\{1, 3\}$; the PC procedure returns only $\{1\}$. The contrast is not that PC is "more conservative" in a loose sense; rather, the two procedures target different nulls. The naive intersection is not calibrated for the partial-conjunction null. A feature can be truly associated in one study and null in another, yet appear in both single-study rejection lists because of a false positive in the null study. The PC procedure instead tests the explicit r-out-of-K replication target. Intersecting separately valid discovery lists is a different operation and need not be calibrated for the replication FDP. The max of feature 3's p-values, $\max(0.03, 0.02) = 0.03$,

is not small enough to certify replication at the joint BH threshold of 0.025 even though both studies individually rejected feature 3.

6. Connections to Regulatory Practice

The closure principle of Chapter 4 gives strong FWER control on hierarchical families through graphical procedures. TreeBH gives FDR control on the same families, but for a different inferential target. Confirmatory regulatory clinical trials primarily use FWER-controlling gatekeeping, because the trial is asking whether *any* primary or secondary endpoint shows a true effect with a small tolerance for false positives over the entire endpoint hierarchy. Genomic, neuroimaging, and benchmark-evaluation pipelines typically prefer FDR-controlling hierarchical procedures, because the goal is to identify many promising findings with a controlled false-discovery proportion at each resolution. The choice between FWER and FDR hierarchical procedures is not a methodological dispute; it is a difference in inferential target.

The relevant regulatory documents include International Council for Harmonisation [66] (ICH E9 addendum on estimands), the EMA *Multiplicity Issues in Clinical Trials* guideline [44], and the FDA *Multiple Endpoints in Clinical Trials* guidance [124]. These documents specify the structural language (primary/secondary/tertiary endpoints, intersection-union and union-intersection tests, gatekeeping) used to organize multi-endpoint confirmatory testing. Their preferred mechanism is FWER-controlling gatekeeping, not FDR-controlling hierarchical procedures: hierarchical FDR procedures of the kind developed in this chapter are appropriate for exploratory and biomarker-discovery contexts, where the analyst trades a tight per-family error rate for a tunable average false-discovery proportion across many findings. Treating these documents as endorsements of hierarchical FDR would overstate their scope.

7. Assumptions in Plain Language

Group BH assumes that the within-null-group p-values make the chosen aggregator super-uniform, and that the resulting group representatives are independent or PRDS across groups; PRDS of the original leaf p-values alone does not automatically imply PRDS of the Simes representatives. The p-filter uses a stronger multilayer joint assumption; a simple sufficient case is mutual independence of all base p-values, with true-null base p-values super-uniform and Simes or weighted-Simes group p-values formed as in the p-filter theorem.

TreeBH uses valid p-values at internal nodes, usually built recursively by applying Simes to the immediate child p-values. Thus it requires the dependence assumptions that make those node-level p-values valid and that make the simple-selection-rule theorem apply; this is what validates the recursive selected-family FDR bound of Theorem 7.6. "Arbitrary within-node dependence" would break ordinary Simes and therefore the TreeBH chain unless one replaces Simes by a Bonferroni or BY-reshaped node-level aggregator.

Partial conjunction tests in the Bonferroni version of Definition 7.10 require only that each true-null per-study p-value is super-uniform; the Bonferroni PC p-value is dependence-robust across studies, by Theorem 7.11. The Simes-style PC p-value is sharper than Bonferroni under conditions ensuring the Simes inequality for the relevant true-null per-study p-values, such as independence or suitable PRDS [103], but loses validity under dependence outside that class. Combination-based global-null p-values (Fisher, Stouffer), typically used in the global-null regime $r = 1$, assume independence across studies in their usual calibration; positive dependence among the combined per-study p-values can inflate the variance of the combined statistic relative to the independent-study calibration and can make these p-values anti-conservative. In replication analyses across genomic consortia or AI benchmarks where study-level p-values may be correlated through shared samples or calibration sets, Bonferroni PC is the safer default;

Simes PC is acceptable when PRDS is plausible; Fisher or Stouffer should be used with their usual independent-study calibration only when across-study independence is verified.

8. Bibliographic Notes

Hierarchical FDR control was introduced by Yekutieli [132]. The modern tree-structured procedure with multi-resolution FDR control is from Bogomolov et al. [23]. The p-filter and its multilayer generalization are due to Barber and Ramdas [7]. Adaptive group-aware procedures for partially ordered side information are from Li and Barber [82]. The original partial conjunction framework is Benjamini and Heller [14]; the replicability extension to multiple features and two-study FDR control is developed by Bogomolov and Heller [22] and Heller and Bogomolov [58]. The connection to regulatory practice in clinical trials is documented in International Council for Harmonisation [66], European Medicines Agency [44], and U.S. Food and Drug Administration [124].

9. Exercises

Basic.

EXERCISE 7.16 (Group BH by hand). Six hypotheses are grouped as $\{1,2,3\}$ and $\{4,5,6\}$ with p-values $(0.028, 0.028, 0.028)$ and $(0.031, 0.40, 0.50)$. Compute the Simes representative for each group and apply BH at $\alpha = 0.05$ on the two representatives. Compare with flat BH on all six p-values. Verify the arithmetic of Example 7.4: flat BH rejects the four smallest p-values (three from Gene 1 plus the smallest p-value of Gene 2), while group BH rejects no group.

EXERCISE 7.17 (Partial conjunction p-value). Two studies report p-values

$$(p_{1,1}, p_{2,1}, p_{3,1}, p_{4,1}) = (0.001, 0.002, 0.03, 0.45), \qquad (p_{1,2}, p_{2,2}, p_{3,2}, p_{4,2}) = (0.0005, 0.5, 0.02, 0.03)$$

for four features. Compute the 2-out-of-2 Bonferroni partial conjunction p-values and apply BH at $\alpha = 0.05$. Compare with BH applied separately within each study and with the naive intersection of the two single-study rejection sets. Verify the arithmetic of Example 7.15: PC rejects only feature 1, while the naive intersection is $\{1,3\}$.

EXERCISE 7.18 (Tree-respecting rejection). Consider a rooted tree with two internal nodes A and B, where A has leaves $\{1,2\}$ and B has leaves $\{3,4\}$. Suppose a tree-aware procedure rejects the internal node A but not B, while flat BH on the four leaves would reject leaves $\{1,3\}$. Under the rule that a leaf may be reported only if every ancestor on its path has been rejected, identify the tree-respecting leaf rejection set and the flat-BH-only rejection.

Intermediate.

EXERCISE 7.19 (Simes within groups). Fill in the proof of Theorem 7.2. In particular, verify the equality

$$\mathbb{P}\left(\bigcup_{k=1}^{m_g} \{U_{(k)} \le tk/m_g\}\right) = t$$

for independent uniform $U_{(1)} \le \cdots \le U_{(m_g)}$ by induction on m_g, or by the area argument used in the original Simes [109] paper.

EXERCISE 7.20 (Bonferroni PC is dependence-robust at $r = K$). For $r = K = 2$, the Bonferroni PC p-value is $p^{(2/2)} = \max(p_1, p_2)$. Show that this PC p-value remains super-uniform under *arbitrary* joint dependence between p_1 and p_2, provided at least one of them is a true-null super-uniform p-value, using the path-wise inequality $\max(p_1, p_2) \ge p_k$ for any true-null index k. Then show that the Simes-style PC formula at $r = K = 2$ (the next exercise with $j \in \{r\}$)

collapses to $p_{(2)} = \max(p_1, p_2)$, so Simes and Bonferroni coincide in the full replication regime; the substantive difference between Simes and Bonferroni PC therefore lives at $r < K$, where Simes can be sharper but also requires PRDS to remain valid.

EXERCISE 7.21 (Simes-style PC p-value: validity and sharpness). The Bonferroni PC p-value of Definition 7.10 can be sharpened under conditions ensuring the Simes inequality for the relevant true-null per-study p-values by the Simes-style PC p-value

$$p_i^{(r/K),\text{Simes}} = \left[\min_{r \le j \le K} \frac{(K - r + 1)\, p_{i,(j)}}{j - r + 1} \right] \wedge 1.$$

Show that this is super-uniform under $H_i^{(r/K)}$, for example when the per-study p-values are independent (or suitable PRDS [103]) and the true-null p-values are super-uniform. Hint: argue that the least favorable case for validity occurs when the $r - 1$ non-null p-values are pushed to zero; in that least-favorable configuration, the $K - r + 1$ largest order statistics $p_{(r)}, \ldots, p_{(K)}$ coincide with the order statistics of the $K - r + 1$ null p-values, and Simes's inequality applied to those null order statistics gives the bound. Then show algebraically that the Simes-style PC p-value is uniformly no larger than the Bonferroni-style PC p-value at every realization: the Simes formula is the minimum over a set whose $j = r$ term equals the Bonferroni value $(K - r + 1)p_{i,(r)}$, so the minimum is $\le$ the Bonferroni value at every realization.

EXERCISE 7.22 (p-filter as BH special case). Verify that the p-filter with $L = 1$ layer at the hypothesis level reduces to ordinary BH. Then verify that the p-filter with $L = 1$ layer using Simes-aggregated group representatives reduces to group BH of Section 2.

Computational.

EXERCISE 7.23 (Genomic pathway simulation). Simulate a three-level hierarchy: 20 pathways, each containing 20 genes (400 genes total), each gene containing 25 variants (10,000 variants total). Generate each variant p-value as $p_i = \Phi(-Z_i)$ where $Z_i \sim \mathcal{N}(\mu_i, 1)$, with $\mu_i = 2.5$ for signal variants and $\mu_i = 0$ for nulls. Place signal in 4 pathways, 5 signal genes per signal pathway (out of 20 genes per pathway), and 5 signal variants per signal gene (out of 25 variants per gene); all other variants are nulls. For TreeBH, compute internal-node p-values recursively from immediate children using Simes; for the p-filter, compute layer group p-values using the same Simes aggregation convention. Implement (a) flat BH at the variant level at $\alpha = 0.05$; (b) TreeBH with per-level targets $(\alpha_{\text{pathway}}, \alpha_{\text{gene}}, \alpha_{\text{variant}}) = (0.05, 0.10, 0.10)$, corresponding to $(\alpha_1, \alpha_2, \alpha_3)$ in the coarsest-to-finest TreeBH order; and (c) the p-filter with layer levels mapped to its finest-to-coarsest convention: $\alpha_{\text{variant}} = 0.10$, $\alpha_{\text{gene}} = 0.10$, and $\alpha_{\text{pathway}} = 0.05$. Average over ≥ 200 replicates and report variant-, gene-, and pathway-level FDR (or recursive selected-family FDR for TreeBH) and power for each procedure. (An extended version of this exercise scales the hierarchy up to genome-wide proportions of $\sim 10^5$ variants; we keep the scale modest here for pedagogical clarity.)

EXERCISE 7.24 (Replication study). Generate three independent batches of $m = 1000$ p-values each. In each study, the per-feature p-value is $p = \Phi(-Z)$ with $Z \sim \mathcal{N}(\mu, 1)$; set $\mu = 0$ for null features and $\mu = 2.5$ for signal features. In the joint design, 100 features are signals in at least two of the three studies ("2-out-of-3 replicated truths"), 100 are signals in exactly one study, and 800 are null in all three. Compare three procedures targeting $r = 2$-out-of-$K = 3$ replication: (a) BH at $\alpha = 0.05$ applied to each study separately, then declaring a feature replicated if it is rejected in at least two of the three study-specific BH analyses, (b) BH applied to the 2-out-of-3 Bonferroni PC p-values, and (c) BH applied to the 2-out-of-3 Simes-style PC p-values (which differ from Bonferroni PC because $r < K$). Average over ≥ 200 replicates and report the realized

replication FDR and power, counting false discoveries as selected features with fewer than two non-null studies and true discoveries as selected features among the 100 replicated-truth features.

EXERCISE 7.25 (Tree depth and power). Generate a three-level balanced tree with 64 leaves (a branching factor of 4 at each level, so depth 3). Leaf p-values are $p_i = \Phi(-Z_i)$ with $Z_i \sim \mathcal{N}(\mu_i, 1)$: $\mu_i = 0$ for nulls and $\mu_i = 2.5$ for signals. Place all signal in one top-level branch, so every leaf below one child of the formal root is a signal and all other leaves are null. Internal-node p-values are formed recursively from immediate children using Simes; in this balanced simulation, one may alternatively compare with a fixed descendant-leaf aggregation rule if it is declared in advance and remains valid. Apply TreeBH with top-tested-layer target $\alpha_1 = 0.05$ and $\alpha_d = 0.10$ for lower tested depths $d \geq 2$, and record leaf-level power. Now repeat with a deeper tree of the same total size: 64 leaves arranged as six tested levels with branching factor 2, keeping all other simulation parameters fixed and using the same per-depth target convention. Average over ≥ 200 replicates and discuss how the additional depth affects leaf-level power and the recursive selected-family FDR.

Advanced.

EXERCISE 7.26 (Fixed layer weights). Suppose a two-layer p-filter uses fixed nonnegative weights a_1, a_2 with $a_1 + a_2 \leq 1$ and layer levels $\alpha_1 = a_1\alpha_{\text{total}}$ and $\alpha_2 = a_2\alpha_{\text{total}}$. If the layer-$\ell$ procedure controls its layer FDR at level α_ℓ, show that the sum of the two layer FDR bounds is at most α_{total}. Then give a concrete reason why choosing a_ℓ from the same p-values used for layer-ℓ testing would require a new validity argument.

EXERCISE 7.27 (TreeBH versus flat BH for high-density signal). Use a four-level binary tree with 16 leaves. Put signals in all four leaves below the leftmost depth-two internal node and set all other leaves to null. Generate independent one-sided p-values from $Z_i \sim N(2.5, 1)$ for signal leaves and $Z_i \sim N(0, 1)$ for null leaves, with internal-node p-values formed recursively by Simes combinations of immediate-child p-values, matching Section 4. Compare flat BH at the leaf level with TreeBH using $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.05, 0.10, 0.10, 0.10)$ over at least 500 Monte Carlo replicates. Report leaf-level power and explain the interpretation gained when a rejected leaf is accompanied by rejected ancestors.

EXERCISE 7.28 (Stouffer at the global null can fail under positive dependence). For two studies with one-sided z-scores $Z_k = \Phi^{-1}(1 - p_k)$, $k = 1, 2$, the *Stouffer global-null combination* (i.e., the $r = 1$, $K = 2$ Stouffer PC) is

$$p^{(1/2),\text{Stouffer}} \;=\; 1 - \Phi\Big(\frac{Z_1 + Z_2}{\sqrt{2}}\Big).$$

Under the global null (both $p_k \sim \text{Unif}(0, 1)$) with *independent* studies, show that $p^{(1/2),\text{Stouffer}}$ is exactly uniform. Then suppose p_1 and p_2 are *perfectly positively dependent*, $p_1 = p_2 = U \sim \text{Unif}(0, 1)$ (so $Z_1 = Z_2 = Z$ with $Z \sim \mathcal{N}(0, 1)$). Show that the Stouffer combination evaluates to $1 - \Phi(\sqrt{2}\, Z)$, and that at the nominal $\alpha = 0.05$ cutoff, $\mathbb{P}(p^{(1/2),\text{Stouffer}} \leq 0.05) = \mathbb{P}(Z \geq 1.645/\sqrt{2}) = \mathbb{P}(Z \geq 1.163) \approx 0.122$, so the Stouffer global-null p-value is anti-conservative by more than a factor of 2. Discuss why this kind of failure does not happen for the Bonferroni global-null p-value $p^{(1/2),\text{Bonf}} = 2\min(p_1, p_2) \wedge 1$, which remains super-uniform under any dependence by the union bound, and relate the contrast to the dependence-robustness theme of Heller and Bogomolov [58].

EXERCISE 7.29 (Fisher at the global null can fail under positive dependence). Define the *Fisher global-null combination* for $K = 2$ studies as

$$p^{(1/2),\text{Fisher}} \;=\; \mathbb{P}_{X \sim \chi^2_4}(X \geq -2\log p_1 - 2\log p_2),$$

i.e., the upper tail of a χ^2_4 distribution evaluated at the Fisher statistic. Under the global null with independent studies, $-2\log p_1 - 2\log p_2 \sim \chi^2_4$ exactly, so $p^{(1/2),\text{Fisher}}$ is uniform. Now suppose the two studies have perfectly positively dependent p-values $p_1 = p_2 = U \sim \text{Unif}(0,1)$, so the Fisher statistic equals $-4\log U$, which has CDF $1 - e^{-x/4}$ on $[0,\infty)$ (it is $4 \cdot \text{Exp}(1)$ and *not* χ^2_4). The χ^2_4 critical value at $\alpha = 0.05$ is ≈ 9.488; compute $\mathbb{P}(-4\log U \geq 9.488) = e^{-9.488/4} \approx 0.093$, which exceeds the nominal 0.05 and demonstrates anti-conservatism. Compare with the Bonferroni global-null p-value $p^{(1/2),\text{Bonf}} = 2\min(p_1, p_2) \wedge 1$, and discuss why combination methods like Fisher/Stouffer cannot be deployed with their usual independent-study calibration when across-study independence is not verifiable.

EXERCISE 7.30 (Sharp PC with verified side information). Heller and Bogomolov [58] discuss sharper procedures that use information about $|\mathcal{N}_i|$ beyond the worst-case bound. Suppose a rule fixed by the study design before looking at the p-values identifies a subset S_i of $K - r + 1$ studies that is guaranteed to consist of true nulls whenever $H_i^{(r/K)}$ holds. Derive the corresponding "conditional" PC p-value – the Bonferroni union bound applied only to the studies in S_i – and show that, under independence across features, BH applied to this conditional PC p-value controls FDR at level α without further inflation. Comment on what could go wrong if S_i is inferred from the same p-values, or if the side information is only an unverified belief: explain why such a procedure cannot in general be justified for FDR control uniformly over the PC null.

CHAPTER 8

E-Values, Safe Testing, and Multiple Testing

Suppose a laboratory is running many experiments at once. Some outcomes are observed today, more data may arrive next week, and the same subjects, batches, or model outputs may be reused across several hypotheses. The analyst wants to make discoveries while the evidence is still being monitored. This is the setting in which ordinary fixed-sample p-values become fragile: if we repeatedly look at the data and then decide whether to continue, the final fixed-sample p-value no longer has the error guarantee suggested by its formula. Here *fixed-sample* means that the p-value formula was calibrated for one prespecified sample size and one planned analysis. It was not calibrated for a workflow in which the sample size or the number of looks is chosen after seeing interim evidence.

This chapter introduces e-values as an alternative evidence scale. The main idea is deliberately simple. A p-value controls a tail probability under the null; an e-value controls an expectation under the null. This expectation budget can be spent in ways that are difficult for p-values: e-values can be multiplied over time when each new factor is conditionally valid, averaged across procedures, pooled across data sources, and used in the e-BH procedure to control FDR under arbitrary dependence.

The chapter is organized around a running question. How should we build an evidence measure that remains valid after monitoring, combining, or selecting among many hypotheses? The answer has three parts:

(1) construct evidence whose null expectation is controlled;
(2) prove that the construction did not spend too much null-evidence budget;
(3) apply a rejection rule whose proof uses only that budget.

1. Setup and Notation

This chapter uses the multiple-testing notation introduced in Chapters 5–6. We recall the formulas needed below, and then introduce the new sequential and e-value notation for this chapter.

For one hypothesis, let X denote the observed data and let $\mathcal{H}_0$ be the null model, represented as a set of probability distributions for X. A simple null has the form $\mathcal{H}_0 = \{P_0\}$. A composite null has the form

$$\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\},$$

where the null parameter set Θ_0 contains more than one value. When densities exist, $p_0(x)$ denotes the density of P_0, $p_\theta(x)$ denotes the density of P_θ, and $q(x)$ denotes the density of an alternative distribution Q. These are density functions, not p-values.

Recall that for multiple testing there are m hypotheses $H_1, \dots, H_m$. The set of true null indices is $\mathcal{I}_0 \subset \{1, \dots, m\}$, with $m_0 = |\mathcal{I}_0|$. For a rejection set $\mathcal{R}$, we write

$$R = |\mathcal{R}|, \qquad V = |\mathcal{R} \cap \mathcal{I}_0|, \qquad \mathrm{FDP} = \frac{V}{R \vee 1}, \qquad \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

The convention $R \vee 1$ makes $\mathrm{FDP} = 0$ when there are no rejections.

For sequential problems, $\mathcal{F}_t$ is the information available after the tth look or batch. A rule is *predictable* at time t if it is $\mathcal{F}_{t-1}$-measurable; in plain language, it is chosen using only past data. A stopping time τ is a time at which the decision to stop is made using the information available so far. A deterministic finite horizon is denoted by K. We reserve t for time or a generic threshold variable, and write $\widehat{t}$ for data-dependent thresholds.

2. Evidence on an Expectation Scale

DEFINITION 8.1 (E-value). A nonnegative random variable $E = E(X)$ is an e-value for $\mathcal{H}_0$ if

$$\sup_{P \in \mathcal{H}_0} \mathbb{E}_P[E] \leq 1.$$

Large values of E are evidence against the null.

Recall from Chapter 2 that a p-value, denoted by the lowercase letter p, is a $[0,1]$-valued statistic satisfying

$$\sup_{P \in \mathcal{H}_0} \mathbb{P}_P(p \leq u) \leq u, \qquad 0 \leq u \leq 1.$$

Small p-values are evidence against the null. Although p is computed from the data, we avoid writing $p(X)$ here because symbols such as $p_0(x)$ and $p_\theta(x)$ denote density functions in this chapter.

The p-value and e-value definitions use different currencies. A p-value promises that small values are rare under the null. An e-value promises that the average amount of null evidence is at most one. The second promise is weaker in some fixed sample problems, but it is algebraically useful because expectations add and conditional expectations multiply.

PROPOSITION 8.2 (Markov calibration). *If E is an e-value for $\mathcal{H}_0$, then*

$$p_E = \min\{1, 1/E\}$$

is a valid p-value for $\mathcal{H}_0$.

PROOF. Fix $P \in \mathcal{H}_0$. For $0 < u \leq 1$,

$$\mathbb{P}_P(p_E \leq u) = \mathbb{P}_P(E \geq 1/u) \leq u\, \mathbb{E}_P[E] \leq u,$$

where the inequality is Markov's inequality. The case $u = 0$ is immediate. The truncation at one only keeps the calibrated p-value inside $[0,1]$. □

The reverse direction is not obtained by $1/p$. If $p \sim \mathrm{Unif}(0,1)$, then $\mathbb{E}[1/p] = \infty$. To convert p-values into e-values, one needs a calibrating function f satisfying $\mathbb{E}[f(U)] \leq 1$ for $U \sim \mathrm{Unif}(0,1)$. This point is worth remembering in applications: p-values and e-values are not just two formats for the same number.

3. Likelihood Ratios, Bayes Factors, and Composite Nulls

The most basic way to construct an e-value is to compare two probability models.

PROPOSITION 8.3 (Likelihood-ratio e-value). *Suppose $\mathcal{H}_0 = \{P_0\}$ is simple and Q is absolutely continuous with respect to P_0. Then*

$$E(X) = \frac{dQ}{dP_0}(X)$$

is an e-value for P_0.

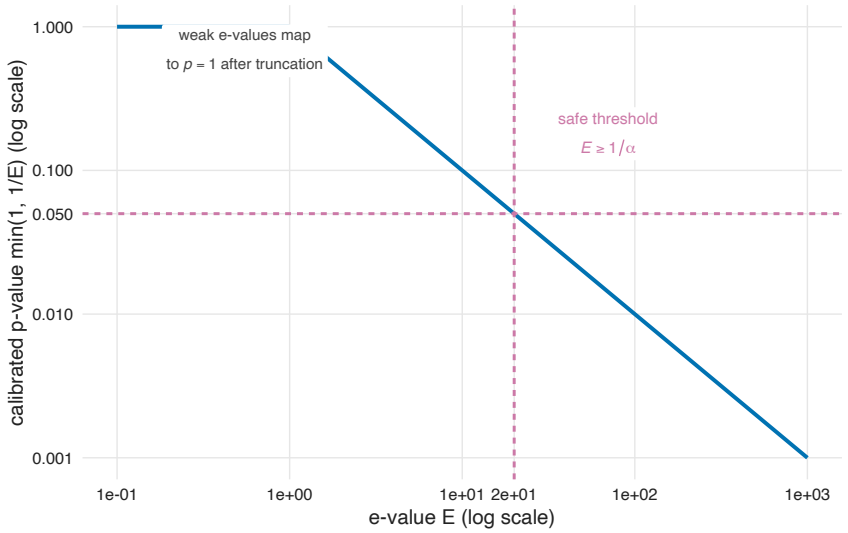

FIGURE 1. Calibration from e-values to p-values on log-log axes. The conversion $p_E = \min\{1, 1/E\}$ is always valid by Markov's inequality. Weak e-values with $E < 1$ all map to $p_E = 1$, so calibration is valid but can lose information.

PROOF. By the definition of the Radon-Nikodym derivative,

$$\mathbb{E}_{P_0}\left[\frac{dQ}{dP_0}(X)\right] = \int \frac{dQ}{dP_0}\, dP_0 = \int dQ = 1.$$

□

If densities are available, the same e-value is $E(X) = q(X)/p_0(X)$. If Q is a mixture over alternatives,

$$q(x) = \int p_\theta(x)\, dG_1(\theta),$$

then $q(X)/p_0(X)$ is a Bayes factor and also an e-value for the simple null. The phrase "simple null" is doing real work. The denominator p_0 is the density of the one null distribution whose expectation must be controlled. More generally, for any prior distribution G on a parameter set, write

$$p_G(x) = \int p_\theta(x)\, dG(\theta),$$

and let Q_G denote the probability distribution with density p_G. This notation is used below for both alternative mixtures and null mixtures.

EXAMPLE 8.4 (Gaussian likelihood ratio). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$. We test

$$H_0 : \mu = 0 \qquad \text{against} \qquad H_1 : \mu = \mu_1.$$

The joint density under μ is

$$p_\mu(x_1, \ldots, x_n) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu)^2\right\}.$$

Therefore

$$\frac{p_{\mu_1}(X_1,\dots,X_n)}{p_0(X_1,\dots,X_n)} = \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(X_j-\mu_1)^2 + \frac{1}{2}\sum_{j=1}^{n}X_j^2\right\}$$
$$= \exp\left\{\mu_1\sum_{j=1}^{n}X_j - \frac{n\mu_1^2}{2}\right\}.$$

This is an e-value for $H_0 : \mu = 0$. Under the alternative $\mu = \mu_1$, its logarithm has positive mean $n\mu_1^2/2$, so the e-value tends to grow.

Composite nulls require more care. Suppose $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$. A Bayes factor of the form

$$\frac{p_{G_1}(X)}{p_{G_0}(X)}$$

averages the null over a prior G_0 on Θ_0. It is not automatically an e-value for the whole composite null. To be an e-value, it must satisfy

$$\mathbb{E}_{P_\theta}\left[\frac{p_{G_1}(X)}{p_{G_0}(X)}\right] \le 1 \qquad \text{for every } \theta \in \Theta_0.$$

Averaging over the null prior gives only

$$\mathbb{E}_{\theta\sim G_0}\mathbb{E}_{P_\theta}\left[\frac{p_{G_1}(X)}{p_{G_0}(X)}\right] = 1,$$

which does not control the worst null value of θ.

One useful research problem is therefore: given an alternative distribution Q, find a null distribution Q_0^* that makes dQ/dQ_0^* valid for the composite null and still powerful under Q. Reverse information projection is one method for choosing such a candidate denominator. Let $\mathcal{G}_0$ be a collection of prior distributions supported on the null parameter set Θ_0, and define

$$\mathcal{Q}_0 = \{Q_G : G \in \mathcal{G}_0\},$$

where Q_G has density p_G. Assume $\mathcal{Q}_0$ is convex: if $Q_a, Q_b \in \mathcal{Q}_0$, then $\lambda Q_a + (1-\lambda)Q_b \in \mathcal{Q}_0$ for $0 \le \lambda \le 1$. Convexity means that if two null mixtures are allowed, then any randomized mixture of those two null mixtures is also allowed. This matters because the optimization below searches over denominators for a likelihood ratio: a convex search set rules out artificial gaps between admissible null mixtures and gives the projection problem the usual "closest point in a set" interpretation. The reverse information projection chooses

$$Q_0^* \in \arg\min_{Q_0\in\mathcal{Q}_0} D(Q \,\|\, Q_0),$$

where D is Kullback-Leibler divergence. The operational interpretation is simple: among the allowed null mixture distributions, choose the one closest to the alternative in the direction relevant for likelihood-ratio growth. After choosing Q_0^*, one must still verify the e-value condition

$$\mathbb{E}_{P_\theta}\left[\frac{dQ}{dQ_0^*}(X)\right] \le 1 \qquad \text{for every } \theta \in \Theta_0.$$

The Gaussian half-line example below shows this verification directly.

Example 8.5 (A one-sided Gaussian composite null). Consider one observation $X \sim N(\mu, 1)$, with

$$H_0 : \mu \le 0 \qquad \text{and} \qquad Q = N(\mu_1, 1), \quad \mu_1 > 0.$$

The null distribution closest to Q in Kullback-Leibler divergence is the boundary distribution $N(0,1)$, because for normals with common variance

$$D\{N(\mu_1,1)\,\|\,N(\mu,1)\} = \frac{(\mu_1-\mu)^2}{2},$$

which is minimized over $\mu \le 0$ at $\mu = 0$. The resulting e-value is

$$E(X) = \exp\{\mu_1 X - \mu_1^2/2\}.$$

For every $\mu \le 0$,

$$\mathbb{E}_{P_\mu}[E(X)] = \exp\{\mu_1\mu\} \le 1.$$

Thus the same likelihood ratio is valid for the whole half-line null, not only for the boundary point.

4. Betting Scores and Safe Tests

The betting language is only a bookkeeping device. No actual money is needed. Imagine that we keep an evidence account whose starting balance is one. The number one means "no accumulated evidence yet," so all later values are measured relative to this neutral starting point. At time t, after seeing the next piece of data, we multiply the current balance by a nonnegative factor E_t. If $E_t = 2$, the account doubles; if $E_t = 1/2$, the account is cut in half; if $E_t = 1$, the new data leave the account unchanged. The cumulative evidence, also called the capital process, is therefore

$$M_t = \prod_{s=1}^{t} E_s, \qquad M_0 = 1.$$

The statistical constraint is that, under the null, this account is not allowed to grow in expectation. Informally, the null model must view the game as fair or unfavorable to the analyst. Thus each new multiplier should have conditional null expectation at most one, given the past information. Then the capital can still go up on some sample paths, but it cannot have positive expected drift under the null. Under a well-chosen alternative, the multipliers tend to be larger than one often enough that the product may grow quickly.

For independent observations with a simple null and a fixed alternative,

$$E_t = \frac{q(X_t)}{p_0(X_t)}$$

is a per-observation likelihood-ratio e-value. The cumulative likelihood ratio is

$$M_t = \prod_{s=1}^{t} \frac{q(X_s)}{p_0(X_s)}.$$

Under P_0, $\mathbb{E}_{P_0}[M_t] = 1$ for each fixed t. A level-α safe test rejects when

$$M_t \ge \frac{1}{\alpha}.$$

For a fixed t, Markov's inequality gives

$$\mathbb{P}_{P_0}\left(M_t \ge \frac{1}{\alpha}\right) \le \alpha.$$

The next section explains why, with the right sequential construction, the same boundary can be monitored over time.

FIGURE 2. Eight simulated paths of $\log_{10} M_t$ for the Gaussian likelihood-ratio e-process with $\alpha = 0.05$, $K = 120$ looks, and alternative mean $\mu_1 = 0.30$. The horizontal line is $\log_{10}(1/\alpha)$. The x-axis is the look index t, which equals the sample size in this one-observation-per-look simulation.

The logarithm of M_t explains what a good e-value is trying to optimize. If $X_1, X_2, \ldots$ are i.i.d. from an alternative Q, then

$$\frac{1}{t} \log M_t = \frac{1}{t} \sum_{s=1}^{t} \log E_s \xrightarrow{Q\text{-a.s.}} \mathbb{E}_Q[\log E_1].$$

So the long-run growth rate is expected log evidence under the alternative, subject to the null constraint $\mathbb{E}_{P_0}[E_1] \leq 1$. This is the statistical content behind betting terminology: we want evidence that cannot grow in expectation under the null but grows multiplicatively under alternatives that matter.

The safe threshold and the fixed-sample Neyman-Pearson threshold solve different problems. In Example 8.4, the fixed-sample Neyman-Pearson test at sample size n rejects when

$$\sum_{j=1}^{n} X_j \geq \sqrt{n}\, z_{1-\alpha}.$$

In terms of the e-value M_n, this is

$$M_n \geq \exp\left\{ \mu_1 \sqrt{n}\, z_{1-\alpha} - \frac{n\mu_1^2}{2} \right\}.$$

The Neyman-Pearson threshold is optimal for the prespecified sample size. The safe threshold $1/\alpha$ is designed for monitoring and continuation. It may be more conservative at a fixed sample size because it is protecting a larger workflow.

5. Optional Continuation and E-Processes

Optional continuation means that future data collection may depend on current evidence. A study may continue only if the interim evidence is promising, or a second group may decide to add data after seeing an inconclusive first result. The key requirement is that the next evidence multiplier must be valid after conditioning on the information already seen.

DEFINITION 8.6 (Conditional e-value). Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. A nonnegative $\mathcal{F}_t$-measurable random variable E_t is a conditional e-value at time t for $\mathcal{H}_0$ if

$$\mathbb{E}_P[E_t \mid \mathcal{F}_{t-1}] \leq 1 \qquad \text{for every } P \in \mathcal{H}_0.$$

PROPOSITION 8.7 (Products of conditional e-values). *If E_t is a conditional e-value at every time t, then*

$$M_t = \prod_{s=1}^{t} E_s, \qquad M_0 = 1,$$

is a nonnegative supermartingale under every $P \in \mathcal{H}_0$.

PROOF. The product M_{t-1} is determined by information available before time t, so it is $\mathcal{F}_{t-1}$-measurable. Therefore

$$\begin{aligned}\mathbb{E}_P[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}_P[M_{t-1}E_t \mid \mathcal{F}_{t-1}] \\ &= M_{t-1}\mathbb{E}_P[E_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}.\end{aligned}$$

Nonnegativity is inherited from the factors E_t. This is exactly the definition of a nonnegative supermartingale. □

Such a process is called an e-process. The name is not just terminology: the whole path is valid, not only a single prespecified time.

THEOREM 8.8 (Ville's inequality). *If $(M_t)_{t\geq 0}$ is a nonnegative supermartingale with $M_0 = 1$, then for every $\alpha \in (0,1)$,*

$$\mathbb{P}\left(\sup_{t\geq 0} M_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

PROOF. Let

$$\tau = \inf\{t : M_t \geq 1/\alpha\},$$

with $\inf \varnothing = \infty$. Fix a deterministic horizon K. The stopped time $\tau \wedge K$ is bounded. For a bounded stopping time $\rho \leq K$, the stopped process $M_{t\wedge\rho}$, $0 \leq t \leq K$, is still a supermartingale; iterating the one-step inequalities gives $\mathbb{E}[M_\rho] \leq \mathbb{E}[M_0]$. Applying this with $\rho = \tau \wedge K$ gives

$$\mathbb{E}[M_{\tau\wedge K}] \leq \mathbb{E}[M_0] = 1.$$

On the event $\{\tau \leq K\}$, the stopped value is at least $1/\alpha$. On the complementary event it is nonnegative. Hence

$$1 \geq \mathbb{E}[M_{\tau\wedge K}] \geq \frac{1}{\alpha}\mathbb{P}(\tau \leq K).$$

Thus $\mathbb{P}(\tau \leq K) \leq \alpha$. The events $\{\tau \leq K\}$ increase to $\{\tau < \infty\}$ as $K \to \infty$, so continuity from below gives the claim. □

What to check in your own research problem:

- What is the filtration $\mathcal{F}_t$, meaning what information is available before the next decision?
- Is each multiplier E_t valid conditionally on the past, or is the whole process M_t otherwise known to be a nonnegative supermartingale?
- Is the reported result a stopped value M_τ, a crossing event $\sup_t M_t \geq 1/\alpha$, or a fixed-time value M_K?

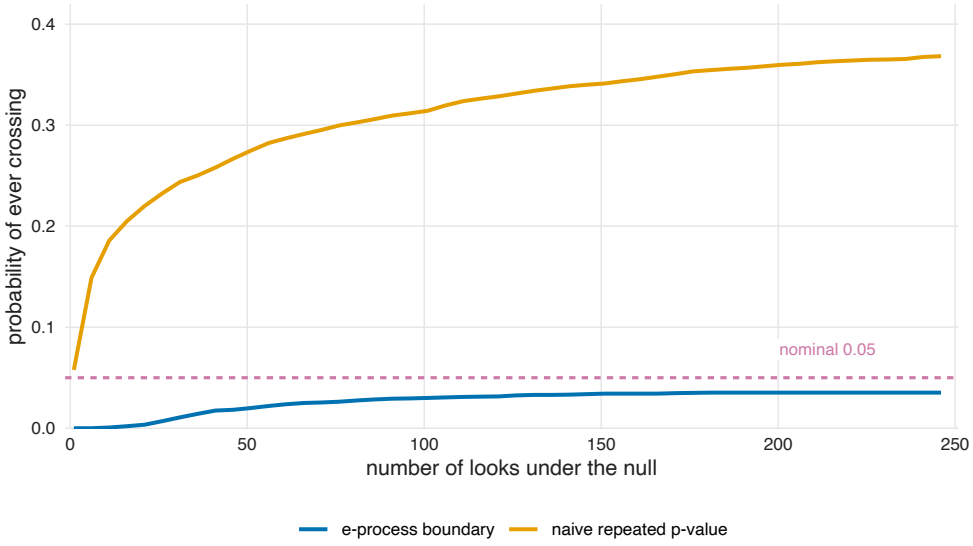

FIGURE 3. A null simulation with 4000 Monte Carlo runs and 250 looks. Repeatedly checking a fixed-sample one-sided Gaussian p-value inflates the probability of ever crossing 0.05. The likelihood-ratio e-process boundary stays controlled because Ville's inequality applies to the whole monitored path.

6. E-BH: FDR Control by Aggregate Null Evidence

Now return to m hypotheses. Suppose hypothesis H_i has a nonnegative evidence score E_i. For individual e-values one often assumes $\mathbb{E}[E_i] \le 1$ for every true null i. For multiple testing, the proof only needs the aggregate null-evidence condition

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \le m. \tag{A}$$

Individual e-values imply (A), since $m_0 \le m$, but (A) is more flexible.

DEFINITION 8.9 (The e-BH procedure). Order the evidence scores from largest to smallest:

$$E_{(1)} \ge E_{(2)} \ge \cdots \ge E_{(m)}.$$

At FDR level α, define

$$\widehat{k} = \max\left\{k \in \{1,\dots,m\} : E_{(k)} \ge \frac{m}{\alpha k}\right\},$$

with $\widehat{k} = 0$ if the set is empty. Reject the hypotheses attached to the $\widehat{k}$ largest e-values.

THEOREM 8.10 (E-BH FDR control). *If $E_1,\dots,E_m$ are nonnegative and satisfy (A), then e-BH controls FDR:*

$$\mathrm{FDR} \le \alpha.$$

In particular, if each true null $i \in \mathcal{I}_0$ satisfies $\mathbb{E}[E_i] \le 1$, then $\mathrm{FDR} \le (m_0/m)\alpha \le \alpha$.

PROOF. Let $\mathcal{R}$ be the e-BH rejection set and $R = |\mathcal{R}|$. If $R = 0$, then FDP $= 0$. Suppose $R > 0$. Every rejected hypothesis i satisfies

$$E_i \ge E_{(R)} \ge \frac{m}{\alpha R}.$$

Therefore, path by path,

$$\frac{\mathbf{1}\{i \in \mathcal{R}\}}{R} \leq \frac{\alpha E_i}{m}.$$

Summing over true nulls gives

$$\mathrm{FDP} = \frac{\sum_{i \in \mathcal{I}_0} \mathbf{1}\{i \in \mathcal{R}\}}{R \vee 1} \leq \frac{\alpha}{m} \sum_{i \in \mathcal{I}_0} E_i.$$

Taking expectations and applying (A),

$$\mathrm{FDR} \leq \frac{\alpha}{m} \mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \leq \alpha.$$

If $\mathbb{E}[E_i] \leq 1$ for every true null, the aggregate expectation is at most m_0, giving the sharper bound $(m_0/m)\alpha$. □

The proof does not condition on other p-values, does not require independence, and does not use PRDS. The random rejection set may depend on all e-values in an arbitrary way; the pathwise inequality above absorbs that dependence before expectations are taken. This is the main reason e-BH is useful when evidence streams share subjects, features, prompts, batches, or model outputs.

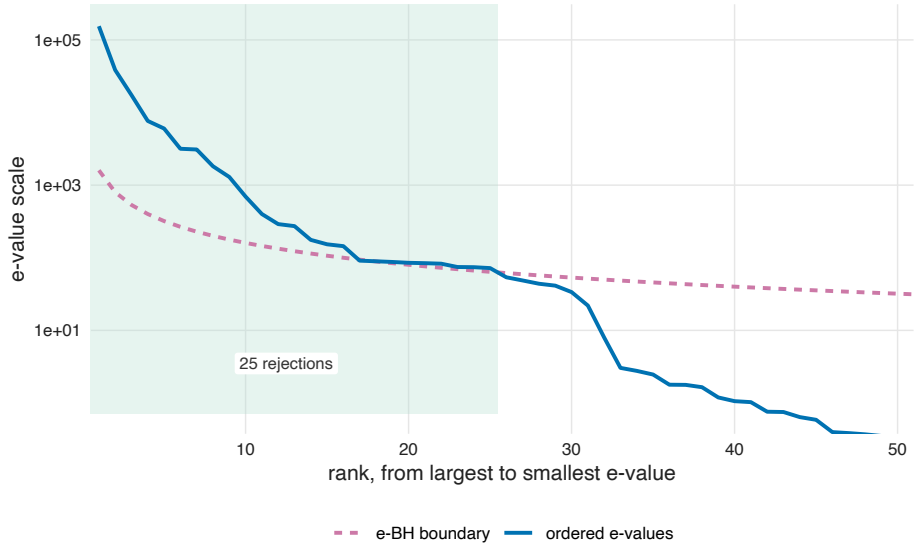

FIGURE 4. The e-BH rule orders e-values from largest to smallest and compares the ordered values with the boundary $m/(\alpha k)$. In the displayed simulation, e-values above the boundary form the rejection set.

What to check in your own research problem:

- Can you prove $\mathbb{E}[E_i] \leq 1$ for each true null, or at least the aggregate bound $\mathbb{E}[\sum_{i \in \mathcal{I}_0} E_i] \leq m$?
- Are the E_i's nonnegative on every data realization?
- Is the goal ordinary FDR for one pooled rejection list, or does the analysis also need a separate error guarantee inside each subgroup or family?

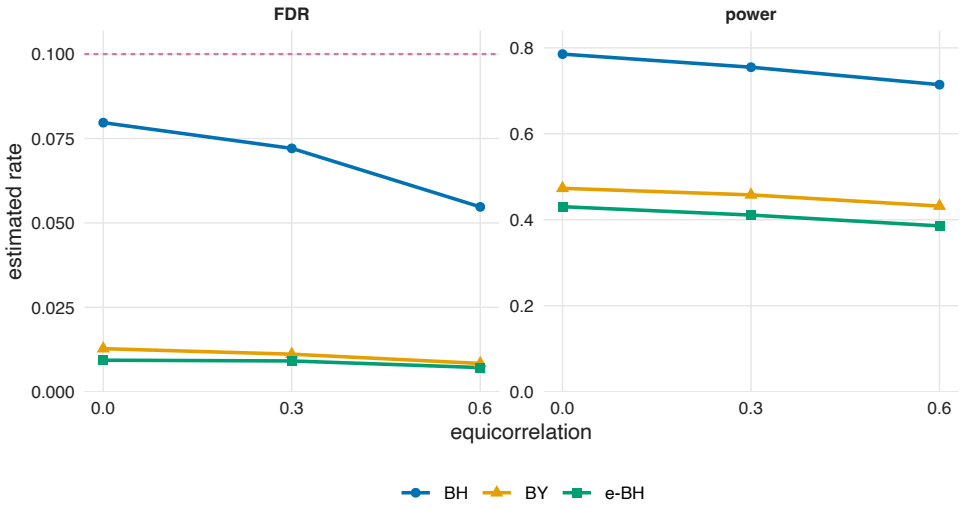

FIGURE 5. Simulation with $m = 400$ hypotheses, 80 nonnulls, equicorrelated Gaussian test statistics, 650 Monte Carlo runs, and $\alpha = 0.1$. The lesson is not that e-BH is always more powerful; it is that e-BH's proof uses an e-value moment condition rather than a p-value dependence condition.

7. How Threshold Rules Create E-Values

Many p-value procedures work by choosing a data-dependent threshold and rejecting everything below it. The e-value viewpoint turns this into a research recipe:

threshold rule
$\Downarrow$
induced evidence vector
$\Downarrow$
aggregate null-evidence proof
$\Downarrow$
e-BH FDR control

The point is not to rename old procedures. The point is to identify the exact budget that must be controlled when designing new procedures.

BH as the first example. Recall from Chapter 5 that BH can be written in threshold form. Let $p_1, \dots, p_m$ be p-values. For a threshold $s \in [0, 1]$, define

$$R(s) = \sum_{i=1}^{m} \mathbf{1}\{p_i \le s\}.$$

The BH threshold can be written as

$$\widehat{t}_{\mathrm{BH}} = \sup\left\{0 < s \le 1 : \frac{ms}{R(s) \vee 1} \le \alpha\right\}.$$

When the set is empty, set $\widehat{t}_{\mathrm{BH}} = 0$. The BH rejection set is

$$\mathcal{R}_{\mathrm{BH}} = \{i : p_i \le \widehat{t}_{\mathrm{BH}}\}.$$

On the event $\widehat{t}_{\mathrm{BH}} > 0$, define the induced evidence score

$$E_i^{\mathrm{BH}} = \frac{1}{\widehat{t}_{\mathrm{BH}}} \mathbf{1}\{p_i \le \widehat{t}_{\mathrm{BH}}\}.$$

If $\widehat{t}_{\mathrm{BH}} = 0$, define $E_i^{\mathrm{BH}} = 0$ for all i.

PROPOSITION 8.11 (The BH rejection set as e-BH). *Applying e-BH at level α to the induced vector $(E_i^{\mathrm{BH}})_{i=1}^m$ gives the same rejection set as BH.*

PROOF. If $\widehat{t}_{\mathrm{BH}} = 0$, both procedures reject nothing. Suppose $\widehat{t}_{\mathrm{BH}} > 0$ and let $\widehat{k} = R(\widehat{t}_{\mathrm{BH}})$. If $\widehat{k} = 0$, then all induced e-values are zero and both procedures reject nothing. Now assume $\widehat{k} > 0$. By the BH threshold definition,

$$\frac{m\widehat{t}_{\mathrm{BH}}}{\widehat{k}} \le \alpha, \qquad \text{so} \qquad \frac{1}{\widehat{t}_{\mathrm{BH}}} \ge \frac{m}{\alpha \widehat{k}}.$$

The $\widehat{k}$ BH rejections have evidence $1/\widehat{t}_{\mathrm{BH}}$, and all non-rejections have evidence 0. Hence the $\widehat{k}$th largest induced e-value crosses the e-BH boundary, while no non-rejected hypothesis can be added because its induced e-value is zero. Thus e-BH selects exactly $\mathcal{R}_{\mathrm{BH}}$. □

This proposition is algebraic. It does not prove BH validity. Validity comes from the budget question:

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i^{\mathrm{BH}}\right] \le m?$$

Under the usual independent-null assumptions for BH, this aggregate bound can be proved by a reverse-time martingale or leave-one-out argument. The intuition is as follows. If a true-null p-value is rejected at a data-dependent threshold $\widehat{t}_{\mathrm{BH}}$, it contributes $1/\widehat{t}_{\mathrm{BH}}$. For a fixed threshold s, a super-uniform null p-value has expected contribution at most $\mathbb{P}(p_i \le s)/s \le 1$. The technical argument shows that the same bound survives when s is the BH threshold chosen from the data, because replacing one rejected true-null p-value by zero does not change the final number of BH rejections in the leave-one-out comparison.

The general threshold template. The BH example suggests a broader template. Suppose a procedure has candidate thresholds $s \in \mathcal{D}$. At threshold s, let

$$R_i(s) \in \{0, 1\}$$

indicate whether hypothesis i would be rejected, and let

$$R(s) = \sum_{i=1}^m R_i(s).$$

Let $\widehat{M}(s)$ be the procedure's estimate or upper bound for the number of false discoveries at threshold s. The procedure chooses

$$\widehat{t} = \sup\left\{ s \in \mathcal{D} : \frac{\widehat{M}(s)}{R(s) \vee 1} \le \alpha \right\}$$

and rejects all i with $R_i(\widehat{t}) = 1$.

The induced evidence formula requires a positive false-discovery budget when there is at least one rejection. We therefore assume

$$R(\widehat{t}) > 0 \quad \Longrightarrow \quad \widehat{M}(\widehat{t}) > 0.$$

If a candidate rule can have $\widehat{M}(\widehat{t}) = 0$ and $R(\widehat{t}) > 0$, it must be modified before using this template; common fixes are to add a $+1$ correction to the budget or to disallow thresholds with zero budget and positive rejections.

The induced evidence vector is

$$E_i = \frac{mR_i(\widehat{t})}{\widehat{M}(\widehat{t})},$$

with the convention $E_i = 0$ for all i when $R(\widehat{t}) = 0$. A rejected hypothesis receives evidence equal to m divided by the estimated false-discovery budget; a non-rejected hypothesis receives zero.

PROPOSITION 8.12 (Threshold rule to e-BH). *For a threshold rule of the form above, e-BH applied to the induced vector $E_i = mR_i(\widehat{t})/\widehat{M}(\widehat{t})$ contains the threshold-rule rejection set. If all non-rejections have $R_i(\widehat{t}) = 0$, the two sets are equal. If the induced vector also satisfies*

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \le m,$$

then e-BH controls FDR at level α.

PROOF. Let $\widehat{r} = R(\widehat{t})$. If $\widehat{r} = 0$, there is nothing to prove. If i is rejected by the threshold rule, then

$$E_i = \frac{m}{\widehat{M}(\widehat{t})}.$$

The threshold definition gives

$$\frac{\widehat{M}(\widehat{t})}{\widehat{r}} \le \alpha, \qquad \text{so} \qquad E_i \ge \frac{m}{\alpha\widehat{r}}.$$

Thus at least $\widehat{r}$ hypotheses cross the e-BH boundary at rank $\widehat{r}$. Non-rejections have induced evidence zero, so e-BH cannot add them when all positive induced evidence belongs to the threshold rejection set. The FDR statement is exactly Theorem 8.10. □

The table below gives examples of the template. Each row should be read in the same way: the middle column supplies the false-discovery budget $\widehat{M}(s)$, the last column supplies the fixed-threshold rejection rule $R_i(s)$, and Proposition 8.12 then gives the induced e-values once the data choose $\widehat{t}$.

Raw p-values are not always the best ranking scale. A researcher may want to borrow information across hypotheses, as in empirical-Bayes or local-FDR ranking, or use external information such as covariates, prior biological knowledge, group structure, or weights learned from independent data. A simple way to represent this is to rank hypothesis i by a transformed score. Before using the p-value p_i, choose a function ϕ_i and define

$$q_i = \phi_i(p_i).$$

Assume ϕ_i is nondecreasing. Then a smaller p-value gives a smaller, or at least no larger, score, so small q_i still means stronger evidence. A score-threshold rule therefore rejects i when $q_i \le s$, that is, when $R_i(s) = \mathbf{1}\{\phi_i(p_i) \le s\}$.

Validity depends on how the score transformation is chosen. In this simple template, ϕ_i is fixed before the procedure uses p_i. If the score transformation is learned from other p-values or from the whole dataset, then the validity proof must account for that learning step, for example through sample splitting, masking, independence, or a direct proof of the aggregate null-evidence bound.

After the ranking score is fixed, we still need a budget $\widehat{M}(s)$. One simple choice is $mg(s)$, where the curve g is chosen before testing. Another choice uses the mirror idea from Chapter 5:

estimate the left-tail null count by looking at very large p-values. Under a well-calibrated null, p-values near 0 and p-values near 1 are both rare tail events. Thus the right tail can sometimes be used as a conservative guide to how many null p-values might fall into the left tail. A large p-value p_j becomes small after the transformation $1 - p_j$, so the count

$$1 + \sum_{j=1}^{m} \mathbf{1}\{\phi_j(1 - p_j) \le s\}$$

counts right-tail observations on the same score scale used for left-tail rejections. The $+1$ is a conservative correction that keeps the estimated budget from being zero. This right-tail idea is useful only when the model or assumptions justify the comparison between the two tails; it is not a free validity guarantee. The raw-p-value right-tail row in the table is the special case $\phi_i(u) = u$.

TABLE 1. Self-contained examples of the threshold template. The customized rows use the score $q_i = \phi_i(p_i)$ defined in the text above the table. In the Storey row, $\widehat{\pi}_0^{\lambda}$ is the null-proportion estimator at tuning parameter λ, recalled from Chapter 6.

Procedure	$\widehat{M}(s)$	$R_i(s)$
BH	ms	$\mathbf{1}\{p_i \le s\}$
Storey	$m\widehat{\pi}_0^{\lambda} s$	$\mathbf{1}\{p_i \le s\},\ s \le \lambda$
Raw p-values, right-tail budget	$1 + \sum_{j=1}^{m} \mathbf{1}\{p_j \ge 1 - s\}$	$\mathbf{1}\{p_i \le s\},\ s < 1/2$
Transformed score, fixed budget	$mg(s)$	$\mathbf{1}\{\phi_i(p_i) \le s\}$
Transformed score, right-tail budget	$1 + \sum_{j=1}^{m} \mathbf{1}\{\phi_j(1 - p_j) \le s\}$	$\mathbf{1}\{\phi_i(p_i) \le s\}$

The useful message, developed by Li and Zhang, is that many p-value threshold rules can be translated into induced e-values once they are written through $R_i(s)$ and $\widehat{M}(s)$ [83, 84]. The remaining mathematical task is to prove the aggregate null-evidence bound

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \le m.$$

Different constructions need different arguments: independent-null leave-one-out arguments for BH-like rules, symmetry or conservativeness arguments for right-tail estimates, and problem-specific arguments for transformed scores.

What to check in your own research problem:

- What is $R_i(s)$, the rejection indicator at a fixed candidate threshold?
- What is $\widehat{M}(s)$, the estimated or upper-bounding number of false discoveries?
- Can you prove the aggregate bound for $E_i = mR_i(\widehat{t})/\widehat{M}(\widehat{t})$, not only for fixed thresholds?

8. Combining and Assembling E-Values

The e-BH proof uses only aggregate null evidence. This makes it possible to combine several evidence sources without redoing the whole FDR proof each time. There are two different problems. The practical question is whether the same hypothesis appears in every evidence source.

Before separating the two problems, define the error rates used when hypotheses are split into groups. Suppose

$$\mathcal{G}_1, \ldots, \mathcal{G}_L$$

are disjoint groups of hypotheses. If a procedure returns a rejection set $\mathcal{R}_\ell \subseteq \mathcal{G}_\ell$ inside group ℓ, define

$$R_\ell = |\mathcal{R}_\ell|, \qquad V_\ell = |\mathcal{R}_\ell \cap \mathcal{I}_0|, \qquad \mathrm{FDR}_\ell = \mathbb{E}\left[\frac{V_\ell}{R_\ell \vee 1}\right].$$

Controlling groupwise FDR at level α means $\mathrm{FDR}_\ell \le \alpha$ for every group ℓ. If a procedure instead returns one pooled rejection set $\mathcal{R}^{\mathrm{pool}}$ across all groups, define

$$R^{\mathrm{pool}} = |\mathcal{R}^{\mathrm{pool}}|, \qquad V^{\mathrm{pool}} = |\mathcal{R}^{\mathrm{pool}} \cap \mathcal{I}_0|, \qquad \mathrm{FDR}^{\mathrm{pool}} = \mathbb{E}\left[\frac{V^{\mathrm{pool}}}{R^{\mathrm{pool}} \vee 1}\right].$$

Controlling overall FDR means $\mathrm{FDR}^{\mathrm{pool}} \le \alpha$. In an assembling problem, the target is this pooled guarantee. The groupwise definition is still useful because the assumptions we impose on each family's e-values also imply groupwise FDR control if the analyst reports family-specific e-BH lists.

Combining means that several analyses speak about the same list $H_1, \dots, H_m$. For example, a genomics study may test the same m genes using a model-based e-value, a permutation-based e-value, and an e-value that borrows pathway information. A clinical monitoring problem may test the same m endpoints using evidence from a primary model and evidence from a robust sensitivity analysis. The final output is still one rejection list among the original m hypotheses; the question is how much weight to give each evidence source.

Assembling means that different analyses cover different hypotheses, and the researcher wants one pooled list at the end. For example, three laboratories may screen disjoint gene panels, or several hospitals may monitor different sets of safety signals. Each laboratory or hospital can construct valid evidence for its own family, but the final scientific action may require ranking all discoveries together for follow-up. The bookkeeping must then account for the family sizes m_ℓ and for the fact that true nulls may be concentrated in the families that receive large weights.

Combining procedures on the same hypotheses. Suppose L procedures all analyze the same m hypotheses. Procedure ℓ produces a nonnegative evidence vector

$$E_1^{(\ell)}, \dots, E_m^{(\ell)}$$

and has already been proved aggregate-valid, meaning that its total expected true-null evidence is at most the number of hypotheses:

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i^{(\ell)}\right] \le m, \qquad \ell = 1, \dots, L.$$

This condition already gives a guarantee for each procedure separately: e-BH applied to $E_1^{(\ell)}, \dots, E_m^{(\ell)}$ controls FDR for procedure ℓ's rejection list. The useful question is whether we can combine the L vectors first and then run e-BH once. The next proposition says yes: with suitable weights, the combined vector is still aggregate-valid, so e-BH on the combined e-values controls FDR for one final rejection list.

Proposition 8.13 (Combining aggregate-valid evidence). *Let*

$$E_i = \sum_{\ell=1}^{L} w_{\ell i} E_i^{(\ell)}, \qquad w_{\ell i} \ge 0.$$

If

$$\sum_{\ell=1}^{L} \max_{1 \le i \le m} w_{\ell i} \le 1,$$

then the combined vector satisfies

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \le m.$$

Consequently, e-BH applied to $E_1, \dots, E_m$ controls FDR at level α.

PROOF. Using linearity of expectation and nonnegative weights,

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] &= \sum_{\ell=1}^{L}\sum_{i\in\mathcal{I}_0} w_{\ell i}\mathbb{E}[E_i^{(\ell)}] \\ &\le \sum_{\ell=1}^{L}\left(\max_i w_{\ell i}\right)\sum_{i\in\mathcal{I}_0}\mathbb{E}[E_i^{(\ell)}] \\ &\le \sum_{\ell=1}^{L}\left(\max_i w_{\ell i}\right) m \le m.\end{aligned}$$

The FDR conclusion follows from Theorem 8.10. □

The max-weight condition is the price of not knowing which hypotheses are true nulls. An average weight can look small over all hypotheses while placing too much weight on the true nulls. The maximum prevents this, so the combined e-values remain valid for a single e-BH run.

Assembling disjoint datasets. Now suppose the hypotheses are split across disjoint families

$$\mathcal{G}_1,\dots,\mathcal{G}_L, \qquad |\mathcal{G}_\ell| = m_\ell, \qquad \sum_{\ell=1}^{L} m_\ell = m.$$

Family ℓ produces evidence $E_i^{(\ell)}$ only for $i \in \mathcal{G}_\ell$, and assume that within that family

$$\mathbb{E}\left[\sum_{i\in\mathcal{G}_\ell\cap\mathcal{I}_0} E_i^{(\ell)}\right] \le m_\ell.$$

This displayed bound has an immediate consequence: if family ℓ runs e-BH only on $\{E_i^{(\ell)} : i \in \mathcal{G}_\ell\}$, then Theorem 8.10 with $m = m_\ell$ gives $\mathrm{FDR}_\ell \le \alpha$. The assembling step builds one pooled list across all families. To do this, choose fixed nonnegative weights, form $E_i = w_{\ell i}E_i^{(\ell)}$ for $i \in \mathcal{G}_\ell$, and run e-BH once on all m assembled e-values. The next proposition gives the condition that makes this pooled run control $\mathrm{FDR}^{\mathrm{pool}}$.

PROPOSITION 8.14 (Assembling disjoint families). *For $i \in \mathcal{G}_\ell$, set*

$$E_i = w_{\ell i}E_i^{(\ell)}, \qquad w_{\ell i} \ge 0.$$

If

$$\sum_{\ell=1}^{L} m_\ell \max_{i\in\mathcal{G}_\ell} w_{\ell i} \le m,$$

then the pooled vector satisfies

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \le m.$$

Thus e-BH controls $\mathrm{FDR}^{\mathrm{pool}}$ *for the pooled rejection list.*

PROOF.

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] &= \sum_{\ell=1}^{L}\sum_{i\in\mathcal{G}_\ell\cap\mathcal{I}_0} w_{\ell i}\mathbb{E}[E_i^{(\ell)}] \\ &\le \sum_{\ell=1}^{L}\left(\max_{i\in\mathcal{G}_\ell} w_{\ell i}\right)\sum_{i\in\mathcal{G}_\ell\cap\mathcal{I}_0}\mathbb{E}[E_i^{(\ell)}] \\ &\le \sum_{\ell=1}^{L} m_\ell \max_{i\in\mathcal{G}_\ell} w_{\ell i} \le m.\end{aligned}$$

□

The proposition proves the assembling result: the one pooled e-BH run controls overall FDR for the pooled rejection list. The family-level aggregate bound imposed before the proposition has the separate consequence that family-specific e-BH runs would control FDR_ℓ. These are different rejection lists; the pooled guarantee does not require reporting the family-specific lists.

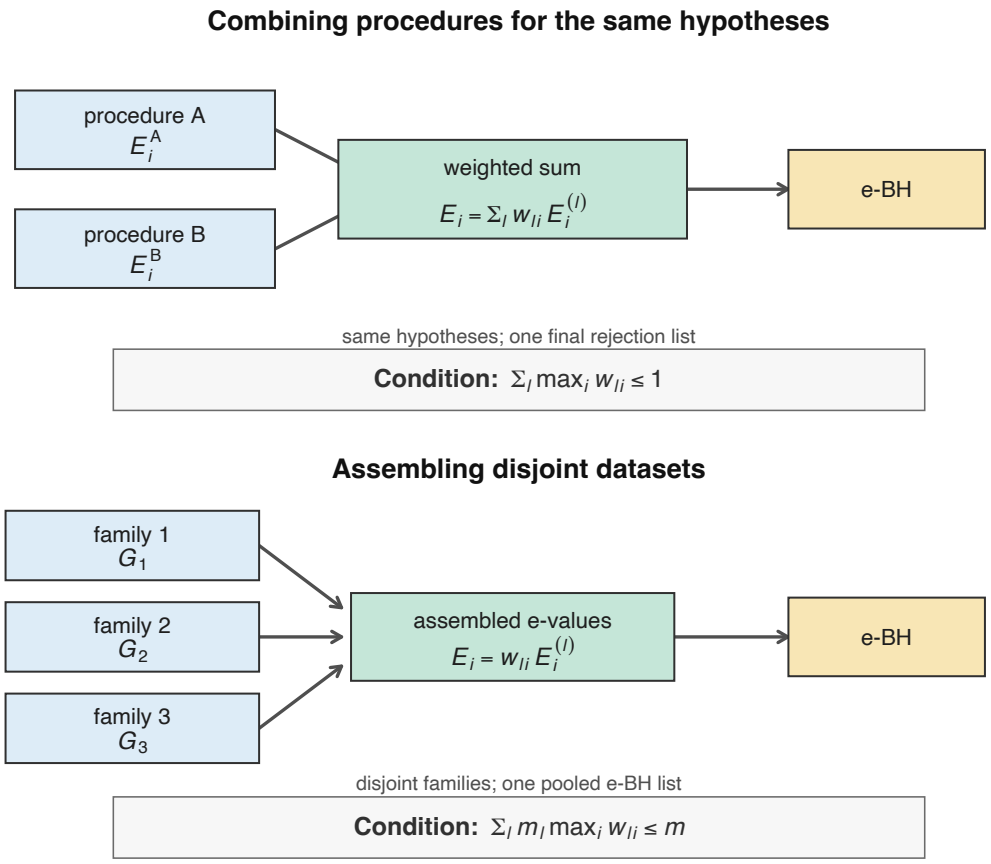

FIGURE 6. Combining evidence from several procedures on the same hypotheses and assembling evidence from disjoint datasets. In assembling, the main output is one pooled e-BH list; family-level aggregate bounds also justify separate family lists when reported. The max-weight conditions protect against unknown true-null placement.

What to check in your own research problem:

- Are you combining several procedures on the same hypotheses, or assembling disjoint families?
- For assembling, can each family prove its own aggregate evidence bound, and can the assembled weights prove the pooled aggregate bound?
- Is the intended output the pooled list? If you also report family-specific lists, state their groupwise FDR guarantees separately.
- Are the weights fixed or chosen using data? If they are data-dependent, what part of the data is allowed to influence them?

9. Online FDR Control

The procedures above handle a fixed family of m hypotheses. Online testing is different. Hypotheses arrive one at a time, and the decision on H_t must be made before seeing future hypotheses. This is the right abstraction for monitoring many experiments, testing safety signals as they appear, or screening a long sequence of scientific questions.

Let $\mathcal{F}_t$ be the information available after the first t tests. At time t, the procedure may use $\mathcal{F}_{t-1}$ to choose how hard to test H_t, but it may not use the current p-value or e-value before choosing the level. Let

$$I_t = \mathbf{1}\{H_t \text{ is rejected}\}, \qquad R_t = \sum_{s=1}^{t} I_s, \qquad V_t = \sum_{s \le t : s \in \mathcal{I}_0} I_s.$$

Here $\mathcal{I}_0 \subseteq \{1, 2, \ldots\}$ denotes the true null times. For a deterministic finite horizon K, the online FDR is

$$\mathrm{FDR}(K) = \mathbb{E}\left[\frac{V_K}{R_K \vee 1}\right].$$

For a data-dependent monitoring time τ, the stopped online FDR is

$$\mathrm{FDR}(\tau) = \mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right].$$

A finite-horizon guarantee asks for $\mathrm{FDR}(K) \le \alpha$ for each fixed K. A stopping-time guarantee asks for the same bound when the time at which we look can depend on the past.

The common design template is simple. Before seeing the current evidence, choose a predictable testing level α_t. For a p-value, reject when $p_t \le \alpha_t$. For an e-value, reject when $E_t \ge 1/\alpha_t$, with the convention that no rejection is made if $\alpha_t = 0$. The problem is to choose the sequence α_t so that early false discoveries do not spend the entire error budget, while genuine discoveries allow the procedure to keep testing with useful power.

P-value online rules: alpha wealth, LORD, and SAFFRON. For p-values, the basic null condition is conditional super-uniformity: for each true null t,

$$\mathbb{P}(p_t \le u \mid \mathcal{F}_{t-1}) \le u, \qquad 0 \le u \le 1.$$

Here predictable means that α_t is chosen from the past information $\mathcal{F}_{t-1}$: it may depend on previous p-values, e-values, rejections, and covariates already revealed, but not on the current p-value p_t before H_t is tested. If H_t is tested at a predictable level α_t, then

$$\mathbb{P}(I_t = 1 \mid \mathcal{F}_{t-1}) \le \alpha_t$$

for a true null. This is the basic one-step inequality. Online FDR rules turn it into a many-step guarantee by deciding how much level can be spent at each time.

Alpha wealth is bookkeeping for this spending. The wealth W_t is not a new error rate; it is an internal balance that limits future testing levels. Start with wealth $W_0 \le \alpha$. Testing at

level α_t spends some wealth. A rejection earns a reward, which permits more testing later. A simplified update is

$$W_t = W_{t-1} - \alpha_t + b_t I_t,$$

where $b_t \geq 0$ is a predictable reward. The statistical idea is: spend a small amount of Type I error budget now, and allow larger future levels only after past discoveries have paid for them. Different online FDR rules differ mainly in how they choose α_t and the rewards.

LORD is the first main example. It chooses testing levels from two sources: initial wealth and rewards from past discoveries. Choose nonnegative numbers $\gamma_1, \gamma_2, \ldots$ with $\sum_{j\geq 1} \gamma_j = 1$. Think of γ_j as a schedule that spreads one unit of wealth over future tests. If $D_1 < D_2 < \cdots$ denote the discovery times revealed so far, a LORD-style level has the following simplified form:

$$\alpha_t = W_0 \gamma_t + \sum_{D_j < t} b_j \gamma_{t-D_j}.$$

The full LORD rule imposes additional constraints on the rewards b_j so the wealth account never overspends. This displayed formula is predictable because it depends only on previous rejections. Operationally, each discovery releases a new stream of future testing levels.

Proof outline for LORD-type rules. For true nulls,

$$\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] \leq \alpha_t.$$

Therefore, for any deterministic horizon K,

$$\mathbb{E}[V_K] \leq \mathbb{E}\left[\sum_{t\leq K : t\in\mathcal{I}_0} \alpha_t\right] \leq \mathbb{E}\left[\sum_{t\leq K} \alpha_t\right].$$

The LORD design makes the cumulative spending $\sum_{t\leq K} \alpha_t$ controlled by initial wealth plus rewards from R_K discoveries. This is why discoveries matter in the denominator of FDR: if there are few discoveries, the rule has little reward and cannot keep spending large levels; if there are many discoveries, the denominator $R_K \vee 1$ is larger. A full LORD theorem turns this accounting into a bound for $\mathbb{E}[V_K/(R_K \vee 1)]$, not just for $\mathbb{E}[V_K]$. Standard LORD FDR theorems also impose assumptions on how true-null p-values relate to the past, commonly independence or conditional super-uniformity strong enough for the rejection history being used. The proof has two moving parts: the p-value assumption controls false rejections at predictable levels, and the wealth equation prevents those predictable levels from growing too fast without discoveries.

SAFFRON is the next idea after LORD. LORD treats every tested hypothesis as a place where alpha might have been wasted. SAFFRON tries to estimate that waste online. Before seeing p_t, choose a predictable candidate threshold λ_t with $\alpha_t \leq \lambda_t \leq 1$, and define

$$C_t = \mathbf{1}\{p_t \leq \lambda_t\}.$$

A candidate is not necessarily rejected; it is a hypothesis whose p-value is small enough to be worth counting as a serious opportunity for rejection. Very large p-values are not candidates. SAFFRON uses the number of candidates and discoveries to estimate how much alpha was effectively spent on nulls, and then reallocates the saved budget to later tests. This is the online analogue of Storey's offline idea: estimate how many hypotheses look null, then use the saved budget for power.

For implementation, track four objects at every time: the predictable test level α_t, the predictable candidate threshold λ_t, the candidate indicator C_t, and the rejection indicator $I_t = \mathbf{1}\{p_t \leq \alpha_t\}$. The full SAFFRON formula specifies how candidates reduce the estimate of wasted alpha and how discoveries replenish wealth. Its proof follows the LORD template, but replaces raw spending $\sum \alpha_t$ by an online estimate of alpha spent on true null candidates.

E-value online rules: a clean e-LOND theorem. E-values give a particularly transparent online theorem. The p-value rules above must carefully ration α_t. For e-values, the test $E_t \geq 1/\alpha_t$ is controlled directly by conditional expectation. We do not need independence between the current evidence and the past; we need only conditional e-value validity. The resulting rule is called e-LOND, the e-value analogue of Levels based On Number of Discoveries.

The design is deliberately simple. Choose a deterministic spending schedule γ_t with total mass at most one. At time t, enlarge the level in proportion to the number of previous discoveries:

$$\alpha_t = \alpha\gamma_t(R_{t-1} + 1).$$

Then reject large e-values. More past discoveries permit a larger current level, but the factor γ_t keeps the total evidence budget summable.

THEOREM 8.15 (A simple e-LOND bound). *Let E_t be nonnegative and $\mathcal{F}_t$-measurable. For every true null t, assume*

$$\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \leq 1.$$

Let $\gamma_t \geq 0$ be deterministic with $\sum_{t\geq 1} \gamma_t \leq 1$. Define the predictable testing level

$$\alpha_t = \alpha\,\gamma_t\,(R_{t-1} + 1).$$

Reject H_t when $E_t \geq 1/\alpha_t$, with no rejection if $\alpha_t = 0$. Then for every stopping time τ,

$$\mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right] \leq \alpha.$$

PROOF. Let $I_t = \mathbf{1}\{E_t \geq 1/\alpha_t\}$. If $t \leq \tau$ and $I_t = 1$, then the final number of rejections by time τ is at least $R_{t-1} + 1$. Therefore,

$$\frac{V_\tau}{R_\tau \vee 1} \leq \sum_{t\in\mathcal{I}_0} \frac{\mathbf{1}\{t \leq \tau\}I_t}{R_{t-1} + 1}.$$

The rejection event implies $E_t \geq 1/\alpha_t$, so, path by path,

$$I_t \leq \alpha_t E_t$$

when $\alpha_t > 0$; when $\alpha_t = 0$, both sides are zero by convention. The factor

$$\frac{\mathbf{1}\{t \leq \tau\}\alpha_t}{R_{t-1} + 1} = \mathbf{1}\{t \leq \tau\}\alpha\gamma_t$$

is $\mathcal{F}_{t-1}$-measurable. Indeed, for a stopping time, $\{t \leq \tau\} = \{\tau \geq t\}$ is determined by whether the procedure had already stopped before time t, and α_t, R_{t-1}, and γ_t are predictable. Using Tonelli's theorem for the nonnegative sum, followed by the tower property and conditional e-value validity, gives

$$\begin{aligned}
\mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right] &\leq \sum_{t\in\mathcal{I}_0} \mathbb{E}\left[\mathbf{1}\{t \leq \tau\}\frac{\alpha_t}{R_{t-1} + 1}E_t\right] \\
&= \sum_{t\in\mathcal{I}_0} \mathbb{E}\left[\mathbf{1}\{t \leq \tau\}\alpha\gamma_t\mathbb{E}(E_t \mid \mathcal{F}_{t-1})\right] \\
&\leq \sum_{t\in\mathcal{I}_0} \mathbb{E}[\mathbf{1}\{t \leq \tau\}\alpha\gamma_t] \leq \alpha\sum_{t\geq 1}\gamma_t \leq \alpha.
\end{aligned}$$

□

What to check in your own research problem: are the testing levels chosen before seeing the current p-value or e-value? For p-value rules, verify conditional super-uniformity under the allowed dependence. For e-value rules, verify conditional e-value validity under each true null and state the intended horizon.

10. Confidence Sequences

Testing asks whether a null value should be rejected. Estimation asks which parameter values remain plausible. In a monitored experiment, the interval should be valid no matter when we look.

DEFINITION 8.16 (Confidence sequence). Let θ be a target parameter. A sequence of sets $(C_t)_{t\geq 1}$ is a $(1-\alpha)$ confidence sequence if

$$\mathbb{P}_\theta\{\theta \in C_t \text{ for every } t \geq 1\} \geq 1-\alpha.$$

The universal quantifier over t is the key difference from ordinary fixed-time confidence intervals. A 95% interval at each fixed time does not imply 95% simultaneous coverage over all times.

THEOREM 8.17 (Confidence sequence by inverting e-processes). *For each candidate value θ_0, suppose $(M_t^{\theta_0})_{t\geq 0}$ is a nonnegative supermartingale under the null model $H_0 : \theta = \theta_0$, with $M_0^{\theta_0} = 1$. Define*

$$C_t = \{\theta_0 : M_t^{\theta_0} < 1/\alpha\}.$$

Then $(C_t)_{t\geq 1}$ is a $(1-\alpha)$ confidence sequence.

PROOF. The true value θ is excluded at some time if and only if

$$M_t^\theta \geq 1/\alpha$$

for at least one t. Since M_t^θ is a nonnegative supermartingale under the true distribution, Ville's inequality gives

$$\mathbb{P}_\theta\left(\sup_{t\geq 1} M_t^\theta \geq 1/\alpha\right) \leq \alpha.$$

Thus the probability of ever excluding the true θ is at most α. □

EXAMPLE 8.18 (A simple bounded-mean confidence sequence). Let $X_1, X_2, \ldots \in [0,1]$ be independent with common mean θ. The fixed-time form of Hoeffding's inequality says that, for each t and each $\eta_t \in (0,1)$,

$$\mathbb{P}_\theta\left(|\bar{X}_t - \theta| > \sqrt{\frac{\log\{2/\eta_t\}}{2t}}\right) \leq \eta_t.$$

Choose $\eta_t = 6\alpha/(\pi^2 t^2)$, so that $\sum_{t=1}^\infty \eta_t = \alpha$. By the union bound,

$$C_t = \left[\bar{X}_t - \sqrt{\frac{\log\{\pi^2 t^2/(3\alpha)\}}{2t}}, \bar{X}_t + \sqrt{\frac{\log\{\pi^2 t^2/(3\alpha)\}}{2t}}\right] \cap [0,1]$$

is a $(1-\alpha)$ confidence sequence. This construction is conservative, but it is fully self-contained and shows the simultaneous-coverage idea.

Betting-style confidence sequences use the e-process inversion theorem more directly. For bounded means, fix a candidate value $\theta_0 \in (0,1)$ and choose predictable bets

$$\lambda_t \in \left[-\frac{1}{1-\theta_0}, \frac{1}{\theta_0}\right].$$

Define

$$M_t^{\theta_0} = \prod_{s=1}^t \{1 + \lambda_s(X_s - \theta_0)\}.$$

The bounds on λ_s keep each factor nonnegative for every $X_s \in [0,1]$. Under $\theta = \theta_0$,

$$\mathbb{E}[1 + \lambda_s(X_s - \theta_0) \mid \mathcal{F}_{s-1}] = 1,$$

so $M_t^{\theta_0}$ is a nonnegative martingale. Inverting these processes over θ_0 gives a confidence sequence. More refined choices of the bets adapt to the observed variance and often give much shorter intervals than the simple union-bound construction.

What to check in your own research problem:

- What null $H_0 : \theta = \theta_0$ is tested for each candidate value?
- What e-process $M_t^{\theta_0}$ is valid under that null?
- Is the reported set C_t obtained by inverting the threshold $M_t^{\theta_0} < 1/\alpha$?
- Is the claim simultaneous over time, or only valid at a prespecified time?

11. Assumptions in Plain Language

The e-value condition is a null expectation bound. For simple nulls, likelihood ratios give canonical examples. For composite nulls, the bound must hold uniformly over the whole null model. A Bayes factor averaged over a null prior is not enough unless it also satisfies the uniform expectation bound.

Safe testing requires a process-level guarantee. A fixed-sample e-value can be used once at a prespecified time. An e-process can be monitored, stopped, and continued because it is a nonnegative supermartingale under the null.

E-BH controls FDR under arbitrary dependence because its proof reduces FDP to aggregate null evidence before taking expectations. Dependence can still affect power and construction of good e-values, but it does not enter the e-BH proof once the aggregate evidence bound is established.

Threshold-to-eBH arguments are design tools. They do not make a procedure valid by algebra alone. The hard step is always the same: prove that the induced evidence vector has acceptable aggregate null expectation.

Online p-value procedures rely on predictable levels and conditional super-uniformity. Online e-value procedures replace conditional super-uniformity by conditional e-value validity. Confidence sequences are the interval analogue of e-processes: invert a family of anytime-valid tests and obtain simultaneous coverage over time.

12. Bibliographic Notes

E-values, calibration, and merging are developed by Vovk and Wang [127]. Safe testing and optional continuation are treated by Grünwald et al. [54]; the betting interpretation is emphasized by Shafer [105] and connected to the broader game-theoretic view by Ramdas et al. [99]. E-BH and its arbitrary-dependence FDR proof are due to Wang and Ramdas [129]. The threshold-to-eBH viewpoint and the aggregation and assembling ideas are based on Li and Zhang [84] and Li and Zhang [83].

Online FDR control begins with alpha-investing [49] and includes LORD [69] and SAFFRON [98]. Online e-value rules such as e-LOND are developed by Xu and Ramdas [131]. Confidence sequences in the modern nonasymptotic form are developed by Howard et al. [62]; betting confidence sequences for bounded means are developed by Waudby-Smith and Ramdas [130].

13. Exercises

Basic.

EXERCISE 8.19 (Markov calibration). Let E be an e-value. Prove that $p_E = \min\{1, 1/E\}$ is a valid p-value. Then give an example showing why $1/p$ is not generally an e-value when $p \sim \mathrm{Unif}(0,1)$.

EXERCISE 8.20 (Gaussian likelihood-ratio e-value). Work through Example 8.4. Derive the likelihood ratio from the normal densities, verify that its null expectation is one, and compute the expected log e-value under $\mu = \mu_1$.

EXERCISE 8.21 (Composite one-sided null). For $X \sim N(\mu, 1)$, test $H_0 : \mu \le 0$ against the design alternative $\mu = \mu_1 > 0$. Verify that

$$E(X) = \exp\{\mu_1 X - \mu_1^2/2\}$$

satisfies $\mathbb{E}_{P_\mu}[E] \le 1$ for every $\mu \le 0$.

EXERCISE 8.22 (Averaging e-values). Suppose $E_1, \dots, E_L$ are e-values for the same null and $a_\ell \ge 0$, $\sum_\ell a_\ell \le 1$. Prove that $\sum_\ell a_\ell E_\ell$ is an e-value.

Intermediate.

EXERCISE 8.23 (Conditional products). Let E_t be nonnegative and $\mathcal{F}_t$-measurable with $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \le 1$. Prove that $M_t = \prod_{s=1}^t E_s$ is a nonnegative supermartingale. Identify exactly where measurability is used.

EXERCISE 8.24 (Ville's inequality). Prove Ville's inequality for a finite deterministic horizon K. Then pass to the infinite horizon by monotone convergence of the crossing events.

EXERCISE 8.25 (E-BH proof). Reproduce the proof of Theorem 8.10. Explain why the proof does not require independence among the e-values.

EXERCISE 8.26 (BH as e-BH). Let $\widehat{t}_{\mathrm{BH}}$ be the BH threshold and define

$$E_i^{\mathrm{BH}} = \widehat{t}_{\mathrm{BH}}^{-1} \mathbf{1}\{p_i \le \widehat{t}_{\mathrm{BH}}\}$$

when $\widehat{t}_{\mathrm{BH}} > 0$, with all induced e-values equal to zero otherwise. Show that e-BH applied to these induced e-values gives the BH rejection set. Which part of the argument is algebraic, and which part is a validity proof?

EXERCISE 8.27 (General threshold template). For a threshold procedure with rejection indicators $R_i(s)$ and estimated false-discovery count $\widehat{M}(s)$, derive the induced e-values

$$E_i = \frac{m R_i(\widehat{t})}{\widehat{M}(\widehat{t})}.$$

State the positive-budget condition needed to avoid division by zero, and explain what aggregate bound must be proved before e-BH can be used for FDR control.

EXERCISE 8.28 (Combining procedures). Suppose two aggregate-valid procedures produce evidence vectors $\{E_i^{(1)}\}$ and $\{E_i^{(2)}\}$. Choose nonconstant weights w_{1i}, w_{2i} satisfying

$$\max_i w_{1i} + \max_i w_{2i} \le 1.$$

Verify the aggregate bound for the combined vector. Construct an example where replacing the max-weight condition by an average-weight condition would overweight the true nulls.

EXERCISE 8.29 (Assembling datasets). Let three disjoint studies have sizes m_1, m_2, m_3. Propose weights that prioritize the first study while satisfying

$$\sum_{\ell=1}^{3} m_\ell \max_{i \in \mathcal{G}_\ell} w_{\ell i} \le m.$$

First verify that the study-level aggregate bounds would allow separate study-level e-BH runs to control groupwise FDR if those lists were reported. Then focus on the assembling goal: prove that the weighted assembled e-values satisfy the pooled aggregate bound, so one pooled e-BH run controls overall FDR.

Computational.

EXERCISE 8.30 (Optional stopping simulation). Reproduce Figure 3. Simulate null Gaussian data, repeatedly check a one-sided fixed-sample p-value, and compare the probability of ever crossing 0.05 with the probability that a Gaussian likelihood-ratio e-process crosses 20.

EXERCISE 8.31 (Dependence simulation). Implement BH, BY, and e-BH for equicorrelated one-sided Gaussian tests. Vary the correlation, signal strength, and number of nonnulls. Report FDR and power, and explain why e-BH's validity statement differs from BH's PRDS validity statement.

EXERCISE 8.32 (LORD and e-LOND simulation). Simulate an online stream with a mixture of null and nonnull hypotheses. Implement a simple LORD-style p-value rule and the e-LOND rule in Theorem 8.15. Track R_t, V_t, and $V_t/(R_t \vee 1)$ over time.

EXERCISE 8.33 (SAFFRON-style candidates). Modify the previous simulation by adding candidate indicators $C_t = \mathbf{1}\{p_t \le \lambda\}$. Compare a rule that spends budget on all tests with a rule that spends more budget after many candidate tests are not rejected. Explain how this illustrates the Storey-like idea behind SAFFRON.

Advanced.

EXERCISE 8.34 (Leave-one-out aggregate bound). Under independent super-uniform null p-values, prove the aggregate bound for the BH-induced e-values in Proposition 8.11. Use a leave-one-out argument: replace one rejected true-null p-value by zero and show why the number of BH rejections does not change on the relevant event.

EXERCISE 8.35 (Flexible threshold rule). Let $q_i = \phi_i(p_i)$, where each ϕ_i is nondecreasing, and let a procedure reject $q_i \le \widehat{t}$ with

$$\widehat{t} = \sup\{s : mg(s)/(R(s) \vee 1) \le \alpha\}.$$

Write the induced e-values and state a sufficient aggregate condition for FDR control. Give the weighted BH special case.

EXERCISE 8.36 (A simple e-LOND proof variant). In Theorem 8.15, replace α_t by $\alpha\gamma_t(a+R_{t-1})$ for a constant $a > 0$. Determine how the proof changes and what condition on a keeps the same FDR bound.

EXERCISE 8.37 (Confidence sequence by inversion). For each θ_0, let $M_t^{\theta_0}$ be a nonnegative supermartingale under $\mathbb{P}_{\theta_0}$ with $M_0^{\theta_0} = 1$. Prove Theorem 8.17. Then explain why simultaneous coverage implies coverage at any stopping time τ.

EXERCISE 8.38 (Hoeffding confidence sequence). Verify Example 8.18. Compare its half-width with the fixed-time Hoeffding interval at a prespecified time t, and explain the price paid for simultaneous validity.

CHAPTER 9

Conditional Randomization and Knockoff Filters

In a regression or supervised learning problem, a feature can be strongly associated with the response for two different reasons. It may carry information that is not available in the other features, or it may merely be correlated with another feature that carries the information. The distinction is central in high-dimensional variable selection. A marginal test answers the question "is X_j associated with Y?" A variable-selection procedure usually wants to answer the more local question "does X_j still matter after X_{-j} is known?"

This chapter studies two model-X tools for that conditional question. The conditional randomization test (CRT) is the direct benchmark: resample the candidate feature from its conditional law given all the other features, and compare the observed statistic with the conditional null distribution. Knockoff filters build reusable negative controls for all features at once. They replace many conditional resampling tests by a single augmented design $[X, \widetilde{X}]$ whose symmetry turns negative knockoff wins into an estimate of false discoveries.

The price of this robustness to the response model is clear. CRTs and model-X knockoffs do not require a correct model for $Y \mid X$, but they do require knowledge, or accurate estimation, of the feature distribution. Fixed-X knockoffs make a different tradeoff: they use exact geometry for a linear fixed-design model and therefore do not need a model for the distribution of the rows of X.

1. Conditional Feature Nulls

Let $X = (X_1, \dots, X_p)$ denote the feature vector and Y the response. For a sample of size n, X_j will also denote the n-vector of observed values of feature j, and X_{-j} the matrix of all remaining features. The model-X null for feature j is

$$H_j : \quad Y \perp\!\!\!\perp X_j \mid X_{-j}.$$

This says that once the other features are known, replacing X_j by another draw from its conditional distribution should not change the distribution of the response.

Definition 9.1 (Conditional feature null). Feature j is conditionally null if

$$\mathcal{L}(Y \mid X_j, X_{-j}) = \mathcal{L}(Y \mid X_{-j})$$

almost surely. Equivalently, $Y \perp\!\!\!\perp X_j \mid X_{-j}$.

In the linear model

$$Y = X\beta + \varepsilon, \qquad \varepsilon \perp X, \qquad \mathbb{E}[\varepsilon \mid X] = 0,$$

with a nondegenerate conditional distribution of $X_j \mid X_{-j}$, this conditional null corresponds to $\beta_j = 0$. The definition is more general, however. It allows nonlinear response models, interactions, and arbitrary prediction algorithms as long as the null is interpreted conditionally.

The conditional formulation is stricter than a marginal association statement. If two features are correlated, a null feature may have a visible marginal relationship with the response because it shadows another feature. That is the failure mode that CRTs and knockoffs are designed to avoid.

2. Conditional Randomization Tests

The conditional randomization test of Candès et al. [30] tests one conditional null at a time. Choose a statistic

$$T_j = T_j(X_j, X_{-j}, Y)$$

that becomes large when feature j appears important. The statistic may be a regression coefficient, a lasso entry time, a random-forest importance score, or the drop in predictive accuracy after removing X_j. The validity of the CRT does not come from the statistic. It comes from comparing the observed statistic with statistics computed after resampling X_j from the correct conditional law.

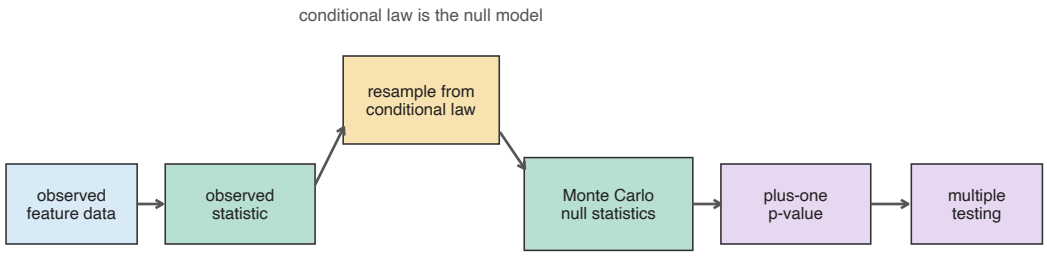

FIGURE 1. Conditional randomization test. The conditional feature law, not a model for $Y \mid X$, supplies the null distribution. The final p-value uses the plus-one Monte Carlo correction.

Given B Monte Carlo draws, the CRT proceeds as follows.

1. Compute $T_j^{\text{obs}} = T_j(X_j, X_{-j}, Y)$.
2. For $b = 1, \dots, B$, draw $X_j^{(b)} \sim \mathcal{L}(X_j \mid X_{-j})$.
3. Compute $T_j^{(b)} = T_j(X_j^{(b)}, X_{-j}, Y)$.
4. Return $p_j = \dfrac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_j^{(b)} \geq T_j^{\text{obs}}\}}{B+1}$.

The p-value is written for large values of T_j. For two-sided or directional statistics, one replaces the comparison by the appropriate tail event.

THEOREM 9.2 (Finite-sample validity of the CRT). *Assume that the conditional law $\mathcal{L}(X_j \mid X_{-j})$ used to resample is correct, and that $H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$ holds. Conditional on (X_{-j}, Y), the observed and resampled statistics $T_j^{\text{obs}}, T_j^{(1)}, \dots, T_j^{(B)}$ are exchangeable. Consequently the plus-one Monte Carlo p-value is super-uniform:*

$$\mathbb{P}(p_j \leq \alpha) \leq \alpha, \qquad 0 \leq \alpha \leq 1.$$

PROOF. Under H_j, the conditional distribution of X_j given (X_{-j}, Y) coincides with $\mathcal{L}(X_j \mid X_{-j})$: the response carries no information about X_j once X_{-j} is fixed. The resampled draws $X_j^{(1)}, \dots, X_j^{(B)}$ come from the same conditional law by construction. Therefore the collection

$$X_j,\ X_j^{(1)}, \dots, X_j^{(B)}$$

is iid conditional on (X_{-j}, Y), and the induced statistics $T_j^{\text{obs}}, T_j^{(1)}, \dots, T_j^{(B)}$ are exchangeable conditional on (X_{-j}, Y). If there are no ties, the rank of T_j^{obs} among the $B+1$ exchangeable statistics is uniform on $\{1, \dots, B+1\}$; the plus-one p-value is the fraction of statistics at least as large as the observed statistic, so it is super-uniform on the grid $\{1/(B+1), \dots, 1\}$. Ties only make the counting rule more conservative. Taking expectations over (X_{-j}, Y) gives the unconditional bound. □

The plus-one correction is not cosmetic. Without it, the smallest possible Monte Carlo p-value would be zero, and exact finite-sample validity would be lost. The correction also makes clear how many null samples are needed if CRT p-values will be fed into a multiple testing procedure. For example, if there are p hypotheses and one hopes to use BH at level q, then B must be large enough that the grid of possible p-values can reach the smallest BH critical values.

3. Why Marginal Resampling Fails

The CRT must resample from $X_j \mid X_{-j}$, not from the marginal law of X_j. The following example is the simplest way to see why.

EXAMPLE 9.3 (A null feature with marginal association). Let

$$(X_1, X_2) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \qquad Y = X_2 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

with ε independent of (X_1, X_2). Then

$$Y \perp\!\!\!\perp X_1 \mid X_2,$$

so feature 1 is conditionally null. Marginally,

$$\operatorname{Cov}(X_1, Y) = \operatorname{Cov}(X_1, X_2) = \rho.$$

Thus X_1 can be strongly correlated with the response even though it adds no information once X_2 is known.

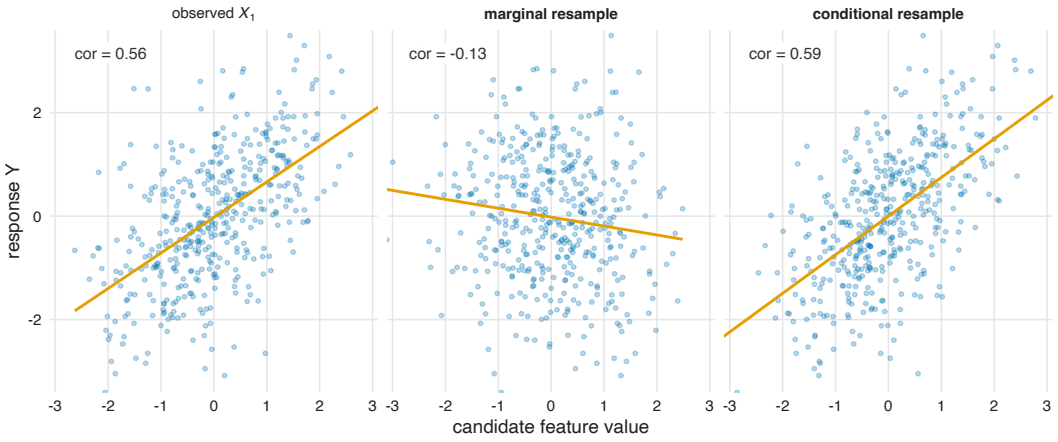

FIGURE 2. A conditionally null feature can be marginally correlated with the response. Marginal resampling breaks the dependence between X_1 and X_2, whereas conditional resampling preserves it.

In Example 9.3, a marginal randomization test would compare the observed association between X_1 and Y with a null distribution obtained by drawing $X_1^* \sim N(0,1)$ independently of

X_2. That null distribution is wrong: it describes a world in which X_1^* no longer tracks X_2. The correct conditional draw is

$$X_1^* \mid X_2 = x_2 \sim N(\rho x_2,\ 1-\rho^2),$$

which preserves the correlation structure that makes X_1 look important marginally.

After valid CRT p-values have been obtained, they can be used with the multiple testing procedures from earlier chapters. The CRT supplies valid p-values for the conditional nulls. The downstream FDR or FWER guarantee then depends on the assumptions of the chosen multiple testing procedure, such as independence, PRDS, or a valid reshaping correction. CRT validity by itself does not make BH valid under arbitrary dependence among the CRT p-values.

4. Fixed-X Knockoffs

We next move from testing one feature at a time to constructing negative controls for all features simultaneously. The original knockoff construction of Barber and Candès [6] is for the fixed-design Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

where $X \in \mathbb{R}^{n\times p}$ is treated as fixed. The null hypothesis is

$$H_j : \beta_j = 0.$$

Assume for the construction below that X has full column rank and that there is room to add p orthogonal directions outside the span of X; the explicit formula below uses $n \geq 2p$. Variants are available when this geometric condition is not met, but the core symmetry is easiest to see in the full-rank case.

Let

$$\Sigma = X^\top X$$

and let $D = \mathrm{diag}(s)$ for a vector $s \in \mathbb{R}^p_+$. A fixed-X knockoff matrix $\widetilde{X} \in \mathbb{R}^{n\times p}$ is designed to satisfy

$$\widetilde{X}^\top \widetilde{X} = \Sigma, \qquad X^\top \widetilde{X} = \Sigma - D.$$

Equivalently,

$$\begin{pmatrix} X^\top X & X^\top \widetilde{X} \\ \widetilde{X}^\top X & \widetilde{X}^\top \widetilde{X} \end{pmatrix} = \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}.$$

Each knockoff column has the same inner products with all other original features as the original column does, except that its inner product with its own original feature is reduced by s_j. When the columns of X are normalized so that $\Sigma_{jj} = 1$, the correlation between X_j and $\widetilde{X}_j$ is $1 - s_j$. Larger s_j gives a more distinguishable negative control and usually improves power.

Proposition 9.4 (Fixed-X knockoff construction). *Suppose X has full column rank, $n - \mathrm{rank}(X) \geq p$, and $D = \mathrm{diag}(s) \succeq 0$ satisfies*

$$2\Sigma - D \succeq 0.$$

Let $U \in \mathbb{R}^{n\times p}$ have orthonormal columns satisfying $U^\top X = 0$, and choose $C \in \mathbb{R}^{p\times p}$ such that

$$C^\top C = 2D - D\Sigma^{-1}D.$$

Then

$$\widetilde{X} = X(I - \Sigma^{-1}D) + UC$$

satisfies

$$X^\top \widetilde{X} = \Sigma - D, \qquad \widetilde{X}^\top \widetilde{X} = \Sigma.$$

PROOF. Since $U^\top X = 0$,

$$X^\top \widetilde{X} = X^\top X(I - \Sigma^{-1}D) = \Sigma - D.$$

Next,

$$\begin{aligned}\widetilde{X}^\top \widetilde{X} &= (I - D\Sigma^{-1})\Sigma(I - \Sigma^{-1}D) + C^\top C \\ &= \Sigma - 2D + D\Sigma^{-1}D + 2D - D\Sigma^{-1}D \\ &= \Sigma.\end{aligned}$$

It remains to check that such a C exists, i.e. that $2D - D\Sigma^{-1}D \succeq 0$. Consider the block Gram matrix

$$G = \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}.$$

On one hand, the congruence transformation $P = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} I & I \\ I & -I \end{smallmatrix}\right)$ brings G to $P^\top G P = \mathrm{diag}(2\Sigma - D,\, D)$, so $G \succeq 0$ iff $2\Sigma - D \succeq 0$ and $D \succeq 0$. On the other hand, the Schur complement of the top-left block gives $G \succeq 0$ iff $\Sigma - (\Sigma - D)\Sigma^{-1}(\Sigma - D) = 2D - D\Sigma^{-1}D \succeq 0$. The first characterization says exactly that the assumed constraints $2\Sigma - D \succeq 0$ and $D \succeq 0$ make the target block Gram matrix positive semidefinite. The Schur-complement characterization then converts that same positive semidefiniteness into $2D - D\Sigma^{-1}D \succeq 0$, so a real matrix C with $C^\top C = 2D - D\Sigma^{-1}D$ exists. □

Figure 3 displays the Gram structure in block form. An analogous block *covariance* matrix will reappear in §7 for Gaussian model-X knockoffs, with Σ reinterpreted as a population covariance rather than the Gram matrix used here.

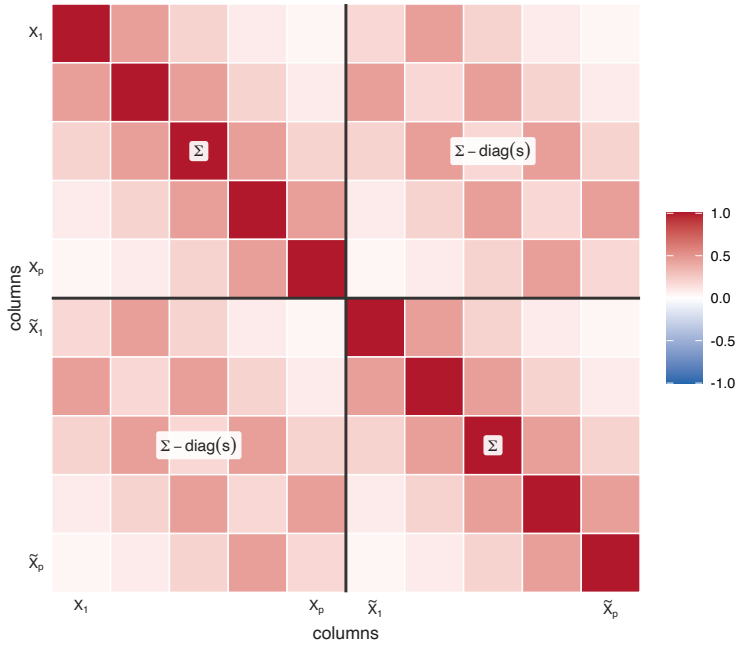

FIGURE 3. The knockoff block structure. Original and knockoff variables have matching within-block dependence; the diagonal gap $\mathrm{diag}(s)$ makes each knockoff a less-than-perfect copy of its original feature. In the fixed-X setting (§4) this is a Gram matrix; in the model-X setting (§7) it is a population covariance.

The vector s is a power parameter. Assume from here that the columns of X are normalized so that $\Sigma_{jj} = 1$; a simple equi-correlated choice is then

$$s_j = \min\{1, 2\lambda_{\min}(\Sigma)\}, \qquad j = 1, \dots, p.$$

A more adaptive choice solves

$$\min_{0 \le s_j \le 1} \sum_{j=1}^{p} |1 - s_j| \qquad \text{subject to} \qquad 2\Sigma - \operatorname{diag}(s) \succeq 0.$$

Both choices express the same principle: make X_j and $\widetilde{X}_j$ as different as the feature correlation structure permits.

5. Knockoff Statistics and Thresholds

After constructing $\widetilde{X}$, fit a variable-selection method to the augmented design

$$[X, \widetilde{X}].$$

The output must be converted into statistics $W_1, \dots, W_p$ satisfying two structural properties.

Definition 9.5 (Sufficiency and antisymmetry). A fixed-X knockoff statistic W_j is sufficient if the vector W depends on the data only through

$$[X, \widetilde{X}]^\top [X, \widetilde{X}], \qquad [X, \widetilde{X}]^\top y.$$

It is antisymmetric if swapping X_j and $\widetilde{X}_j$ flips the sign of W_j and leaves the other statistics unchanged except for the corresponding swap.

The standard example comes from the lasso path. Let Z_j be the largest penalty value at which X_j enters the lasso model fit on $[X, \widetilde{X}]$, and let $\widetilde{Z}_j$ be the analogous entry time for $\widetilde{X}_j$. Define the signed-max statistic

$$W_j = \max\{Z_j, \widetilde{Z}_j\} \operatorname{sign}(Z_j - \widetilde{Z}_j).$$

Large positive W_j means the original feature beats its knockoff strongly. Large negative W_j means the knockoff beats the original feature. For null features, the two events should be symmetric.

Let

$$\mathcal{W} = \{|W_j| : \ |W_j| > 0, \ j = 1, \dots, p\}.$$

The knockoff+ threshold at target FDR level q is

$$T_+ = \min \left\{ t \in \mathcal{W} : \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le q \right\},$$

with $T_+ = \infty$ if the set is empty. The selected variables are

$$\widehat{S} = \{j : W_j \ge T_+\}.$$

A boundary observation: the numerator in the threshold inequality is at least 1, so for the ratio to fall below q we need $\#\{j : W_j \ge t\} \ge 1/q$ at the chosen t. When the target level satisfies $q < 1/p$, *no* value of t can satisfy the inequality, and knockoff+ returns $\widehat{S} = \varnothing$ deterministically. This is the finite-sample analogue of the no-discovery regime in BH at very stringent FDR targets. The version without the plus one,

$$T = \min \left\{ t \in \mathcal{W} : \frac{\#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le q \right\},$$

is often more powerful but controls a slightly modified FDR criterion. The plus-one numerator is what gives ordinary FDR control.

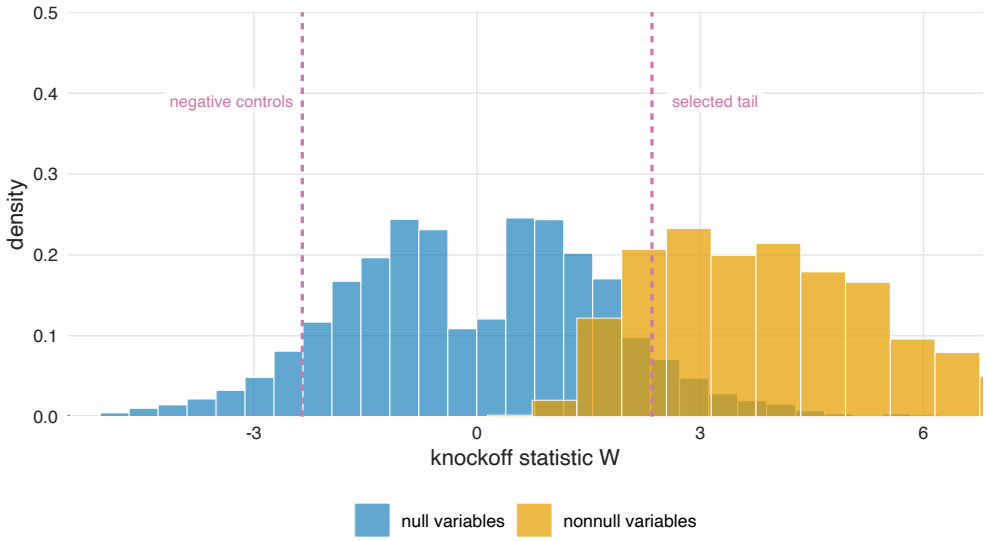

FIGURE 4. Knockoff signs. For true null variables, positive and negative W_j's are symmetric; negative knockoff wins therefore estimate the number of false positive original wins. Nonnull variables should shift mass toward the positive tail.

THEOREM 9.6 (Fixed-X knockoff+ FDR control). *In the fixed-design Gaussian linear model, suppose $\widetilde{X}$ satisfies the fixed-X knockoff Gram equations and W is sufficient and antisymmetric. Then the knockoff+ selection rule satisfies*

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0,\ j \in \widehat{S}\}}{|\widehat{S}| \vee 1}\right] \leq q.$$

PROOF. The Gram equations make each null original feature and its knockoff exchangeable in the Gaussian likelihood, conditional on all nonnull features and on the absolute values of the null statistics. Sufficiency ensures that no other aspect of the fixed design enters the statistic. Antisymmetry then implies that, for null j, the sign of W_j is a fair coin conditional on the magnitudes and on the nonnull statistics; the null signs are also mutually independent given this conditioning.

For a threshold t, write

$$V^+(t) = \#\{j : \beta_j = 0,\ W_j \geq t\}, \qquad V^-(t) = \#\{j : \beta_j = 0,\ W_j \leq -t\}.$$

Let $\mathcal{F}_t$ be the σ-algebra generated by the magnitudes $\{|W_j|\}$, the nonnull statistics, and the null-sign data for thresholds above t. As t decreases through the distinct values $|W_{(1)}| > |W_{(2)}| > \cdots$, the conditional-coin-flip property implies that $V^+(t)/(1 + V^-(t))$ is a supermartingale with respect to the reverse filtration $(\mathcal{F}_t)$. By optional stopping at the data-dependent threshold T_+,

$$\mathbb{E}\left[\frac{V^+(T_+)}{1 + V^-(T_+)}\right] \leq 1.$$

The observable knockoff+ stopping rule guarantees

$$\frac{1 + \#\{j : W_j \leq -T_+\}}{|\widehat{S}| \vee 1} = \frac{1 + \#\{j : W_j \leq -T_+\}}{\#\{j : W_j \geq T_+\} \vee 1} \leq q,$$

and $V^-(T_+) \le \#\{j : W_j \le -T_+\}$ gives $1 + V^-(T_+) \le 1 + \#\{j : W_j \le -T_+\}$. Combining these two facts,

$$\begin{aligned}\mathbb{E}\left[\frac{V^+(T_+)}{|\widehat{S}| \vee 1}\right] &= \mathbb{E}\left[\frac{V^+(T_+)}{1+V^-(T_+)} \cdot \frac{1+V^-(T_+)}{|\widehat{S}| \vee 1}\right] \\ &\le \mathbb{E}\left[\frac{V^+(T_+)}{1+V^-(T_+)} \cdot \frac{1+\#\{j : W_j \le -T_+\}}{|\widehat{S}| \vee 1}\right] \\ &\le q\,\mathbb{E}\left[\frac{V^+(T_+)}{1+V^-(T_+)}\right] \le q.\end{aligned}$$

□

The stopped-ratio step is the only unfamiliar part of the proof. A useful finite-dimensional analogy is this: after conditioning on the magnitudes of the null W_j's, reveal their signs from large $|W_j|$ to small $|W_j|$. At each unrevealed null coordinate the sign is a fair coin, so negative signs are an unbiased proxy for positive null signs. The ratio $V^+(t)/(1+V^-(t))$ is the optional-stopping version of this proxy. The plus one in the denominator prevents division by zero and is exactly what turns the heuristic "negative null wins estimate positive null wins" into a finite-sample FDR bound.

6. Model-X Knockoffs

Fixed-X knockoffs are tied to the linear fixed-design model with target $\beta_j = 0$. Model-X knockoffs change both ingredients. The rows X_i are treated as iid draws from an unknown population distribution, the response model is unrestricted, and the target is the conditional independence null $Y \perp\!\!\!\perp X_j \mid X_{-j}$ from §2 [30].

The two targets are inferentially different even when their algebra looks similar. Under the fixed-X linear model, $\beta_j = 0$ refers to the population-level partial regression coefficient with respect to the columns of the fixed design. Under model-X, $Y \perp X_j \mid X_{-j}$ refers to a distributional statement that does not depend on linearity. They coincide in the special case of a correctly specified Gaussian linear model with random X, but in general one can hold without the other.

Notation also resets here. In §4, $\Sigma = X^\top X$ was the Gram matrix of a deterministic design; in the present section, $\Sigma = \mathrm{Cov}(X)$ is the population covariance of the random feature vector. The algebraic identities that follow are the same, but the statistical meaning is different.

Definition 9.7 (Model-X knockoff). A random vector $\widetilde{X} = (\widetilde{X}_1, \dots, \widetilde{X}_p)$ is a model-X knockoff copy of X if it satisfies:

$$[X, \widetilde{X}]_{\mathrm{swap}(S)} \overset{d}{=} [X, \widetilde{X}] \qquad \text{for every } S \subseteq \{1, \dots, p\},$$

where $\mathrm{swap}(S)$ exchanges X_j and $\widetilde{X}_j$ for all $j \in S$, and

$$\widetilde{X} \perp\!\!\!\perp Y \mid X.$$

For a sample, the construction is applied row by row, or more generally to the joint feature array, without using the responses.

The first condition is pairwise exchangeability. It is stronger than matching marginal distributions. A row permutation or column shuffle generally does not work: it may give $\widetilde{X}_j$ the same marginal distribution as X_j, but it does not preserve the conditional relationship between X_j and X_{-j} for the same observational unit. Knockoffs are coordinatewise negative controls, not arbitrary fake features.

LEMMA 9.8 (Null-swap property). *Let $\widetilde{X}$ be a model-X knockoff copy of X. If*

$$Y \perp\!\!\!\perp X_S \mid X_{-S},$$

then

$$([X, \widetilde{X}], Y) \stackrel{d}{=} ([X, \widetilde{X}]_{\mathrm{swap}(S)}, Y).$$

PROOF. Two ingredients are needed: pairwise exchangeability of $[X, \widetilde{X}]$ and the response-independence condition $\widetilde{X} \perp\!\!\!\perp Y \mid X$ from the model-X knockoff definition.

Pairwise exchangeability gives the equality in distribution for $[X, \widetilde{X}]$ before looking at Y. Response independence says that $\widetilde{X}$ was generated without using Y, conditional on X, so the conditional law $\mathcal{L}(Y \mid X, \widetilde{X})$ equals $\mathcal{L}(Y \mid X)$. The joint null condition $Y \perp\!\!\!\perp X_S \mid X_{-S}$ says that the conditional law of Y given X depends on the features only through X_{-S}. Swapping X_j with its knockoff for $j \in S$ leaves X_{-S} unchanged and therefore does not change the conditional law of the response. Combining the feature exchangeability with this conditional response invariance gives the joint equality. □

For a set S of individually null coordinates, the joint condition in the lemma follows under the usual positivity/intersection assumptions for conditional independence. Without such a regularity condition, individual statements $Y \perp X_j \mid X_{-j}$ need not automatically combine into $Y \perp X_S \mid X_{-S}$, so the subset-swap statement should be read with that caveat.

Once the null-swap property holds, the same sign-flip logic used for fixed-X knockoffs applies. A model-X feature statistic may be produced by any algorithm, including random forests, neural networks, penalized regressions, or screening scores, as long as it has the swap property: swapping $(X_j, \widetilde{X}_j)$ flips the sign of W_j and leaves the other statistics appropriately unchanged.

THEOREM 9.9 (Model-X knockoff+ FDR control). *Assume $\widetilde{X}$ is an exact model-X knockoff copy of X, generated without using Y. Let W be any statistic satisfying the knockoff swap property. Assume also that the conditional law is regular enough that individual nulls combine into joint subset nulls, i.e. for every subset S of true nulls, $Y \perp\!\!\!\perp X_S \mid X_{-S}$. Then the knockoff+ threshold controls FDR for the conditional nulls*

$$H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$$

at level q.

PROOF. By Lemma 9.8, swapping any subset of null originals with their knockoffs leaves the joint distribution of the data and response unchanged. The swap property of W converts these data swaps into sign flips of the corresponding null statistics. Conditional on the magnitudes of the null statistics and on the nonnull statistics, the null signs are therefore exchangeable fair signs. The same stopped-ratio argument used in Theorem 9.6 proves the result. □

7. Gaussian Model-X Knockoffs

For the Gaussian model-X construction of Candès et al. [30], when

$$X \sim N(0, \Sigma),$$

knockoffs are explicit. Choose $D = \mathrm{diag}(s)$ such that

$$D \succeq 0, \qquad 2\Sigma - D \succeq 0.$$

Define the joint Gaussian distribution

$$\begin{pmatrix} X \\ \widetilde{X} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}\right).$$

The block covariance is invariant under swapping any coordinate pair $(X_j, \widetilde{X}_j)$, and therefore the resulting $\widetilde{X}$ is a model-X knockoff copy.

Equivalently, one can sample conditionally:

$$\widetilde{X} \mid X = x \sim N\left((I - D\Sigma^{-1})x,\ 2D - D\Sigma^{-1}D\right),$$

where x is treated as a column vector. The conditional covariance is positive semidefinite precisely under the same feasibility condition $2\Sigma - D \succeq 0$. As in fixed-X knockoffs, larger s_j makes $\widetilde{X}_j$ less correlated with X_j, and hence usually improves the ability to distinguish original features from their knockoffs.

8. SCIP and Structured Feature Laws

Gaussian knockoffs are only one example. For a general known feature distribution, the sequential conditional independent pairs (SCIP) construction of Candès et al. [30] builds knockoffs one coordinate at a time. At step j, it samples

$$\widetilde{X}_j \sim \mathcal{L}(X_j \mid X_{-j}, \widetilde{X}_1, \ldots, \widetilde{X}_{j-1}),$$

where the conditional law is computed under the joint distribution implied by the previous SCIP steps. The construction is recursive, but its goal is simple: after each step, the partially augmented vector remains exchangeable under swaps of the completed coordinate pairs.

The usefulness of SCIP depends on whether the needed conditional laws are tractable. Markov models are a useful case because conditional distributions are local. Suppose $X_1, \ldots, X_p$ form a first-order Markov chain with density

$$p(x_1) \prod_{j=2}^{p} p(x_j \mid x_{j-1}).$$

For an interior coordinate, the ordinary full conditional has density proportional to

$$p(x_j \mid x_{j-1})p(x_{j+1} \mid x_j).$$

In SCIP, once $\widetilde{X}_{j-1}$ has been generated, the conditional for $\widetilde{X}_j$ also includes the factor linking the previous knockoff to the current coordinate. Schematically, the interior density is proportional to

$$p(\widetilde{x}_j \mid x_{j-1})p(x_{j+1} \mid \widetilde{x}_j)p(\widetilde{x}_{j-1} \mid \widetilde{x}_j),$$

with boundary modifications at $j = 1$ and $j = p$. This locality is what makes exact knockoff sampling feasible for some non-Gaussian structured features. Sesia et al. [104] carry this out for hidden Markov models in genome-wide association studies, where the linkage-disequilibrium structure between SNPs is well approximated by a Markov chain on chromosomes.

9. Computation and Approximation

CRTs are conceptually direct but computationally expensive. Testing p features with B Monte Carlo resamples requires roughly $p(B + 1)$ evaluations of the feature-importance statistic. If the statistic itself is a cross-validated predictive model, this can be prohibitive. Several variants reduce this cost: Liu et al. [86] introduce a *distillation* CRT that screens out features and amortizes intermediate computations, often matching the power of the naive CRT with one or two orders of magnitude fewer statistic evaluations. Knockoffs pay a different cost: construct one augmented design and compute one set of feature statistics, but accept the need to build a valid knockoff generator.

Figure 5 should not be read as a universal ranking of methods. Power depends heavily on the signal structure, the feature correlations, the statistic, and the number of hypotheses. The stable

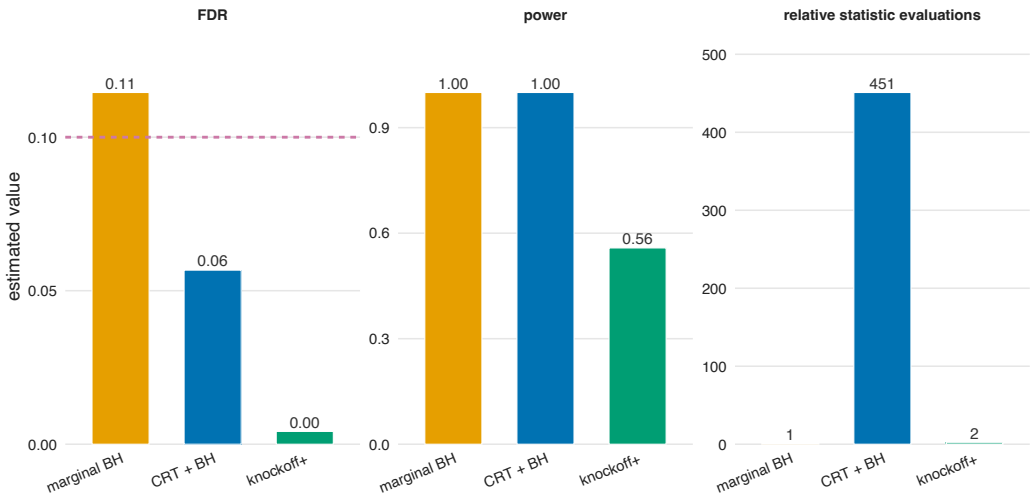

FIGURE 5. A small Gaussian model-X simulation with correlated features. CRT uses $B = 450$ conditional resamples per feature; knockoff+ uses one knockoff copy and a signed marginal-correlation statistic. The right panel shows relative statistic evaluations, not wall-clock time.

lesson is computational: CRT is the exact one-feature-at-a-time benchmark, while knockoffs amortize conditional information across the whole family.

In practice, the feature distribution is rarely known exactly. Model-X knockoffs are then approximate. One may fit a Gaussian graphical model, a hidden Markov model (104), a copula model, a normalizing flow, or a deep generative model trained to match the swap-invariance condition (101), and then sample knockoffs from the fitted distribution. Approximate knockoffs should be treated as a modeling step, not as a theorem. Useful diagnostics include:

swap discrepancies:	compare $[X, \widetilde{X}]$ with its coordinate-swapped versions,
two-sample distances:	MMD, energy distance, or Wasserstein distance,
adversarial checks:	train a classifier to distinguish swapped from unswapped pairs,
conditional checks:	inspect whether $\widetilde{X}_j$ preserves dependence on X_{-j}.

These diagnostics do not by themselves prove FDR control, but they reveal the failure modes that matter: if the null signs are not close to symmetric, the negative knockoff wins no longer estimate false discoveries reliably. Barber et al. [8] give a quantitative robustness result: the FDR is bounded by q plus a term that depends on a KL-style divergence between the true and approximating feature distributions, so small modeling errors translate into small FDR inflation.

10. Multilayer Knockoffs

The model-X knockoff filter also extends to structured families of variables. Katsevich and Sabatti [71] construct the *multilayer knockoff filter* as a single procedure that simultaneously controls FDR at several resolutions through one joint threshold vector, not as a separate filter run independently at each layer. The exposition below is a simplified description meant to convey the mechanism; the formal algorithm and FDR theorem belong to Katsevich and Sabatti [71], and the toy group-statistic construction we use as an illustration is not a substitute for the explicit construction in that paper.

Group nulls and group statistics. Let layer ℓ define groups $\mathcal{G}_1^{(\ell)}, \dots, \mathcal{G}_{G_\ell}^{(\ell)}$ of variables, and let $W_1, \dots, W_p$ be the variable-level model-X knockoff statistics. Knockoff theory equips the family $\{W_j\}_{j=1}^p$ with the *conditional* sign-symmetry property: the joint distribution of the signs of the null W_j's, conditional on their absolute values $\{|W_j|\}_j$ and on the non-null W_j's, is distributed as i.i.d. symmetric Bernoulli. A layer-ℓ group $\mathcal{G}_g^{(\ell)}$ is a *group null* if every variable in $\mathcal{G}_g^{(\ell)}$ is a variable-level conditional null. This matches the intersection-null convention used for hierarchical testing, now applied to model-X conditional null hypotheses.

The multilayer knockoff filter requires layer-ℓ group statistics $\{W_g^{(\ell)}\}_{g=1}^{G_\ell}$ that inherit the conditional sign-symmetry under group nulls. As intuition only, in the special case of disjoint groups within each layer and assuming the constituent W_j signs are conditionally i.i.d. symmetric under the group null, one toy construction is the signed sum

$$W_g^{(\ell)} \;=\; \sum_{j \in \mathcal{G}_g^{(\ell)}} W_j, \qquad g = 1, \dots, G_\ell;$$

in this special disjoint setting, global sign-flipping of the group maps the sum to its negative with the same conditional probability, so the signed sum is sign-symmetric under the group null. In overlapping multilayer settings or under more general dependence, this signed-sum statistic is *not* guaranteed to satisfy the formal flip-sign and compatibility conditions required by Katsevich and Sabatti [71]; the multilayer filter and the FDR theorem below apply to any group statistic satisfying those formal conditions. Zeros are excluded from both tails at any positive threshold; ties at non-zero values are broken by a fixed deterministic tie-breaking rule that preserves conditional sign symmetry of the null statistics.

Joint FDP estimate and joint thresholds. Given a threshold vector $\boldsymbol{\tau} = (\tau_1, \dots, \tau_L)$, the multilayer rejection set at layer ℓ is

$$\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau}) \;=\; \{g : W_g^{(\ell)} \geq \tau_\ell \text{ and the cross-layer coherence condition holds at } \boldsymbol{\tau}\},$$

where the coherence condition links rejections across layers via the partition maps: a variable is rejected only if each of its enclosing groups at every coarser layer is also rejected at the joint vector $\boldsymbol{\tau}$. Let $\pi_\ell(j)$ denote the layer-ℓ group containing variable j, and write $\pi_\ell(A) = \{\pi_\ell(j) : j \in A\}$ for the layer-ℓ groups hit by a variable set A. The crucial multilayer FDP estimate is a *joint* estimate, not the marginal per-layer knockoff+ estimate: it tracks layer-ℓ negative $W_g^{(\ell)}$'s only among groups whose status at the other layers is consistent with the threshold vector $\boldsymbol{\tau}$, giving an FDP-style ratio whose denominator is $\#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau})$. Roughly, the negative group statistics among groups coherent with the other-layer thresholds play the role of false-discovery proxies, in analogy with the standard knockoff+ FDP estimator. The explicit form is spelled out in Katsevich and Sabatti [71, §3]; the essential property is that, for each ℓ, the joint estimate $\widehat{\mathrm{FDP}}^{(\ell)}(\boldsymbol{\tau})$ is a knockoff-style overestimate of the true layer-ℓ FDP in the supermartingale sense needed for the optional-stopping argument that controls $\mathrm{FDR}^{(\ell)}$.

The multilayer knockoff filter searches, using the threshold-search rule specified in Katsevich and Sabatti [71], for a coordinate-wise small threshold vector $\boldsymbol{\tau}$ such that the layer-wise FDP estimates are simultaneously below their targets. The formal theorem includes a correction parameter c in the MKF(c)+ procedure; using $c = c_{\mathrm{kn}}$ gives target-level worst-case control, while the uncorrected $c = 1$ version has the constant-factor theoretical bound stated below. The final discovery set $\widehat{R}$ is the variable-level set consistent with the joint rejection vector $\{\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau})\}_{\ell=1}^L$; the layer-$\ell$ FDP associated with $\widehat{R}$ is computed by mapping $\widehat{R}$ back to layer-ℓ groups via π_ℓ, and the layer-ℓ FDR is the expectation of this FDP.

THEOREM 9.10 (Multilayer knockoff FDR control, cited result [71]). *Assume the model-X exchangeability conditions of this chapter and suppose each layer-ℓ group statistic $W_g^{(\ell)}$ satisfies the formal flip-sign and compatibility conditions of Katsevich and Sabatti [71]. For the MKF(c)+ procedure of that paper,*

$$\mathrm{FDR}^{(\ell)}(\widehat{R}) \;\le\; \frac{c_{\mathrm{kn}}}{c}\,\alpha_\ell, \qquad c_{\mathrm{kn}} = 1.93, \qquad \textit{for every layer } \ell.$$

In particular, running the corrected version with $c = c_{\mathrm{kn}}$, or running the uncorrected $c = 1$ version at levels $\alpha_\ell / c_{\mathrm{kn}}$, gives the formal guarantee $\mathrm{FDR}^{(\ell)} \le \alpha_\ell$.

Proof idea. Each layer-ℓ group statistic, by construction, has the conditional sign-symmetry property that drives the standard knockoff+ supermartingale argument. The joint FDP estimate $\widehat{\mathrm{FDP}}^{(\ell)}(\boldsymbol{\tau})$ is, layer by layer, a knockoff-style overestimate of the true layer-ℓ FDP in the supermartingale sense, and the simultaneous stopping rule on the multilayer filter is a valid stopping time for the relevant filtration at every layer. Optional stopping then gives the constant-factor bound in the theorem; using the correction constant converts that bound into target-level control. The bookkeeping that verifies the joint supermartingale property – the heart of the multilayer construction – is in Katsevich and Sabatti [71]; we do not reproduce it here.

In practice, Katsevich and Sabatti [71] also study and recommend the uncorrected version based on simulations, but the worst-case theorem contains the c_{kn} constant. The distinction matters when the sentence is read as a formal guarantee rather than a description of empirical behavior.

REMARK 9.11 (Joint thresholds are essential). The multilayer guarantee crucially does *not* come from running separate layer-wise filters and intersecting the resulting lifted rejection sets. A marginally valid layer-wise procedure need not remain calibrated for the post-intersection target: shrinking a rejection set can drop true discoveries while retaining false ones, inflating the FDP of the intersected output. The simultaneous threshold-choice step in the multilayer filter is what links the layers correctly; intersection on its own is a common fallacy. Exercise 9.24 gives an explicit counterexample.

When variables within a group are tightly correlated (LD blocks in genomics, for example), group-level discoveries are typically the scientifically meaningful target, and the multilayer filter's joint thresholding delivers them at a controlled per-layer FDR while still reporting variable-level findings inside each rejected group.

11. Assumptions in Plain Language

The procedures in this chapter target several related statistical nulls. The CRT and model-X knockoffs both target the conditional independence null $Y \perp\!\!\!\perp X_j \mid X_{-j}$ without assuming a parametric model for $Y \mid X$; their validity rests on knowing or accurately approximating the feature distribution. Fixed-X knockoffs target the linear-model coefficient null $\beta_j = 0$ inside a Gaussian fixed-design model; their validity rests on geometric properties of the design and on Gaussian errors, not on a feature distribution.

The CRT is exact when the conditional law $\mathcal{L}(X_j \mid X_{-j})$ is correct and the resampling is performed conditionally on the observed X_{-j}. The statistic can be arbitrarily complicated, but the conditional resampling law cannot be wrong without changing the null distribution.

Fixed-X knockoffs are exact for the Gaussian linear fixed-design model when the knockoff matrix satisfies the Gram equations and the feature statistic is sufficient and antisymmetric. The construction is geometric and depends on the design X; it does not claim validity for arbitrary random-design conditional nulls.

Model-X knockoffs require a knockoff copy that is pairwise exchangeable with X and generated without using Y. Row shuffling, marginal resampling, or poorly fitted generative models can break this exchangeability. Approximate knockoffs, fit by graphical models, normalizing flows, or deep generators, are a modeling choice, not a theorem; robustness to approximation is studied by Barber et al. [8] and related generator-based constructions by Romano et al. [101].

Multilayer knockoffs inherit the assumptions of model-X knockoffs: a correct or accurately estimated feature law, and valid pairwise exchangeability. No separate within-layer or across-layer independence assumption is added, but the layer-specific group statistics must satisfy the formal flip-sign and compatibility conditions of Katsevich and Sabatti [71]. The formal worst-case statement includes the $c_{\mathrm{kn}} = 1.93$ correction constant in Theorem 9.10.

12. Bibliographic Notes

Fixed-X knockoffs and the original knockoff filter were introduced by Barber and Candès [6]. Model-X knockoffs, the CRT viewpoint, Gaussian model-X construction, and SCIP were developed by Candès et al. [30]. Robustness to approximate knockoffs is studied by Barber et al. [8]. Metropolized knockoff sampling provides one route to exact or controlled sampling for complex distributions [12]. Hidden Markov model knockoffs for genome-wide association studies are due to Sesia et al. [104], and deep generative knockoffs are developed by Romano et al. [101]. Computationally efficient CRT variants, including the distillation CRT used to amortize statistic evaluations across features, are studied by Liu et al. [86]. A frontier extension to privacy-constrained inference is the differentially private model-X knockoff filter of Pournaderi and Xiang [97], which uses a Johnson–Lindenstrauss projection to privatize the knockoff matrix while preserving approximate exchangeability up to bounds determined by the projection dimension. Multilayer and group-aware knockoffs that exploit structured families are developed in Katsevich and Sabatti [71].

13. Exercises

Basic.

Exercise 9.12 (The CRT p-value). Write the CRT algorithm for testing $H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$. Prove that the plus-one Monte Carlo p-value is super-uniform when the conditional law $X_j \mid X_{-j}$ is known exactly.

Exercise 9.13 (Conditional resampling). In Example 9.3, compute $\mathrm{Cov}(X_1, Y)$ and verify that $Y \perp\!\!\!\perp X_1 \mid X_2$. Derive the conditional law $X_1 \mid X_2 = x_2$, and explain why marginal resampling of X_1 is invalid.

Exercise 9.14 (Pairwise exchangeability). State pairwise exchangeability for a model-X knockoff copy $\widetilde{X}$. For $p = 2$, write out explicitly the four swap identities corresponding to $S = \varnothing, \{1\}, \{2\}, \{1, 2\}$.

Intermediate.

Exercise 9.15 (Fixed-X Gram verification). Let

$$\widetilde{X} = X(I - \Sigma^{-1}D) + UC, \qquad C^\top C = 2D - D\Sigma^{-1}D,$$

with $U^\top X = 0$ and $U^\top U = I_p$. Verify directly that

$$X^\top \widetilde{X} = \Sigma - D, \qquad \widetilde{X}^\top \widetilde{X} = \Sigma.$$

Exercise 9.16 (Feasibility of s). Show that the condition $2\Sigma - D \succeq 0$ is equivalent to the positive semidefiniteness of $2D - D\Sigma^{-1}D$ when D is positive definite. How should the argument be modified when some $s_j = 0$?

Exercise 9.17 (Signed-max statistic). Let Z_j and $\widetilde{Z}_j$ be lasso entry times for the original and knockoff variables. Show that

$$W_j = \max\{Z_j, \widetilde{Z}_j\} \operatorname{sign}(Z_j - \widetilde{Z}_j)$$

is antisymmetric under swapping X_j and $\widetilde{X}_j$.

Exercise 9.18 (Knockoff+ threshold). For

$$W = (4.0,\ 2.2,\ -1.7,\ 0.8,\ -0.6,\ 0.4)$$

and target $q = 0.2$, compute the knockoff and knockoff+ thresholds by hand, using the candidate thresholds in $\{|W_j| : |W_j| > 0\}$. Identify the rejected coordinates under each rule. Explain why the plus-one version can make no discoveries when the number of strong positive statistics is too small relative to $1/q$.

Exercise 9.19 (Gaussian model-X conditionals). Starting from the block covariance

$$\begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix},$$

derive the conditional distribution

$$\widetilde{X} \mid X = x \sim N\left((I - D\Sigma^{-1})x,\ 2D - D\Sigma^{-1}D\right).$$

Computational.

EXERCISE 9.20 (CRT simulation). Simulate the Gaussian example in Figure 2. Test the null feature X_1 using a marginal randomization test and a CRT. Estimate the rejection probability of both tests under the conditional null.

EXERCISE 9.21 (Gaussian knockoff implementation). Write a function that generates Gaussian model-X knockoffs for $X \sim N(0, \Sigma)$. Verify empirically that the sample covariance of $(X, \widetilde{X})$ is close to the target block covariance. Then apply the knockoff+ threshold to a simulated sparse linear model.

EXERCISE 9.22 (CRT versus knockoff cost). Reproduce Figure 5 with different values of B, p, and the feature correlation. Report FDR, power, and runtime. Explain when CRT is computationally plausible and when knockoffs are more attractive.

EXERCISE 9.23 (Fixed-X versus model-X targets under misspecification). Generate data with a strongly nonlinear response, say $Y = \sin(2X_k) + \varepsilon$, and a null feature X_j strongly correlated with the nonnull X_k. Apply (a) fixed-X knockoffs in a linear model and (b) Gaussian model-X knockoffs. Compare the rejection sets and explain why the fixed-X null ($\beta_j = 0$ in the misspecified linear model) and the model-X null ($Y \perp X_j \mid X_{-j}$) can disagree when the linear model is wrong. Verify that under a correctly specified Gaussian linear model with random design, the two nulls coincide for every coordinate.

Advanced.

EXERCISE 9.24 (Multilayer knockoff intersection – a fallacy). This is a stylized rejection-set counterexample; constructing actual knockoff statistics with this exact behavior is not required. Specify a probability model in which two separately valid layer-wise FDR procedures control their marginal FDRs at level α, but the intersection $R = \widehat{R}^{(1)} \cap \widehat{R}^{(2)}_{\text{lifted}}$ has expected variable-level FDP exceeding α. A workable design: take four variables grouped into two layer-2 groups ($\{1, 2\}$ and $\{3, 4\}$), where variables 1 and 3 are true signals and variables 2 and 4 are nulls; both layer-2 groups are then non-null, so a layer-2 rejection of either group is a true group discovery.

Fix $\alpha = 0.05$ and pick a probability p with $\alpha < p < 2\alpha$, e.g. $p = 0.08$; the lower bound makes the intersection fail, while the upper bound keeps the layer-1 marginal FDR below α. Arrange the joint distribution of the variable-level and group-level procedures so that exactly one of the following two events occurs in every realization:

- Event A (with probability p): at layer 1 the procedure rejects $\{1, 4\}$ (one signal, one null), and at layer 2 the procedure rejects only group $\{3, 4\}$ (a non-null group). The lifted layer-2 rejection set is $\{3, 4\}$, and the intersection is $R = \{1, 4\} \cap \{3, 4\} = \{4\}$ – a null variable.
- Event B (with probability $1 - p$): no rejections at either layer.

Then the marginal layer-1 FDR is $p \cdot (1/2) = p/2 < \alpha$ and the marginal layer-2 FDR is $0 \leq \alpha$, so each per-layer procedure respects α; but the intersection's variable-level FDR is $p \cdot 1 = p > \alpha$. Verify these computations and use the example to explain why the multilayer knockoff filter chooses one joint threshold vector rather than running separate filters and intersecting their lifted outputs.

EXERCISE 9.25 (Row shuffling is not a knockoff). Generate correlated Gaussian features and create a fake knockoff by permuting the rows of each feature column. Check pairwise exchangeability empirically by comparing $[X, \widetilde{X}]$ with a coordinate-swapped version. Describe which dependence relationships are broken.

EXERCISE 9.26 (SCIP for a Markov chain). Consider a discrete first-order Markov chain

$$p(x_1) \prod_{j=2}^{p} p(x_j \mid x_{j-1}).$$

Derive the full conditional distribution of X_j given X_{-j} for an interior coordinate. Then write the corresponding SCIP update, including the factor involving the previous knockoff.

EXERCISE 9.27 (Approximate knockoff diagnostics). Fit a misspecified Gaussian knockoff generator to non-Gaussian features. Compute at least two swap-discrepancy diagnostics, such as MMD, energy distance, Wasserstein distance, or an adversarial classification accuracy. Then run knockoff+ and discuss how diagnostic failure relates to empirical FDR.

CHAPTER 10

Conformal Prediction and Conformal P-Values

Prediction intervals are often reported as though the fitted model were correct. In modern applications the fitted model may be a random forest, a neural network, a lasso, a quantile regression engine, or an ensemble selected by cross-validation. Even when the predictor is useful, it is rarely credible to treat the entire fitted model as known and correctly specified.

Conformal prediction separates two tasks. The first task is prediction: build an algorithm that produces small errors. The second task is calibration: convert the algorithm's errors on exchangeable calibration data into a prediction set with finite-sample coverage. The coverage statement is distribution-free and marginal over the next observation. Model quality still matters, but it matters through the size and shape of the prediction set, not through the validity of the rank argument.

1. The Target

Let

$$Z_i = (X_i, Y_i), \qquad i = 1, \dots, n+1,$$

where $X_i \in \mathcal{X}$ is a covariate and $Y_i \in \mathcal{Y}$ is a response. The goal is to construct a set

$$C_n(X_{n+1}) = C(D_n, \alpha, X_{n+1})$$

such that

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha.$$

The probability is over the training data and the test point. This is *marginal* coverage, not conditional coverage at each fixed covariate value.

A trivial procedure can satisfy the display by returning $\mathcal{Y}$ with probability $1 - \alpha$ and the empty set otherwise. Conformal prediction is interesting because it can wrap a useful prediction algorithm and produce sets that adapt to the actual difficulty of the prediction problem.

Throughout this chapter, the core assumption is exchangeability:

$$(Z_1, \dots, Z_{n+1}) \stackrel{d}{=} (Z_{\pi(1)}, \dots, Z_{\pi(n+1)})$$

for every permutation π. Independent identically distributed data imply exchangeability, but exchangeability is the property used by the proofs.

2. The Rank Argument

The essential idea appears before covariates enter. Suppose

$$Y_1, \dots, Y_{n+1}$$

are exchangeable real-valued observations. Let

$$k = \lceil (1 - \alpha)(n + 1) \rceil$$

and let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the first n observations. Set

$$q_n = Y_{(k)}$$

when $k \le n$, with $q_n = \infty$ if $k = n + 1$. Then

$$\mathbb{P}(Y_{n+1} \le q_n) \ge 1 - \alpha.$$

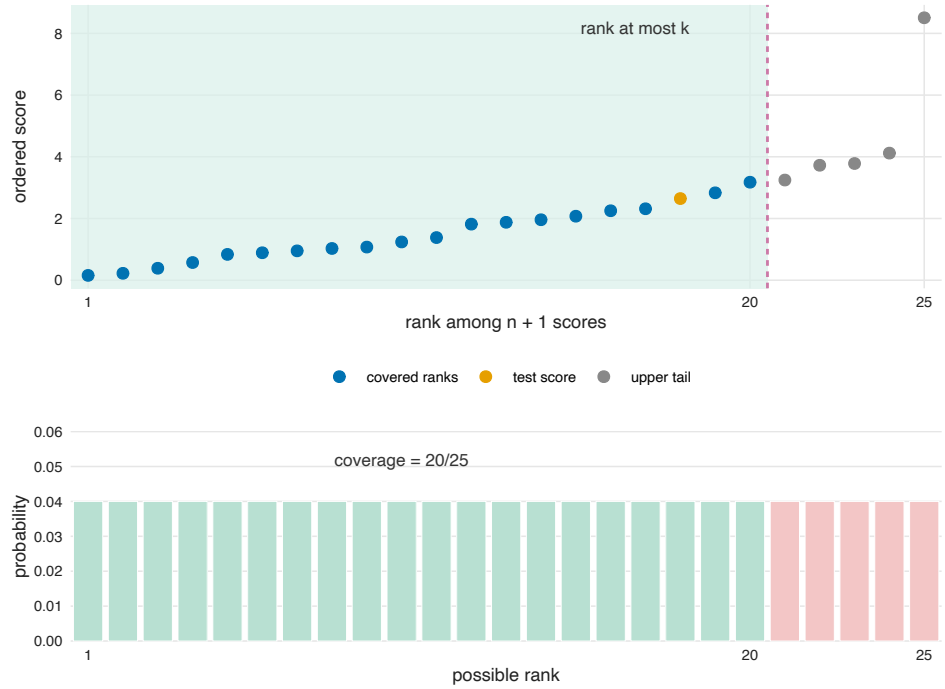

FIGURE 1. The conformal rank argument. Under exchangeability, the test score rank among the $n + 1$ scores is uniform. The finite-sample correction uses $k = \lceil (1 - \alpha)(n + 1) \rceil$, not simply the empirical $1 - \alpha$ quantile of the first n scores. The displayed example takes $n = 24$ and $\alpha = 0.2$, giving $k = 20$ and coverage exactly $20/25$.

PROPOSITION 10.1 (Finite-sample quantile correction). *If $Y_1, \dots, Y_{n+1}$ are exchangeable, then the quantile rule above satisfies*

$$\mathbb{P}(Y_{n+1} \le q_n) \ge 1 - \alpha.$$

If there are no ties almost surely and $k \le n$, then

$$1 - \alpha \le \mathbb{P}(Y_{n+1} \le q_n) = \frac{k}{n+1} < 1 - \alpha + \frac{1}{n+1}.$$

PROOF. Let R_{n+1} be the rank of Y_{n+1} among all $Y_1, \dots, Y_{n+1}$, with any deterministic or random tie-breaking rule that is symmetric among tied observations. Exchangeability implies that R_{n+1} is uniform on $\{1, \dots, n + 1\}$ when there are no ties, and the non-randomized rank event is conservative with ties. The event $Y_{n+1} \le Y_{(k)}$ is the same as saying that Y_{n+1} is among the k smallest of the $n + 1$ observations. Hence its probability is at least $k/(n + 1)$, which is at least $1 - \alpha$. In the no-ties case the probability is exactly $k/(n + 1)$, giving the upper bound. □

The same proof applies to any exchangeable scores. Conformal prediction is mostly a careful way of producing scores for the calibration points and the test point so that this rank proof remains valid.

3. Full Conformal Prediction

A score function assigns a numerical measure of disagreement or nonconformity [128]. For regression, a common choice is

$$S(x, y; \widehat{f}) = |y - \widehat{f}(x)|.$$

Smaller scores mean better conformity. In full conformal prediction, for each candidate value y at the test covariate X_{n+1}, one augments the data with (X_{n+1}, y), fits or refits the prediction rule in a symmetric way, and computes scores

$$R_1(y), \dots, R_n(y), R_{n+1}(y).$$

The conformal p-value for the candidate response is

$$p_y = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{R_i(y) \geq R_{n+1}(y)\}}{n+1},$$

for nonconformity scores where larger is worse. The full conformal prediction set is

$$C_n(x) = \{y : p_y > \alpha\}.$$

Full conformal is conceptually clean because all $n+1$ scores are produced symmetrically. It can also be expensive: in regression, the algorithm may need to be refit for many candidate values y. Split conformal gives up some statistical efficiency in exchange for a much simpler computation.

4. Split Conformal Prediction

Split conformal prediction, as treated for regression by Lei et al. [78], divides the observed data into a proper training set D_1 and a calibration set D_2, with

$$|D_1| = n_1, \qquad |D_2| = n_2, \qquad n_1 + n_2 = n.$$

Fit a predictor $\widehat{f}_{n_1}$ using only D_1. For $i \in D_2$, compute calibration residuals

$$R_i = |Y_i - \widehat{f}_{n_1}(X_i)|.$$

Let

$$k = \lceil (1-\alpha)(n_2 + 1) \rceil$$

and let

$$\widehat{q} = R_{(k)}$$

be the kth smallest calibration residual, with $\widehat{q} = \infty$ if $k = n_2 + 1$. The split conformal interval is

$$C_n(x) = [\widehat{f}_{n_1}(x) - \widehat{q},\ \widehat{f}_{n_1}(x) + \widehat{q}].$$

THEOREM 10.2 (Split conformal coverage). *Assume $Z_1, \dots, Z_{n+1}$ are exchangeable and the split is fixed in advance or chosen independently of the data. Then the split conformal interval satisfies*

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha.$$

If the calibration and test residuals have no ties almost surely, then the coverage is less than

$$1 - \alpha + \frac{1}{n_2 + 1}.$$

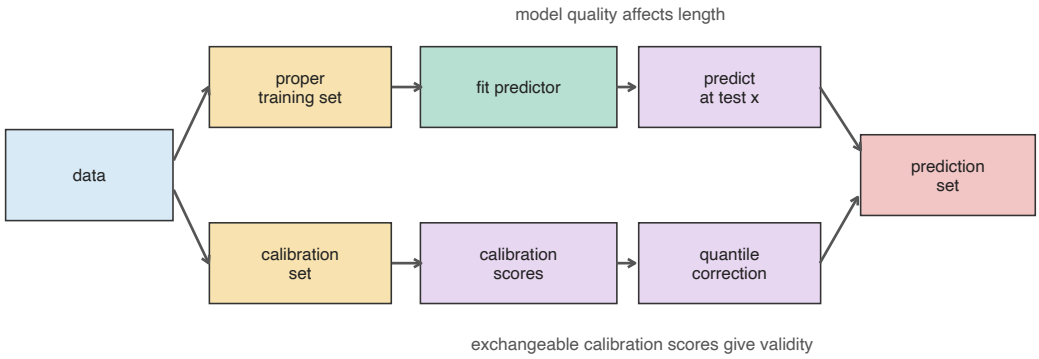

FIGURE 2. Split conformal prediction. The fitted predictor is trained on the proper training set; validity comes from the exchangeability of the calibration scores and the future test score conditional on that fitted predictor.

PROOF. Condition on the proper training set D_1. The fitted function $\widehat{f}_{n_1}$ is then fixed. The calibration observations $\{Z_i : i \in D_2\}$ and the test observation Z_{n+1} remain exchangeable, so the scores

$$\{R_i : i \in D_2\}, \qquad R_{n+1} = |Y_{n+1} - \widehat{f}_{n_1}(X_{n+1})|$$

are exchangeable. The event $Y_{n+1} \in C_n(X_{n+1})$ is exactly $R_{n+1} \le \widehat{q}$. Proposition 10.1 applied to these $n_2 + 1$ scores gives the conditional coverage statement. Averaging over D_1 gives marginal coverage. The no-ties upper bound follows from the same proposition. □

The proof also covers general nonconformity scores. Let

$$S(x, y) = S(x, y; \widehat{f}_{n_1})$$

be any score for which smaller values are more conforming. Compute

$$R_i = S(X_i, Y_i), \qquad i \in D_2,$$

and define

$$C_n(x) = \{y : S(x, y) \le R_{(k)}\}.$$

The rank proof is unchanged. The art is to choose S so that the resulting set is useful.

5. Randomization and Ties

The ordinary conformal set is conservative because the coverage can only move on the grid $\{1/(n_2 + 1), \dots, 1\}$, and ties make it more conservative. Auxiliary randomization can produce exact coverage.

Let W have distribution function F, and let $U \sim \mathrm{Unif}(0, 1)$ independent of W. Define the randomized distribution transform

$$F^*(W) = F(W-) + U\{F(W) - F(W-)\}.$$

Then

$$F^*(W) \sim \mathrm{Unif}(0, 1).$$

When F is continuous, $F^*(W) = F(W)$. At atoms, the uniform randomization spreads the mass across the jump.

For split conformal prediction, let $R_{n+1} = S(X_{n+1}, Y_{n+1})$, and let F_{n_2+1} be the empirical distribution of the $n_2 + 1$ scores formed by the calibration scores and R_{n+1}. Draw a single

$U \sim \mathrm{Unif}(0,1)$ independent of $D_1, D_2, (X_{n+1}, Y_{n+1})$, and use it once to randomize the test rank. The randomized conformal set can be written as

$$C_n^*(x) = \{y : \frac{1}{n_2+1} \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) < S(x,y)\} + \frac{U}{n_2+1}\left(1 + \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) = S(x,y)\}\right) \leq 1-\alpha\}.$$

Conditional on D_1, the randomized rank of the test score is exactly uniform, and therefore

$$\mathbb{P}\{Y_{n+1} \in C_n^*(X_{n+1}) \mid D_1\} = 1-\alpha.$$

This exactness is useful with discrete scores, classification labels, or other settings where ties are common.

6. Conformal P-Values

Conformal scores can also be used for testing.

A brief comment on sign conventions before we start. In the prediction-set sections above, the score $S(x,y) = |y - \widehat{f}(x)|$ is a *nonconformity* score: *smaller* values indicate better agreement with the fitted model. In the conformal p-value literature, it is also common to work with a *conformity* score, where *larger* values indicate better conformity. The two conventions differ only by a sign, and either is fine as long as the comparison in the p-value formula is reversed accordingly. We use the conformity convention in this section because the "small p-value means unusual" interpretation is then verbal-direct.

Let

$$Z_1, \ldots, Z_n \sim P$$

be a reference sample, and let S be a conformity score where larger values mean better conformity. For a new point z, define

$$p(z) = \frac{1 + \#\{1 \leq i \leq n : S(Z_i) \leq S(z)\}}{n+1}.$$

Small p-values indicate that z conforms poorly relative to the reference sample. If $Z_{n+1} \sim P$ and $S(Z)$ is continuous under P, then

$$p(Z_{n+1}) \in \left\{\frac{1}{n+1}, \ldots, 1\right\}$$

is exactly uniform on this grid.

For the absolute-residual score

$$S(x,y) = -|y - \widehat{\mu}(x)|,$$

where $\widehat{\mu}$ is trained on data independent of the reference sample, the inequality $S(Z_i) \leq S(z)$ is $-|Y_i - \widehat{\mu}(X_i)| \leq -|y - \widehat{\mu}(x)|$, which flips under the negation to $|y - \widehat{\mu}(x)| \leq |Y_i - \widehat{\mu}(X_i)|$. The p-value therefore becomes

$$p(x,y) = \frac{1 + \#\{1 \leq i \leq n : |y - \widehat{\mu}(x)| \leq |Y_i - \widehat{\mu}(X_i)|\}}{n+1}.$$

The split conformal exclusion event is essentially a small conformal p-value: if the candidate response lies outside the conformal interval, then its residual is larger than the calibrated residual quantile, and the corresponding conformal p-value is at most α, up to the finite grid.

Now suppose we have m new test points

$$Z_{n+1}, \ldots, Z_{n+m},$$

which are mutually independent and independent of the reference sample. Define

$$p_j = p(Z_{n+j}), \qquad j = 1, \dots, m.$$

The p-values are dependent because they share the same reference sample. Under continuous scores, however, they have a positive dependence structure that is compatible with BH.

THEOREM 10.3 (PRDS property of conformal p-values). *Assume $S(Z)$ is continuous for $Z \sim P$, and assume that $Z_{n+1}, \dots, Z_{n+m}$ are mutually independent and independent of the reference sample. For test points whose null hypotheses $H_{0,j} : Z_{n+j} \sim P$ are true, the conformal p-values are PRDS on the true-null subset.*

PROOF. Consider first p_1 and assume $Z_{n+1} \sim P$. Let

$$S_{(1)} \le \cdots \le S_{(n+1)}$$

be the order statistics of

$$S(Z_1), \dots, S(Z_n), S(Z_{n+1}).$$

Condition on these order statistics and on the scores of the remaining test points. If $p_1 = k/(n+1)$, then $S(Z_{n+1}) = S_{(k)}$, and the reference scores used to compute $p_2, \dots, p_m$ are the order statistics with $S_{(k)}$ removed. Increasing k, equivalently increasing p_1, removes a larger reference score. This cannot decrease any of the other conformal p-values, because each p_j counts how many reference scores are less than or equal to $S(Z_{n+j})$.

In symbols, fix another test score $r = S(Z_{n+j})$ and remove $S_{(k)}$ from the ordered list to form the reference sample used for p_j. The resulting count is

$$N_j(k) = \#\{\ell : S_{(\ell)} \le r\} - \mathbf{1}\{S_{(k)} \le r\}.$$

If k is increased, the indicator $\mathbf{1}\{S_{(k)} \le r\}$ can only decrease, so $N_j(k)$ can only increase. Hence $p_j(k) = \{1 + N_j(k)\}/(n+1)$ is nondecreasing in k, and therefore nondecreasing in $p_1 = k/(n+1)$.

Thus, for fixed auxiliary information, the vector $(p_1, \dots, p_m)$ is coordinatewise increasing in p_1. Under the null, the rank of $S(Z_{n+1})$ is uniform and independent of the order statistics and of the other test scores. Therefore, for any increasing set D,

$$\mathbb{P}\{(p_1, \dots, p_m) \in D \mid p_1 = x\}$$

is increasing in x. The same argument applies to any true-null p-value, which is the PRDS property. □

This theorem is the bridge back to Chapter 6. Conformal p-values are not independent, but under the stated continuous-score null model they satisfy the positive dependence condition under which BH controls FDR.

One technical caveat is worth flagging. Theorem 6.5 in Chapter 6 was stated for true-null p-values that are continuous uniform. Conformal p-values are discrete: they take values on the grid $\{1/(n+1), \dots, 1\}$ and are exactly uniform on this grid. The proof of Theorem 6.5 uses two ingredients: super-uniformity of true-null p-values ($\mathbb{P}(p_i \le t) \le t$ for all t) and the PRDS conditioning inequality, neither of which requires continuous uniformity. A uniform grid distribution is super-uniform, since $\mathbb{P}(p_i \le t) = \lfloor (n+1)t \rfloor/(n+1) \le t$, and the PRDS argument applies verbatim to the discrete law. Hence applying BH to the conformal p-values at FDR target q controls FDR at level q, with the bound FDR $\le m_0 q/m$ preserved. We use q for the FDR target in this paragraph to distinguish it from the conformal miscoverage level α used earlier in the chapter.

7. Conformalized Quantile Regression

Split conformal intervals based on absolute residuals have constant half-width $\widehat{q}$. They are valid, but they do not adapt to heteroscedasticity. If the noise variance is small for some covariates and large for others, a constant-width interval is too wide in easy regions and too narrow in hard regions.

Quantile regression estimates conditional quantiles. For a conditional distribution

$$F(y \mid X = x) = \mathbb{P}(Y \le y \mid X = x),$$

the τth conditional quantile is

$$q_\tau(x) = \inf\{y : F(y \mid X = x) \ge \tau\}.$$

The pinball loss for estimating this quantile is

$$\rho_\tau(y, b) = \begin{cases} \tau(y - b), & y > b, \\ (1 - \tau)(b - y), & y \le b. \end{cases}$$

Classical quantile regression estimates $q_\tau(x)$ by minimizing an average pinball loss, possibly with regularization.

Conformalized quantile regression (CQR) of Romano et al. [100] wraps estimated lower and upper quantiles with conformal calibration. Split the data into D_1 and D_2. On D_1, fit

$$\widehat{q}_{\alpha/2}(x), \qquad \widehat{q}_{1-\alpha/2}(x).$$

We assume throughout that the fits are noncrossing, in the sense that $\widehat{q}_{\alpha/2}(x) \le \widehat{q}_{1-\alpha/2}(x)$ for every x; a quantile-regression fitter that does not guarantee this should be post-processed (for example, by sorting or by isotonic adjustment) before calibration. For each $i \in D_2$, define the calibration score

$$A_i = \max\{\widehat{q}_{\alpha/2}(X_i) - Y_i,\ Y_i - \widehat{q}_{1-\alpha/2}(X_i)\}.$$

The score is positive when Y_i falls outside the fitted interval and negative when it lies inside with slack. Let

$$\widehat{Q} = \text{the } \lceil (1 - \alpha)(n_2 + 1) \rceil \text{th smallest value of } \{A_i : i \in D_2\}.$$

The CQR prediction set is

$$\widetilde{C}_n(x) = \{y : A(x, y) \le \widehat{Q}\} = [\widehat{q}_{\alpha/2}(x) - \widehat{Q},\ \widehat{q}_{1-\alpha/2}(x) + \widehat{Q}],$$

where the right-hand expression collapses to the displayed interval whenever $\widehat{q}_{\alpha/2}(x) - \widehat{Q} \le \widehat{q}_{1-\alpha/2}(x) + \widehat{Q}$, that is, whenever $\widehat{Q} \ge [\widehat{q}_{\alpha/2}(x) - \widehat{q}_{1-\alpha/2}(x)]/2$. The calibrated width is

$$\{\widehat{q}_{1-\alpha/2}(x) - \widehat{q}_{\alpha/2}(x)\} + 2\widehat{Q},$$

so calibration expands or contracts the uncalibrated quantile envelope by $2\widehat{Q}$. Under the noncrossing condition above the interval is nonempty automatically when $\widehat{Q} \ge 0$. Since CQR scores can be negative (a fitted interval already covers Y_i with slack), $\widehat{Q}$ can be negative. A negative correction indicates that the fitted intervals are wider than needed on the calibration sample; very negative corrections should be interpreted cautiously because they may produce empty intervals for some x.

THEOREM 10.4 (CQR coverage). *If $Z_1, \dots, Z_{n+1}$ are exchangeable and the quantile models are fitted using only D_1, then*

$$\mathbb{P}\{Y_{n+1} \in \widetilde{C}_n(X_{n+1})\} \ge 1 - \alpha.$$

If the calibration and test scores A_i have no ties almost surely, then the coverage is less than $1 - \alpha + 1/(n_2 + 1)$.

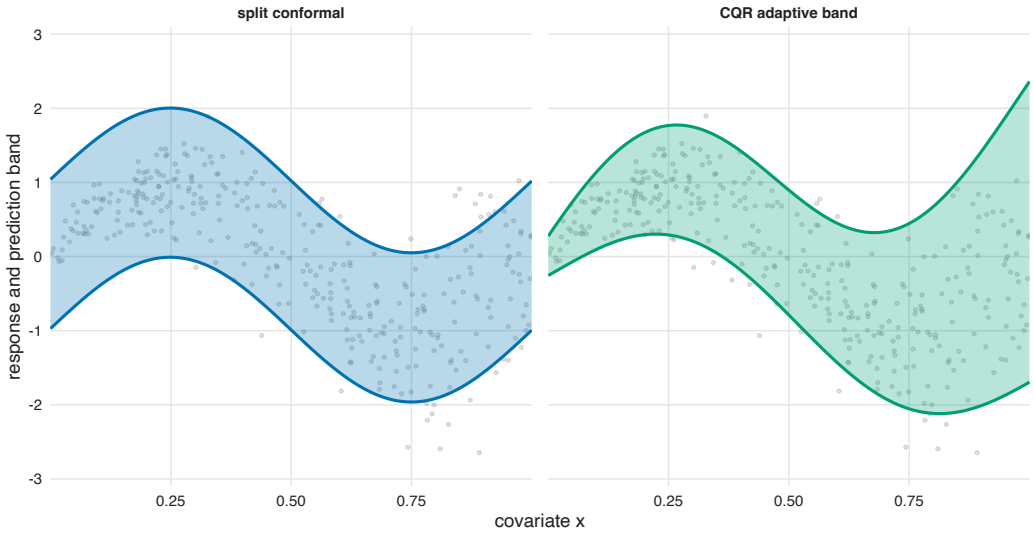

FIGURE 3. Split conformal and CQR adaptive bands in a heteroscedastic regression simulation. Split conformal uses a constant residual quantile; CQR calibrates a covariate-dependent lower and upper band.

PROOF. Conditional on D_1, the fitted quantile functions are fixed. The calibration scores $\{A_i : i \in D_2\}$ and the test score

$$E_{n+1} = \max\{\widehat{q}_{\alpha/2}(X_{n+1}) - Y_{n+1},\ Y_{n+1} - \widehat{q}_{1-\alpha/2}(X_{n+1})\}$$

are exchangeable. The event $Y_{n+1} \in \widetilde{C}_n(X_{n+1})$ is equivalent to $E_{n+1} \le \widehat{Q}$. The rank argument gives the result. □

The conformal calibration is what gives finite-sample marginal validity. The quality of the quantile regression determines how short and locally adaptive the interval is.

8. Jackknife and Jackknife+

Sample splitting is computationally simple, but it may waste data. If D_1 is small, the predictor may be poor. If D_2 is small, the calibration quantile may be unstable. Leave-one-out methods try to use the data more efficiently.

Let $\widehat{f}_{-i}$ be the predictor trained on all observations except (X_i, Y_i). Define the leave-one-out residual

$$R_i^{\mathrm{LOO}} = |Y_i - \widehat{f}_{-i}(X_i)|.$$

The ordinary jackknife interval uses the full-data predictor $\widehat{f}$ and the empirical quantile of the leave-one-out residuals:

$$[\widehat{f}(X_{n+1}) - \widehat{q}_{\mathrm{LOO}},\ \widehat{f}(X_{n+1}) + \widehat{q}_{\mathrm{LOO}}].$$

This interval can work well for stable algorithms, but it does not have a universal finite-sample coverage guarantee. The problem is asymmetry: the residuals are computed using leave-one-out fits, while the interval is centered at the full-data fit. For unstable algorithms, the mismatch can be severe.

Jackknife+ [9] changes the endpoints. For the test covariate X_{n+1}, form

$$L_i = \widehat{f}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}, \qquad U_i = \widehat{f}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}.$$

Let

$$q_{\text{low}} = \text{the } \lfloor \alpha(n+1) \rfloor \text{th smallest value of } L_1, \ldots, L_n,$$

and

$$q_{\text{up}} = \text{the } \lceil (1-\alpha)(n+1) \rceil \text{th smallest value of } U_1, \ldots, U_n,$$

with two boundary conventions when the rank falls outside $\{1, \ldots, n\}$: set $q_{\text{low}} = -\infty$ when $\lfloor \alpha(n+1) \rfloor = 0$, and $q_{\text{up}} = +\infty$ when $\lceil (1-\alpha)(n+1) \rceil = n+1$ (the latter occurs when $\alpha < 1/(n+1)$). The jackknife+ interval is

$$[q_{\text{low}}, q_{\text{up}}],$$

which is the half-line $[q_{\text{low}}, \infty)$ when only the upper boundary triggers, the half-line $(-\infty, q_{\text{up}}]$ when only the lower triggers, and the whole line $\mathbb{R}$ when both trigger.

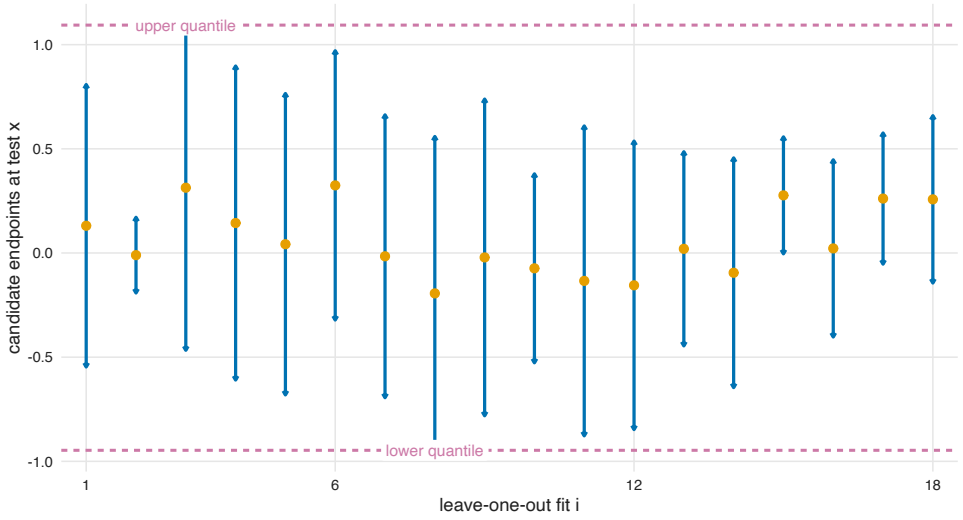

FIGURE 4. Jackknife+ endpoints. Each leave-one-out fit gives a lower and upper candidate endpoint at the test covariate. The final interval is formed from quantiles of these endpoint collections.

THEOREM 10.5 (Jackknife+ coverage, cited result). *Under exchangeability and algorithmic symmetry of the fitting rule, the jackknife+ interval satisfies*

$$\mathbb{P}\{Y_{n+1} \in [q_{\text{low}}, q_{\text{up}}]\} \geq 1 - 2\alpha.$$

Proof idea. The upper-tail event $Y_{n+1} > q_{\text{up}}$ and the lower-tail event $Y_{n+1} < q_{\text{low}}$ are bounded separately, and a union bound combines them.

Consider the upper tail. Augment the dataset to include the test point as the $(n+1)$-th observation, and define the leave-one-out residuals $\widehat{R}_i = |Y_i - \widehat{f}_{-i}(X_i)|$ for $i = 1, \ldots, n+1$, where each $\widehat{f}_{-i}$ is fit on the augmented dataset with index i removed. By algorithmic symmetry of the fitting rule and exchangeability of the augmented data, $\widehat{R}_1, \ldots, \widehat{R}_{n+1}$ are exchangeable. Apply the rank argument of Proposition 10.1 to a comparison statistic linking $\widehat{R}_{n+1}$ to the $\widehat{R}_i$ and to the leave-one-out predictions $\widehat{f}_{-i}(X_{n+1})$; the upper-tail probability is bounded by α. The lower-tail bound is analogous.

This route is the construction in Barber et al. [9]; the details, including the precise comparison statistic and the symmetric ranking that yields the α bound on each tail, are in their Section 3.

The factor 2α comes from controlling the lower and upper tail errors separately. Stronger statements are possible under additional stability or with related cross-validation variants, but the

robust message is that jackknife+ restores a finite-sample guarantee that the ordinary jackknife does not have in general.

9. When Exchangeability Fails

Conformal prediction is often described as distribution-free. The phrase means distribution-free under exchangeability. If the calibration data and the future test point are drawn from different distributions, the rank of the test score among the calibration scores need not be uniform.

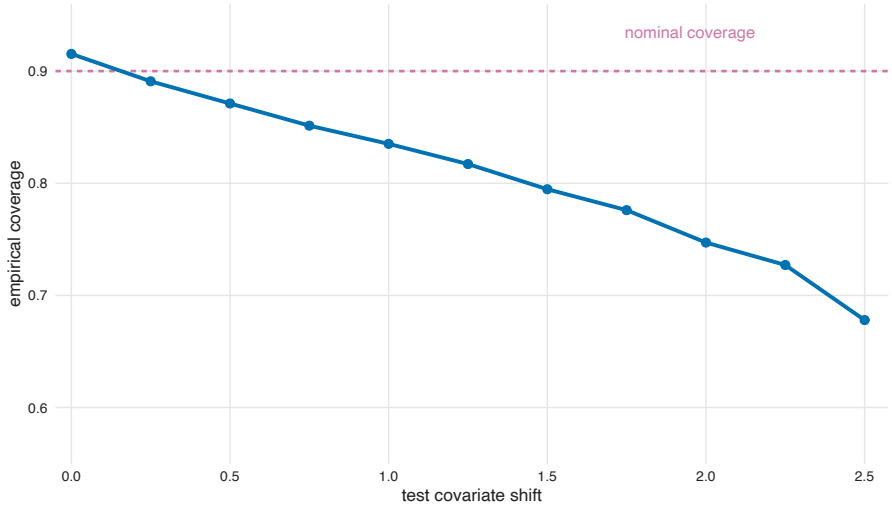

FIGURE 5. A split conformal interval calibrated under one covariate distribution can undercover after test covariate shift. The issue is not model misspecification alone; the exchangeable-rank argument no longer applies.

Weighted and distribution-shift conformal methods modify the calibration weights or the target coverage statement when the shift structure is known or estimable. Those methods require extra assumptions. Without such structure, finite-sample conditional or shifted-distribution coverage cannot be obtained for free.

Weighted conformal prediction. When the test distribution differs from the calibration distribution by a known likelihood-ratio function $w(x, y) = \mathrm{d}P_{\text{test}}/\,\mathrm{d}P_{\text{cal}}$, weighted conformal prediction [123] replaces the empirical CDF of calibration scores by a weighted distribution that includes an atom at $+\infty$:

$$\widehat{\mathcal{F}}_y^w = \sum_{i=1}^{n} \frac{w(X_i, Y_i)}{\sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y)}\, \delta_{S_i} + \frac{w(X_{n+1}, y)}{\sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y)}\, \delta_{+\infty}.$$

The finite-score part of this distribution has CDF

$$\widehat{F}_n^w(s) = \frac{\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbf{1}\{S_i \le s\}}{\sum_{i=1}^{n} w(X_i, Y_i) + w(X_{n+1}, y)}.$$

The corresponding weighted prediction interval recovers exact $(1-\alpha)$ marginal coverage under the shifted distribution, provided the weight function is known. In practice the weights must be estimated, and approximation error appears in the coverage gap. Barber et al. [10] extend this idea to arbitrary distribution drift between calibration and test points, bounding the coverage gap by a quantitative function of the total variation between empirical-weight measures. The

construction does not require the shift to be characterized by a single density ratio: it allows adversarial calibration weights chosen by the user, with a coverage penalty proportional to the maximal weight deviation from the exchangeable case.

Conformal e-prediction. A different generalization is offered by Gauthier et al. [52]. Replace the rank statistic of the score by an e-process, and define a prediction set by inverting an e-value test of $H_0 : Y_{n+1} = y$:

$$C(X_{n+1}) = \{y \in \mathcal{Y} : E_n(y) < 1/\alpha\}.$$

The result is a *conformal e-prediction* set whose marginal coverage is guaranteed by Markov's inequality applied to the e-process. Three consequences follow. First, the construction is anytime-valid: the e-process view immediately extends to a sequence of calibration points indexed by t, giving a prediction set that remains valid as more calibration data arrive. Second, the analyst can dynamically adjust the miscoverage level α post hoc, because the cutoff $1/\alpha$ is applied to the e-value rather than embedded in a rank. Third, fixed-size prediction sets with data-dependent coverage become natural objects: rather than fix α and report a variable-sized set, fix the set size and report the achieved miscoverage e-value as an evidence summary. These properties make conformal e-prediction the natural meeting point of Chapter 8 and Chapter 10.

10. Conformal Risk Control and Learn-Then-Test

The conformal guarantee controls miscoverage, the expectation of a 0/1 loss. Many modern deployments care about quantities that are not miscoverage: the false-negative rate of a medical-image segmentation, the intersection-over-union of a detected object, the token-level F1 score of a generated summary, or the maximum sub-group loss of a fairness-constrained predictor. Angelopoulos et al. [3] introduce *conformal risk control* (CRC), which generalizes coverage control to the expectation of any bounded, monotone loss function.

THEOREM 10.6 (Conformal risk control). *Let $L_\lambda(X, Y)$ be a loss function depending on a real-valued parameter λ and assume that, for every realization of the data, $\lambda \mapsto L_\lambda(X, Y)$ is non-increasing, right-continuous, and $0 \le L_\lambda(X, Y) \le B$. Fix a target loss level $\alpha \in [0, B]$, and assume the feasible set below is nonempty. Let $\widehat{\lambda}$ be the smallest λ such that*

$$\frac{1}{n+1}\left(\sum_{i=1}^{n} L_\lambda(X_i, Y_i) + B\right) \le \alpha.$$

Then the test-point loss satisfies

$$\mathbb{E}[L_{\widehat{\lambda}}(X_{n+1}, Y_{n+1})] \le \alpha$$

under exchangeability of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$.

PROOF. Write $Z_i = (X_i, Y_i)$, and define the empirical surrogate

$$R_n(\lambda) = \frac{1}{n+1}\left(\sum_{i=1}^{n} L_\lambda(X_i, Y_i) + B\right).$$

The map $\lambda \mapsto R_n(\lambda)$ is non-increasing because $\lambda \mapsto L_\lambda(X_i, Y_i)$ is non-increasing for every i. Define $\widehat{\lambda} = \inf\{\lambda : R_n(\lambda) \le \alpha\}$; the infimum is achieved by right-continuity, so $R_n(\widehat{\lambda}) \le \alpha$.

To justify the test-point expectation, use a leave-one-out symmetrization. For each $i \in \{1, \ldots, n+1\}$, let $\widehat{\lambda}^{(-i)}$ be the threshold obtained by applying the same rule to all observations except Z_i:

$$\frac{1}{n+1}\left(\sum_{\ell \ne i} L_\lambda(Z_\ell) + B\right) \le \alpha.$$

By exchangeability,

$$\mathbb{E}[L_{\widehat{\lambda}}(Z_{n+1})] = \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}[L_{\widehat{\lambda}^{(-i)}}(Z_i)].$$

It remains to show the deterministic inequality

$$\sum_{i=1}^{n+1} L_{\widehat{\lambda}^{(-i)}}(Z_i) \le (n+1)\alpha.$$

Let i_* be an index for which $\widehat{\lambda}^{(-i_*)}$ is smallest. Since losses are non-increasing in λ,

$$L_{\widehat{\lambda}^{(-i)}}(Z_i) \le L_{\widehat{\lambda}^{(-i_*)}}(Z_i) \qquad \text{for every } i.$$

The defining inequality for $\widehat{\lambda}^{(-i_*)}$ gives

$$\sum_{i\neq i_*} L_{\widehat{\lambda}^{(-i_*)}}(Z_i) + B \le (n+1)\alpha.$$

Because $L_{\widehat{\lambda}^{(-i_*)}}(Z_{i_*}) \le B$, the full sum at $\widehat{\lambda}^{(-i_*)}$ is at most $(n+1)\alpha$. Combining the last two displays proves the deterministic inequality, and hence $\mathbb{E}[L_{\widehat{\lambda}}(X_{n+1}, Y_{n+1})] \le \alpha$. □

The proof uses exactly the same exchangeability device as the split-conformal rank argument: a symmetric function of the data inherits the exchangeable behavior of the data, which lets the calibration sample certify a bound that transfers to the test point. When $L_\lambda(X,Y) = \mathbf{1}\{Y \notin C_\lambda(X)\}$ and C_λ is the prediction set indexed by quantile level λ, CRC reduces to ordinary split conformal coverage. For richer losses such as $1 - \mathrm{IoU}(C_\lambda(X), Y)$, $1 - \mathrm{F1}(C_\lambda(X), Y)$, or per-token false-negative rates, the same theorem yields an expected-loss guarantee with no change in the proof — only the requirement that the loss be monotone in λ and uniformly bounded.

Beyond monotone losses: Learn Then Test. CRC requires monotonicity of the loss in the calibration parameter. Many operational losses are not monotone: a fairness-adjusted error metric may rise and fall with calibration parameter, and token-F1 in a multi-step generation need not respond monotonically to a single threshold. The *Learn-Then-Test* (LTT) framework of Angelopoulos et al. [2] reduces calibration of an arbitrary risk-controlling parameter to a multiple-testing problem.

Fix a finite grid $\Lambda = \{\lambda_1, \dots, \lambda_K\}$. For each candidate λ, test the null hypothesis

$$H_\lambda : \mathbb{E}[L_\lambda(X,Y)] > \alpha$$

on the calibration data of size n using a concentration-based p-value p_λ. For losses bounded in $[0, B]$, Hoeffding's inequality gives the closed-form valid p-value

$$p_\lambda^{\mathrm{Hoeff}} = \exp\!\left(-\frac{2n}{B^2}(\alpha - \bar{L}_\lambda)_+^2\right), \qquad \bar{L}_\lambda = \frac{1}{n}\sum_{i=1}^{n} L_\lambda(X_i, Y_i),$$

which is super-uniform under H_λ (a direct application of Hoeffding to $\bar{L}_\lambda$ bounded around its mean). Sharper Bernstein- or e-value-based p-values can be used when the loss has small variance or when the experimenter prefers anytime-valid certificates.

Apply BH (or any FWER-controlling procedure of Chapter 4) at level δ to the family $\{p_\lambda : \lambda \in \Lambda\}$, and report the rejected set $\widehat{\Lambda} = \{\lambda : p_\lambda \le \tau_\delta\}$, where τ_δ is the BH cutoff.

THEOREM 10.7 (LTT validity). *Suppose each p_λ is a valid super-uniform p-value for H_λ under the calibration sample. Apply Bonferroni at level δ to the family $\{p_\lambda : \lambda \in \Lambda\}$, so $\widehat{\Lambda}_{Bonf} = \{\lambda : p_\lambda \le \delta/K\}$. Then for any selection rule $\widehat{\lambda} \in \widehat{\Lambda}_{Bonf}$,*

$$\mathbb{P}\Big(\mathbb{E}[L_{\widehat{\lambda}}(X,Y)] \le \alpha\Big) \ge 1 - \delta.$$

PROOF. The event "$\widehat{\Lambda}_{\mathrm{Bonf}}$ contains a null λ" is exactly the event "some $p_\lambda \le \delta/K$ for $\lambda \in \Lambda$ with $\mathbb{E}[L_\lambda] > \alpha$". By the union bound,

$$\mathbb{P}\Big(\widehat{\Lambda}_{\mathrm{Bonf}} \text{ contains a null}\Big) \le \sum_{\lambda : H_\lambda} \mathbb{P}(p_\lambda \le \delta/K) \le K \cdot \delta/K = \delta,$$

using super-uniformity of each p_λ under H_λ. Hence with probability at least $1-\delta$, every λ in the rejected set is *not* a null, i.e., $\mathbb{E}[L_\lambda(X,Y)] \le \alpha$. The bound transfers to any selection rule $\widehat{\lambda}$ that returns an element of $\widehat{\Lambda}_{\mathrm{Bonf}}$. □

Using BH instead of Bonferroni preserves the rejection structure but controls FDR rather than FWER on the calibration grid; the resulting guarantee is that the proportion of rejected λ that fail to satisfy $\mathbb{E}[L_\lambda] \le \alpha$ is bounded in expectation by δ, which can be more informative when the grid is large. The excess-risk bound for a bounded loss over a grid of size K calibrated with Hoeffding p-values is of order $B\sqrt{\log(K)/n}$, which is the minimax rate when the loss is not constrained to be monotone.

The interplay between LTT and the rest of this book is direct. LTT applies a multiple-testing procedure from earlier chapters (BH, Bonferroni, or e-BH on e-value calibration tests) to a calibration grid. Risk-controlling prediction sets in the sense of Bates et al. [11] are a special case in which the test is constructed from a known concentration inequality.

11. Assumptions in Plain Language

Exchangeability is the engine of conformal validity. Once the calibration scores and the test score are exchangeable, the proof is a rank proof. The prediction algorithm may be complicated or misspecified; this affects interval length and adaptivity, not the marginal coverage guarantee.

The guarantee is marginal over the next draw. It does not say that

$$\mathbb{P}\{Y_{n+1} \in C_n(x) \mid X_{n+1} = x\} \ge 1 - \alpha$$

for every x. Distribution-free conditional coverage at every covariate value is impossible without weakening the requirement or adding structure. Randomization can remove finite-grid conservatism and handle ties exactly, but it does not repair nonexchangeability. Jackknife+ uses data more efficiently than split conformal, but it pays with extra computation and, in its most general form, a $1-2\alpha$ coverage guarantee.

12. Bibliographic Notes

The modern conformal prediction framework is developed in Vovk et al. [128] and surveyed by Shafer and Vovk [106]. Distribution-free predictive inference for regression and split conformal methods are treated by Lei et al. [78]. Conformalized quantile regression is due to Romano et al. [100]. Jackknife+ and related cross-validation methods are developed by Barber et al. [9]. Conformal prediction under covariate shift through density-ratio reweighting is developed by Tibshirani et al. [123]; the broader treatment of conformal prediction beyond exchangeability is in Barber et al. [10]. The PRDS property of conformal p-values and BH validity with a shared calibration sample, used in our Theorem 10.3, are established in Bates et al. [13]. An accessible pedagogical introduction covering the breadth of the field is Angelopoulos and Bates [1].

Conformal risk control extends the conformal program from miscoverage to arbitrary monotone bounded losses; the core construction is from Angelopoulos et al. [3], and risk-controlling prediction sets in the dual form are from Bates et al. [11]. The Learn-Then-Test reduction, which uses multiple testing of calibration hypotheses to handle non-monotone losses, is from Angelopoulos et al. [2]. Conformal e-prediction, which replaces the rank step of conformal prediction with an

e-process to obtain data-dependent coverage and anytime-valid prediction sets, is developed by Gauthier et al. [52].

13. Exercises

Basic.

EXERCISE 10.8 (Rank proof). Let $Y_1, \dots, Y_{n+1}$ be exchangeable and let

$$k = \lceil (1-\alpha)(n+1) \rceil.$$

Prove that Y_{n+1} is among the k smallest observations with probability at least $1-\alpha$. Then show that, with no ties, the probability is exactly $k/(n+1)$.

EXERCISE 10.9 (Split conformal interval). Write the split conformal algorithm for absolute residual scores. Identify which data are used to train the predictor and which data are used to compute the residual quantile.

EXERCISE 10.10 (Quantile index). For $n_2 = 49$ calibration points and $\alpha = 0.1$, compute the conformal rank $k = \lceil (1-\alpha)(n_2+1) \rceil$. In the no-ties case where Proposition 10.1 gives equality $k/(n_2+1)$, report the exact coverage and verify it lies in the canonical band $[1-\alpha,\, 1-\alpha+1/(n_2+1))$.

EXERCISE 10.11 (Conformal p-value). Suppose S is a conformity score, where larger means better conformity. Write the conformal p-value for a new point z. Explain how the formula changes if S is instead a nonconformity score, where larger means worse.

Intermediate.

EXERCISE 10.12 (General scores). Let $S(x, y; \widehat{f})$ be any nonconformity score computed after training $\widehat{f}$ on D_1. Prove split conformal coverage for

$$C_n(x) = \{y : S(x, y; \widehat{f}) \le R_{(k)}\}.$$

Here $R_{(k)}$ is the kth order statistic of the calibration scores with $k = \lceil (1-\alpha)(n_2+1) \rceil$; if $k > n_2$, use $R_{(k)} = \infty$.

EXERCISE 10.13 (Randomized ranks). Let W have cdf F, and define

$$F^*(W) = F(W-) + U\{F(W) - F(W-)\}, \qquad U \sim \mathrm{Unif}(0,1).$$

Prove that $F^*(W) \sim \mathrm{Unif}(0,1)$. Then translate this result into exact randomized conformal coverage for discrete scores.

EXERCISE 10.14 (CQR score). Derive the CQR score

$$A_i = \max\{\widehat{q}_{\alpha/2}(X_i) - Y_i,\ Y_i - \widehat{q}_{1-\alpha/2}(X_i)\}.$$

Show that $Y_{n+1} \in \widetilde{C}_n(X_{n+1})$ is equivalent to $E_{n+1} \le \widehat{Q}$.

EXERCISE 10.15 (Conformal p-values and intervals). For the score $S(x, y) = -|y - \widehat{\mu}(x)|$, show algebraically that if a candidate y lies outside the split conformal interval, then its conformal p-value is at most α, up to the finite-sample grid.

Computational.

EXERCISE 10.16 (Heteroscedastic simulation). Reproduce Figure 3. Simulate heteroscedastic data and compare the average length and empirical coverage of absolute-residual split conformal and CQR-style conformal intervals.

EXERCISE 10.17 (Ordinary jackknife failure). Construct a simulation with $X \sim \mathrm{Unif}[-1,1]$, $Y = \sin(4X) + \varepsilon$, and $\varepsilon \sim N(0, 0.2^2)$. Use one-nearest-neighbor regression as the unstable fitting algorithm. Compare ordinary jackknife, split conformal, and jackknife+ intervals over at least 1000 test points and 200 training samples. Report coverage and average length.

EXERCISE 10.18 (Jackknife+ endpoints). For the dataset

$$(X_i, Y_i) \in \{(0, 1.0), (1, 1.8), (2, 3.2), (3, 3.9), (4, 5.1)\}, \qquad X_{n+1} = 2.5,$$

use leave-one-out least-squares simple linear regression with an intercept. Compute $R_i^{\mathrm{LOO}} = |Y_i - \widehat{f}_{-i}(X_i)|$, $\widehat{f}_{-i}(X_{n+1})$, $L_i = \widehat{f}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}$, and $U_i = \widehat{f}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}$ by hand or with a short script. For $\alpha = 0.2$, compute the jackknife+ interval from the order statistics of $\{L_i\}$ and $\{U_i\}$.

EXERCISE 10.19 (Covariate shift). Reproduce Figure 5. Vary the amount of test covariate shift and report empirical coverage. Then try a weighted conformal correction when the density ratio is known.

Advanced.

EXERCISE 10.20 (PRDS proof). Fill in the details of Theorem 10.3. In particular, prove that removing a larger order statistic from the shared reference sample cannot decrease the other conformal p-values.

EXERCISE 10.21 (Beyond marginal coverage). Give an example where split conformal has valid marginal coverage but poor conditional coverage for a subset of the covariate space. Explain why this does not contradict Theorem 10.2.

EXERCISE 10.22 (Conformal p-values and BH). State precisely the dependence structure between the test points that Theorem 10.3 actually requires. Then prove that applying BH to the conformal p-values for m such test points controls FDR at level α, and explain why this property does *not* extend to arbitrary joint dependence among the test points: identify what goes wrong in the PRDS proof if the test points are allowed to be adversarially dependent.

EXERCISE 10.23 (Conformal risk control proof). Prove Theorem 10.6 by reduction to the rank argument. Assume L_λ is monotone non-increasing in λ and bounded by B, and let $\widehat{\lambda}$ be the smallest threshold satisfying the calibration inequality. Show that the test point's contribution to the empirical average is at most B, and conclude using the exchangeability of $(L_\lambda(X_i, Y_i))$ across the calibration points and the test point.

EXERCISE 10.24 (Learn-Then-Test grid effect). For LTT applied to a grid of size K with Hoeffding p-values from n calibration points, derive the excess-risk bound $\mathbb{E}[L_{\widehat{\lambda}}] - \alpha = O(\sqrt{\log K / n})$ for the rejected parameter set. Discuss the trade-off between grid resolution and excess risk, and explain why CRC achieves a rate of $O(\log K / n)$ for monotone losses.

EXERCISE 10.25 (Weighted split conformal under known density ratio). Let the calibration data come from P_{cal} and the test point from P_{test} with known density ratio $w(x, y)$. Derive the weighted prediction set

$$\widehat{C}_n^w(x) = \{y : S(x, y) \le \widehat{Q}_{x,y}^w(\alpha)\}$$

where $\widehat{Q}^{w}_{x,y}(\alpha)$ is the weighted $(1-\alpha)$ empirical quantile of the discrete distribution that places mass

$$\frac{w(X_i, Y_i)}{\sum_{j \in D_2} w(X_j, Y_j) + w(x, y)}$$

on each calibration score $S(X_i, Y_i)$, and mass

$$\frac{w(x, y)}{\sum_{j \in D_2} w(X_j, Y_j) + w(x, y)}$$

at $+\infty$. Prove the resulting marginal coverage guarantee.

CHAPTER 11

Debiased Lasso and High-Dimensional Confidence Intervals

The Lasso is one of the standard tools for prediction and variable screening when the number of features p is comparable to or much larger than the sample size n. Its success is also the reason it is easy to misuse inferentially. The ℓ_1 penalty shrinks estimates toward zero; that shrinkage is useful for prediction, but it creates bias that is not negligible on the $n^{-1/2}$ scale used by ordinary confidence intervals.

This chapter studies a different inferential target. We do not ask whether a variable survived model selection, and we do not pretend that the selected model is fixed. Instead, for a pre-specified coordinate j, or for a low-dimensional collection of coordinates, we ask for confidence intervals and tests for the coefficient β_j in a high-dimensional linear model. The main device is to correct the Lasso by adding a nearly orthogonal score term. The correction is called debiasing, desparsifying, or one-step estimation in the literature.

1. Model and Target

We work with the linear model

$$y = X\beta + \varepsilon, \qquad X \in \mathbb{R}^{n\times p}, \quad y, \varepsilon \in \mathbb{R}^n, \quad \beta \in \mathbb{R}^p.$$

The ith row of X is $x_i^\top$, and the jth column is X_j. The empirical Gram matrix is

$$\widehat{\Sigma} = \frac{X^\top X}{n}.$$

When rows of X are random with covariance matrix Σ, the precision matrix is denoted by

$$\Gamma = \Sigma^{-1}.$$

Many papers use Θ for the precision matrix; in this book Γ is the global notation, and M will denote a generic empirical approximate inverse of $\widehat{\Sigma}$.

Let

$$S = \{j : \beta_j \neq 0\}, \qquad s = |S| = \|\beta\|_0.$$

The sparse high-dimensional regime allows p to exceed n, but assumes that s is small enough relative to n and $\log p$. The target in this chapter is the original coefficient β_j, not a coefficient in a data-selected least-squares model. Post-selection targets are treated in Chapter 12.

2. Regularization and the Lasso

When $p > n$, least squares is not uniquely defined without further structure. Regularization imposes such structure by trading empirical fit against a complexity measure. The three most common constraints are

$$\|b\|_0 = \sum_{j=1}^p \mathbf{1}\{b_j \neq 0\}, \qquad \|b\|_1 = \sum_{j=1}^p |b_j|, \qquad \|b\|_2^2 = \sum_{j=1}^p b_j^2.$$

Best subset selection uses the ℓ_0 constraint and searches directly over sparse supports. It is statistically natural but computationally difficult. Ridge regression uses an ℓ_2 penalty and is

stable under collinearity, but it does not usually produce exact zeros. The Lasso [121] uses an ℓ_1 penalty:

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - Xb\|_2^2 + \lambda \|b\|_1 \right\}. \tag{5}$$

The ℓ_1 ball is convex and has corners, so the optimization problem is tractable and the solution is often sparse.

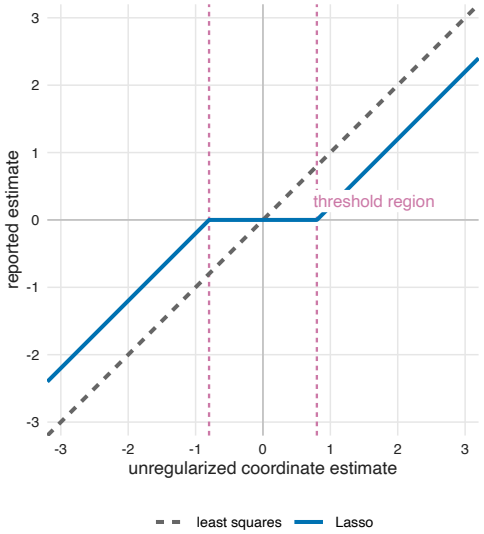

FIGURE 1. Lasso shrinkage in the orthonormal one-coordinate problem. The soft-thresholding rule removes small coordinates and shifts large coordinates toward zero. This bias remains visible on the scale used for ordinary confidence intervals.

The same shrinkage that helps prediction makes direct inference difficult. In the orthonormal one-coordinate problem, the Lasso estimate is

$$\widehat{\beta}_j = \operatorname{sign}(z_j)(|z_j| - \lambda)_+,$$

where $z_j = X_j^\top y/n$ is the least-squares coordinate. This soft-thresholding map is not centered at the truth. Therefore $\widehat{\beta}_j$ is not approximately normal around β_j unless the penalty is asymptotically negligible on the $n^{-1/2}$ scale, which would destroy the regularization needed in high dimensions.

3. What Lasso Consistency Gives

The Lasso can still estimate a sparse vector well in prediction and ℓ_1 norm. These rates are the starting point for debiasing.

DEFINITION 11.1 (Compatibility condition). For the true support S with $|S| = s$, the empirical Gram matrix $\widehat{\Sigma} = X^\top X/n$ satisfies the compatibility condition with constant $\phi_0 > 0$ if, for every $u \in \mathbb{R}^p$ such that $\|u_{S^c}\|_1 \le 3\|u_S\|_1$,

$$\|u_S\|_1^2 \le \frac{s}{\phi_0^2}\, u^\top \widehat{\Sigma} u.$$

The condition says that sparse directions cannot be nearly invisible to the design. It is weaker than requiring all $p \times p$ eigenvalues of $\widehat{\Sigma}$ to be bounded away from zero, which is impossible

when $p > n$. The constant 3 in the cone restriction $\|u_{S^c}\|_1 \le 3\|u_S\|_1$ is not arbitrary: it is the cone that the Lasso error $h = \widehat{\beta} - \beta$ lives in whenever the tuning parameter satisfies the score event $\lambda \ge 2\|X^\top \varepsilon/n\|_\infty$, as the proof of Theorem 11.2 below makes precise. Other constants (e.g., $c \ge 1$ replacing 3) give an analogous restricted eigenvalue condition with different proof constants downstream.

THEOREM 11.2 (A basic Lasso error bound). *Assume the compatibility condition in Definition 11.1. If the tuning parameter satisfies*

$$\lambda \ge 2 \left\| \frac{X^\top \varepsilon}{n} \right\|_\infty ,$$

then the Lasso estimator in (5) obeys

$$\frac{1}{n}\|X(\widehat{\beta} - \beta)\|_2^2 + \lambda \|\widehat{\beta} - \beta\|_1 \le C \, \frac{s\lambda^2}{\phi_0^2}$$

for a universal constant C. In particular, if $\lambda \asymp \sigma\sqrt{(\log p)/n}$, then

$$\|\widehat{\beta} - \beta\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}} \right), \qquad \frac{1}{n}\|X(\widehat{\beta} - \beta)\|_2^2 = O_p\left(\frac{s \log p}{n} \right).$$

PROOF. The optimality of $\widehat{\beta}$ implies the basic inequality obtained by comparing the objective at $\widehat{\beta}$ and at β. With $h = \widehat{\beta} - \beta$,

$$\frac{1}{2n}\|Xh\|_2^2 \le \frac{\varepsilon^\top X h}{n} + \lambda(\|\beta\|_1 - \|\beta + h\|_1).$$

The score event bounds the stochastic term by

$$\left| \frac{\varepsilon^\top X h}{n} \right| \le \left\| \frac{X^\top \varepsilon}{n} \right\|_\infty \|h\|_1 \le \frac{\lambda}{2}\|h\|_1.$$

Since $\beta_{S^c} = 0$, decomposability of the ℓ_1 norm gives

$$\|\beta\|_1 - \|\beta + h\|_1 \le \|h_S\|_1 - \|h_{S^c}\|_1.$$

Combining these two displays yields

$$\frac{1}{2n}\|Xh\|_2^2 + \frac{\lambda}{2}\|h_{S^c}\|_1 \le \frac{3\lambda}{2}\|h_S\|_1.$$

Dropping the nonnegative prediction term shows $\|h_{S^c}\|_1 \le 3\|h_S\|_1$, so h lies in the compatibility cone. The compatibility condition then gives

$$\|h_S\|_1 \le \frac{\sqrt{s}}{\phi_0} \frac{\|Xh\|_2}{\sqrt{n}}.$$

Let $P = \|Xh\|_2^2/n$. The previous two displays imply

$$\frac{P}{2} \le \frac{3\lambda}{2} \frac{\sqrt{s}}{\phi_0} \sqrt{P},$$

and hence $P \le 9s\lambda^2/\phi_0^2$. Finally,

$$\|h\|_1 \le 4\|h_S\|_1 \le \frac{4\sqrt{s}}{\phi_0}\sqrt{P} \le \frac{12 s\lambda}{\phi_0^2}.$$

Multiplying the last display by λ and adding the prediction bound gives the stated inequality with a universal constant, for instance $C = 21$. The probabilistic rates follow by substituting $\lambda \asymp \sigma\sqrt{(\log p)/n}$. □

Theorem 11.2 is an estimation theorem, not an inferential theorem. Its prediction-error bound is consistent under the usual condition $s \log p/n \to 0$. Its ℓ_1-error bound is consistent under the stronger condition

$$s\sqrt{\frac{\log p}{n}} \to 0, \qquad \text{equivalently} \qquad \frac{s^2 \log p}{n} \to 0.$$

Debiased inference needs more than either consistency statement: after multiplication by $\sqrt{n}$, the bias remainder must still vanish.

4. Why the Lasso Limit Is Not Enough

The nonstandard behavior of the Lasso is already visible when p is fixed. Suppose p is fixed, $X^\top X/n \xrightarrow{P} \Sigma_0$, and

$$\frac{X^\top \varepsilon}{\sqrt{n}} \xrightarrow{d} W, \qquad W \sim N(0, \sigma^2 \Sigma_0).$$

If $\sqrt{n}\,\lambda_n \to \lambda_0$, then the limit of $\sqrt{n}(\widehat{\beta} - \beta)$ is the minimizer of a random convex function of the form

$$V(u) = \frac{1}{2} u^\top \Sigma_0 u - u^\top W + \lambda_0 \sum_{j=1}^{p} \left[u_j \operatorname{sign}(\beta_j) \mathbf{1}\{\beta_j \neq 0\} + |u_j| \mathbf{1}\{\beta_j = 0\} \right],$$

up to constants determined by the normalization of the Lasso objective [73]. This limit is not generally centered normal. If $\beta_j = 0$, the absolute value term creates a kink at the origin; if $\beta_j \neq 0$, the linear penalty term shifts the center. High-dimensional debiased methods should therefore be viewed as new estimators for inference, not as ordinary normal approximations to the raw Lasso.

5. Debiasing by an Approximate Inverse

Let $M \in \mathbb{R}^{p \times p}$ be a matrix, computed from X, whose rows nearly invert the empirical Gram matrix. Write $m_j^\top$ for the jth row of M. The debiased Lasso estimator is

$$\widehat{b} = \widehat{\beta} + M \frac{X^\top (y - X\widehat{\beta})}{n}. \tag{6}$$

Substituting $y = X\beta + \varepsilon$ gives the exact decomposition

$$\widehat{b} - \beta = M \frac{X^\top \varepsilon}{n} + (I - M\widehat{\Sigma})(\widehat{\beta} - \beta). \tag{7}$$

For coordinate j,

$$\sqrt{n}(\widehat{b}_j - \beta_j) = \frac{m_j^\top X^\top \varepsilon}{\sqrt{n}} + \sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta} - \beta), \tag{8}$$

where e_j is the jth standard basis vector.

The first term in (8) is a score term. If $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$, then conditionally on X,

$$\frac{m_j^\top X^\top \varepsilon}{\sqrt{n}} \sim N\left(0, \sigma^2 m_j^\top \widehat{\Sigma} m_j\right).$$

For non-Gaussian errors, the same normal approximation follows from a Lindeberg-type central limit theorem under standard moment conditions. The second term is the price of using an approximate inverse rather than the exact inverse of $\widehat{\Sigma}$, which typically does not exist when $p > n$.

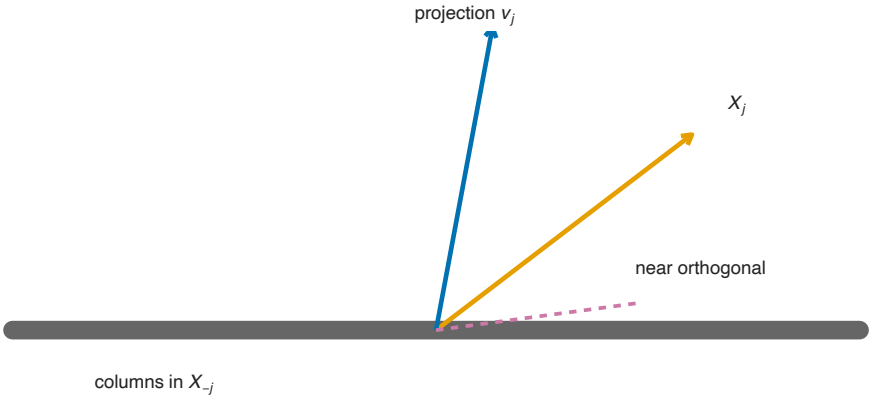

FIGURE 2. The debiasing direction should keep unit sensitivity to the target column X_j while being nearly orthogonal to nuisance columns X_{-j}. This geometry is the coordinate version of Neyman orthogonalization.

The remainder is controlled by Holder's inequality:

$$\left|\sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta} - \beta)\right| \le \sqrt{n}\, \|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \|\widehat{\beta} - \beta\|_1. \tag{9}$$

If

$$\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$$

and Theorem 11.2 gives $\|\widehat{\beta} - \beta\|_1 = O_p(s\sqrt{(\log p)/n})$, then the scaled remainder is

$$O_p\left(\frac{s \log p}{\sqrt{n}}\right).$$

Thus a common sufficient condition for coordinatewise normality is

$$\frac{s \log p}{\sqrt{n}} \to 0. \tag{10}$$

This condition is stronger than the condition needed for prediction consistency, because inference must remove bias on the $n^{-1/2}$ scale.

THEOREM 11.3 (Coordinatewise debiased normality). *Fix a coordinate j. Suppose the Lasso satisfies $\|\widehat{\beta} - \beta\|_1 = O_p(s\sqrt{(\log p)/n})$, the row m_j is X-measurable and satisfies $\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty = O_p(\sqrt{(\log p)/n})$, and $s \log p/\sqrt{n} \to 0$. Suppose also that*

$$0 < c \le m_j^\top \widehat{\Sigma} m_j \le C < \infty$$

with probability tending to one, that the noise satisfies $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$ (or a Lindeberg-type sub-Gaussian moment condition together with the vanishing-influence requirement

$$\frac{\max_{1 \le i \le n} |(X m_j)_i|}{\|X m_j\|_2} \to 0$$

in probability). Then

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j)}{\sigma \sqrt{m_j^\top \widehat{\Sigma} m_j}} \xrightarrow{d} N(0, 1).$$

PROOF. Use the decomposition (8). The leading score term is conditionally Gaussian under Gaussian errors:

$$\frac{m_j^\top X^\top \varepsilon/\sqrt{n}}{\sigma\sqrt{m_j^\top \widehat{\Sigma} m_j}} \mid X \sim N(0,1).$$

Under the stated non-Gaussian alternative, the same standardized score is asymptotically normal by the Lindeberg central limit theorem, because the vanishing-influence condition prevents any single observation from dominating the sum. The remainder satisfies

$$\left|\sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta}-\beta)\right| \le \sqrt{n}\, O_p\left(\sqrt{\frac{\log p}{n}}\right) O_p\left(s\sqrt{\frac{\log p}{n}}\right) = O_p\left(\frac{s\log p}{\sqrt{n}}\right) = o_p(1).$$

Dividing by $\sigma\sqrt{m_j^\top \widehat{\Sigma} m_j}$, which is bounded away from zero and infinity with probability tending to one, keeps the remainder $o_p(1)$. Slutsky's theorem completes the argument. □

If $\widehat{\sigma} \xrightarrow{P} \sigma$, Slutsky's theorem replaces σ in the studentized statistic by $\widehat{\sigma}$, and a $1-\alpha$ confidence interval for β_j is

$$\widehat{b}_j \pm z_{1-\alpha/2}\,\widehat{\sigma}\sqrt{\frac{m_j^\top \widehat{\Sigma} m_j}{n}}. \tag{11}$$

The scaled Lasso [117] is one common route to estimating σ. Residual-based estimates can also work, but only under assumptions that make the residual degrees of freedom and model bias sufficiently small.

Read Figure 3 from left to right as the sparsity burden increases: the main check is whether the empirical coverage curve stays near the nominal 95 percent line while interval lengths remain reasonable.

6. Nodewise Lasso and Precision Approximation

It remains to construct rows m_j that nearly invert $\widehat{\Sigma}$. A useful population identity comes from Gaussian graphical models. Let

$$W = (W_1, \dots, W_p)^\top \sim N(0, \Sigma), \qquad \Gamma = \Sigma^{-1},$$

where W is an abstract population row used only to derive the precision identity below; it should not be confused with the studentized statistic Z_j^{db} introduced in Section 8. The population regression of W_j on W_{-j} has coefficient vector

$$\gamma_j^\star = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j},$$

and residual variance

$$\tau_j^2 = \Sigma_{jj} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}.$$

The block inverse formula gives

$$\Gamma_{jj} = \frac{1}{\tau_j^2}, \qquad \Gamma_{j,-j} = -\frac{(\gamma_j^\star)^\top}{\tau_j^2}. \tag{12}$$

Thus each row of the precision matrix can be recovered from a regression of one feature on the remaining features. When $p \gg n$, those regressions are again high-dimensional, so they are regularized by Lasso.

For each j, define the nodewise Lasso estimator

$$\widehat{\gamma}_j \in \arg\min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n}\|X_j - X_{-j}\gamma\|_2^2 + 2\lambda_j\|\gamma\|_1 \right\}. \tag{13}$$

FIGURE 3. A simulation for $n = 120$, $p = 70$, Gaussian AR(1) designs with correlations $\rho = 0.1$ or 0.55, target $\beta_1 = 0.55$, $\sigma = 1$, and 500 Monte Carlo repetitions per setting. Naive intervals centered at the Lasso estimate undercover because of shrinkage. Debiased score intervals are much closer to the nominal 95 percent level, with oracle support intervals included only as a benchmark.

(The factor in front of the squared-loss term is a normalization choice; the common implementation in the `glmnet` package uses $1/(2n)$ instead of $1/n$. Multiplying the displayed objective by $1/2$ leaves the minimizer unchanged and gives the usual solver form with penalty coefficient λ_j. Calibrating λ_j therefore requires matching the entire objective convention of the solver, not just comparing the raw penalty coefficient in isolation.) Let $\widehat{c}_j \in \mathbb{R}^p$ be the vector with

$$(\widehat{c}_j)_j = 1, \qquad (\widehat{c}_j)_k = -\widehat{\gamma}_{j,k} \quad (k \neq j),$$

where $\widehat{\gamma}_{j,k}$ is the coefficient assigned to column k in the regression excluding column j. Set

$$\widehat{\tau}_j^2 = \frac{1}{n}\|X_j - X_{-j}\widehat{\gamma}_j\|_2^2 + \lambda_j\|\widehat{\gamma}_j\|_1.$$

This combination is the one given by the KKT identity for the nodewise program: $X_j^\top(X_j - X_{-j}\widehat{\gamma}_j)/n = \widehat{\tau}_j^2$, so the diagonal condition $(\widehat{\Gamma}\widehat{\Sigma})_{jj} = 1$ holds exactly. To see the identity, write $\widehat{r}_j = X_j - X_{-j}\widehat{\gamma}_j$. The KKT conditions for (13) give

$$\frac{X_{-j}^\top \widehat{r}_j}{n} = \lambda_j \widehat{\kappa}_j, \qquad \widehat{\kappa}_j \in \partial\|\widehat{\gamma}_j\|_1,$$

so $\widehat{\gamma}_j^\top \widehat{\kappa}_j = \|\widehat{\gamma}_j\|_1$. Since $X_j = \widehat{r}_j + X_{-j}\widehat{\gamma}_j$,

$$\frac{X_j^\top \widehat{r}_j}{n} = \frac{\|\widehat{r}_j\|_2^2}{n} + \widehat{\gamma}_j^\top \frac{X_{-j}^\top \widehat{r}_j}{n} = \frac{\|\widehat{r}_j\|_2^2}{n} + \lambda_j\|\widehat{\gamma}_j\|_1 = \widehat{\tau}_j^2.$$

The nodewise precision-row estimator is

$$\widehat{\Gamma}_j^\top = \frac{\widehat{c}_j^\top}{\widehat{\tau}_j^2}. \tag{14}$$

Stacking these rows gives $\widehat{\Gamma}$, which can be used as M in (6). Under sparsity of the precision rows and regularity of the design, nodewise Lasso satisfies

$$\|\widehat{\Gamma}\widehat{\Sigma} - I\|_{\max} := \max_{j,k} |(\widehat{\Gamma}\widehat{\Sigma} - I)_{jk}| = O_p\left(\sqrt{\frac{\log p}{n}}\right),$$

which is the approximation needed in (9); see Meinshausen and Bühlmann [91], van de Geer et al. [125], and Dezeure et al. [35].

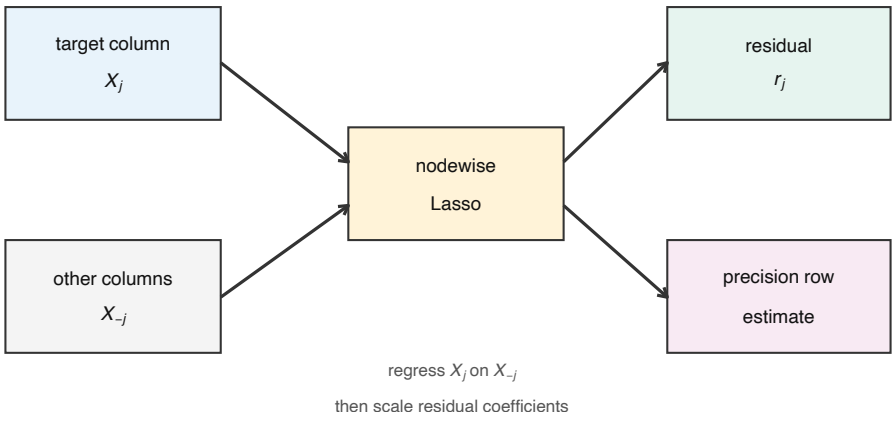

FIGURE 4. Nodewise Lasso estimates the direction used to orthogonalize the target coordinate. Regressing X_j on X_{-j} produces a residual direction and, after scaling, an estimated row of the precision matrix Γ.

7. The Optimal-Projection View

The same construction can be understood without starting from a precision matrix. Fix a target coordinate j and rewrite the model as

$$y - X_{-j}\beta_{-j} = X_j\beta_j + \varepsilon.$$

The left side is not observed because β_{-j} is unknown. Replacing it by the Lasso gives

$$\widehat{\eta}_j = y - X_{-j}\widehat{\beta}_{-j} = X_j\beta_j + \varepsilon + X_{-j}(\beta_{-j} - \widehat{\beta}_{-j}).$$

Let $v_j \in \mathbb{R}^n$ be a projection direction satisfying $v_j^\top X_j = n$. Define

$$\widetilde{\beta}_j = \frac{v_j^\top \widehat{\eta}_j}{n}.$$

Then

$$\sqrt{n}(\widetilde{\beta}_j - \beta_j) = \frac{v_j^\top \varepsilon}{\sqrt{n}} + \frac{v_j^\top X_{-j}(\beta_{-j} - \widehat{\beta}_{-j})}{\sqrt{n}}. \tag{15}$$

The first term has conditional variance $\sigma^2\|v_j\|_2^2/n$. The second term is bounded by

$$\sqrt{n}\left\|\frac{v_j^\top X_{-j}}{n}\right\|_\infty \|\widehat{\beta}_{-j} - \beta_{-j}\|_1.$$

Thus v_j should have moderate Euclidean norm and small correlations with the nuisance columns X_{-j}, while keeping unit sensitivity to X_j. The nodewise residual direction is one way to produce such a v_j.

Javanmard and Montanari [68] formulated this tradeoff row by row. Their program lives in parameter space rather than sample space: parametrize the sample-space direction as $v_j = Xm_j$ for a vector $m_j \in \mathbb{R}^p$. Then

$$v_j^\top X/n = m_j^\top X^\top X/n = m_j^\top \widehat{\Sigma},$$

so requiring $v_j^\top X_j/n \approx 1$ and small $v_j^\top X_k/n$ for $k \neq j$ is exactly the requirement $m_j^\top \widehat{\Sigma} \approx e_j^\top$. In these coordinates the Javanmard–Montanari row program reads

$$\widehat{m}_j \in \arg\min_{m \in \mathbb{R}^p} m^\top \widehat{\Sigma} m \quad \text{subject to} \quad \|m^\top \widehat{\Sigma} - e_j^\top\|_\infty \leq \eta, \tag{16}$$

with $\eta \asymp \sqrt{(\log p)/n}$. The program is trivially feasible when $\eta \geq 1$, because $m = 0$ then satisfies the constraint $\|e_j\|_\infty \leq \eta$. That feasibility is not statistically useful: the standard rates come from the much smaller regime $\eta \asymp \sqrt{(\log p)/n}$, which is feasible under sparse-precision designs. Setting η too small (e.g. $\eta = 0$ with $p > n$ and singular $\widehat{\Sigma}$) leaves an empty feasible set. The constraint controls the bias remainder, and the objective controls the asymptotic variance $\sigma^2 m_j^\top \widehat{\Sigma} m_j$, which equals $\sigma^2 \|v_j\|_2^2/n$ under $v_j = Xm_j$. This makes the statistical tradeoff explicit: more exact orthogonalization reduces bias but may require a high-variance direction.

8. Many Coordinates and Multiple Testing

For a set of pre-specified coordinates J, the debiased estimator produces z-statistics

$$Z_j^{\mathrm{db}} = \frac{\sqrt{n}\,\widehat{b}_j}{\widehat{\sigma}\sqrt{m_j^\top \widehat{\Sigma} m_j}}, \qquad j \in J,$$

for testing $H_{0j} : \beta_j = 0$. Two-sided p-values can be formed as

$$p_j = 2\{1 - \Phi(|Z_j^{\mathrm{db}}|)\}.$$

If only a few coordinates are tested, coordinatewise normality may be enough. If many coordinates are tested, the approximation must be uniform over the tested set, and the dependence among the debiased scores matters for FWER or FDR guarantees.

For an index set J whose size grows with n, a precise uniform Gaussian approximation requires tail conditions, anti-concentration, covariance estimation control, and logarithmic factors that depend on the theorem being used. As a useful scaling heuristic, one needs two strengthenings of Theorem 11.3. First, the remainder bound (9) must be controlled uniformly: $\max_{j \in J} \|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty = O_p(\sqrt{(\log p)/n})$, and the sparsity-product condition is pushed toward $s(\log p)\sqrt{\log |J|}/\sqrt{n} \to 0$, up to theorem-specific log factors. Second, the leading score vector $(m_j^\top X^\top \varepsilon/\sqrt{n})_{j \in J}$ must satisfy a high-dimensional Gaussian approximation, valid simultaneously over J; standard results in this direction extend the high-dimensional CLT of Chernozhukov et al. [33] to debiased Lasso statistics.

Once uniform approximation holds, Bonferroni adjustments require only the marginal validity of the p-values, paid for by a $\sqrt{\log |J|}$ factor in the critical value. BH-type FDR analyses require stronger control of the joint behavior or PRDS-style dependence assumptions, as in Chapter 6.

Bootstrap-assisted simultaneous inference. Bonferroni is conservative because it ignores the dependence among the debiased z-statistics, which can be substantial when the design has correlated columns. Zhang and Cheng [136] develop a bootstrap-assisted procedure that automatically incorporates this dependence. The construction has three steps. First, form the leading score vector $W_j = m_j^\top X^\top \varepsilon/\sqrt{n}$ for $j \in J$, so that the debiased statistic decomposes as $\sqrt{n}\,\widehat{b}_j = W_j + o_p(1)$ uniformly in j. Second, generate bootstrap weights ξ_i (multiplier bootstrap) and form

$$W_j^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \, (m_j^\top X_i) \, \widehat{\varepsilon}_i,$$

where $\widehat{\varepsilon}_i$ are Lasso residuals. Third, use the bootstrap distribution of $\max_{j \in J} |W_j^*|$ to calibrate simultaneous critical values for the debiased statistics. The procedure is valid when the dimension of J is allowed to grow exponentially with n, provided the sparsity and design conditions of debiasing hold. The simultaneous $(1-\alpha)$ confidence intervals it produces are typically much narrower than the Bonferroni intervals when the design has substantial column correlation, because the bootstrap captures the correlation structure directly rather than bounding it through a worst-case dependence argument. The resulting intervals also support max-statistic tests of composite hypotheses such as $H_0 : \beta_j = 0$ for all $j \in J$ and step-down procedures for FWER control on the full coordinate set.

It is important not to reinterpret these p-values as post-selection p-values for the variables chosen by the Lasso. If the question is "among the variables selected by an algorithm, which selected coefficients are significant in the selected model?", the target and conditioning event are different; Chapter 12 treats that problem.

9. Assumptions in Plain Language

The assumptions behind debiased Lasso are structural. Sparsity of β keeps the Lasso estimation error small enough that a bias correction can remove it. Compatibility or restricted-eigenvalue conditions ensure that sparse signals are visible through the design. Sparsity of the relevant precision rows, or a successful projection program, makes approximate orthogonalization feasible. Noise assumptions justify the normal approximation and variance formula.

The strongest-looking condition,

$$s \log p/\sqrt{n} \to 0,$$

has a clear interpretation: after multiplying by $\sqrt{n}$, the product of the Lasso ℓ_1 error and the approximate-inverse error must vanish. The condition is not a technical nuisance; it is exactly what keeps the regularization bias from contaminating the confidence interval.

10. Failure Modes and Diagnostics

Debiased Lasso can fail for several reasons.

- If β is too dense, the Lasso ℓ_1 error is too large and the bias remainder in (9) is not negligible.
- If the design is highly collinear, the approximate inverse can have large variance or fail to satisfy the required ℓ_∞ approximation.
- If the precision rows are dense, nodewise Lasso may not estimate Γ accurately even when β itself is sparse.
- If the noise variance is poorly estimated, the interval width in (11) is wrong even when the center is asymptotically normal.

- If the linear model is misspecified, the target may become a projection coefficient rather than a structural coefficient; the variance formula may also need heteroscedasticity-robust modification.

In practice, one should examine the stability of intervals over reasonable tuning choices, the size of estimated standard errors, the empirical correlations $m_j^\top \widehat{\Sigma} - e_j^\top$, and the plausibility of sparsity for both β and the relevant precision rows.

11. Limits of Adaptivity

The previous sections describe a positive result: under enough sparsity and design regularity, one can obtain honest coordinatewise inference for pre-specified low-dimensional targets. There are also fundamental limits. High-dimensional confidence intervals cannot, in general, be simultaneously short, honest over a large sparse model, and fully adaptive to a much smaller unknown sparsity class. Cai and Guo [26] give minimax lower bounds and adaptivity results for confidence intervals in high-dimensional linear regression. Their results clarify why apparently automatic confidence intervals should be interpreted carefully: extra assumptions such as stronger sparsity, beta-min separation, known support structure, or sample splitting may be needed to obtain the desired length and coverage together.

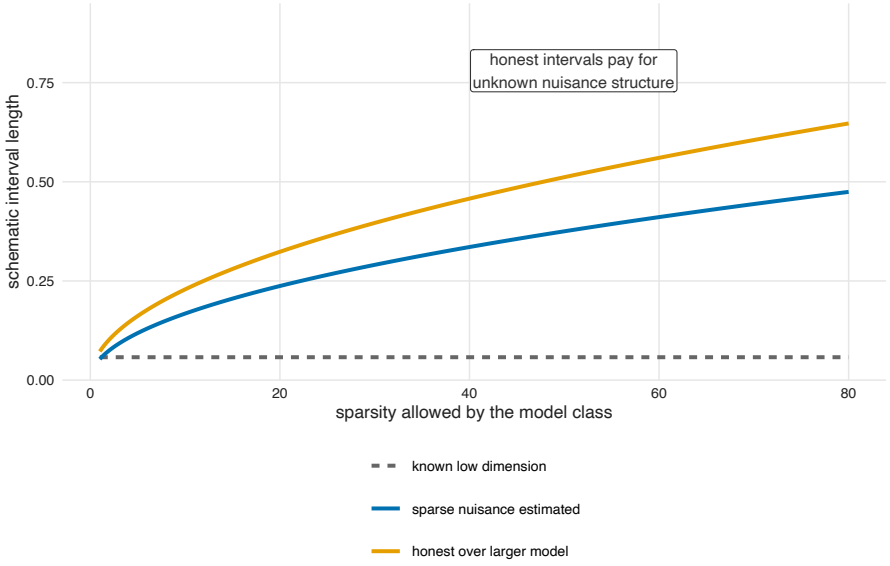

FIGURE 5. A schematic view of adaptivity limits. If the nuisance structure were known, interval length would behave like a low-dimensional standard error. Honest inference over larger sparse classes generally pays extra length for the unknown nuisance structure. The figure is illustrative, not a numerical theorem.

12. Bibliographic Notes

The Lasso was introduced by Tibshirani [121]. Fixed-dimensional asymptotics for Lasso-type estimators were developed by Knight and Fu [73]. The standard textbook reference for Lasso theory, compatibility/restricted-eigenvalue conditions, and high-dimensional regression foundations is Bühlmann and van de Geer [25]. Debiased and desparsified Lasso methods for high-dimensional confidence intervals and tests were developed in Zhang and Zhang [133],

Javanmard and Montanari [68], and van de Geer et al. [125]. The nodewise Lasso construction is closely tied to Gaussian graphical model estimation by Meinshausen and Bühlmann [91]. Dezeure et al. [35] provide a broad account of high-dimensional inference methods and software, while Cai and Guo [26] give minimax and adaptivity limits for high-dimensional confidence intervals. Uniform Gaussian approximations of debiased Lasso statistics over many coordinates extend the maximum-of-sums theory of Chernozhukov et al. [33]. The bootstrap-assisted simultaneous-inference procedure introduced in Section 8 is from Zhang and Cheng [136], who prove its validity for an index set whose size may grow exponentially with n and demonstrate substantial efficiency gains over Bonferroni when the design has correlated columns.

A practically important direction not pursued in detail here is the *streaming* version of debiased Lasso. In modern high-dimensional GLMs with online or batch-arriving observations, the storage and computational demands of forming the full $p \times p$ precision-matrix surrogate become prohibitive. Han et al. [57] show that combining stochastic gradient descent updates with one-pass debiasing achieves asymptotically optimal confidence-interval coverage with strictly $O(p)$ memory. An alternative anytime-valid approach uses confidence sequences (Chapter 8) directly on the debiased Lasso target, replacing fixed-sample intervals with intervals that remain valid under optional stopping.

13. Exercises

Basic.

Exercise 11.4 (Soft-thresholding bias). Consider the orthonormal model with $X^\top X/n = I_p$. Show that the Lasso solution is

$$\widehat{\beta}_j = \operatorname{sign}(z_j)(|z_j| - \lambda)_+, \qquad z_j = X_j^\top y/n.$$

For a nonzero coordinate, compute the shrinkage bias in the regime $|\beta_j| \gg \lambda$. For a zero coordinate, show that the estimator has mean zero by symmetry but has a point mass at zero and a non-Gaussian sampling distribution. Explain why both features make ordinary Wald intervals centered at the Lasso estimate unreliable.

Exercise 11.5 (Debiasing decomposition). Starting from

$$\widehat{b} = \widehat{\beta} + MX^\top(y - X\widehat{\beta})/n,$$

derive (7). Identify the leading stochastic term and the remainder term.

Exercise 11.6 (Coordinate interval). Assume $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$ and M is fixed given X. Derive the conditional variance of $\sqrt{n}(\widehat{b}_j - \beta_j)$ after ignoring the bias remainder, and obtain the confidence interval (11).

Intermediate.

Exercise 11.7 (Remainder control). Prove the bound

$$\left|\sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta} - \beta)\right| \le \sqrt{n}\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \|\widehat{\beta} - \beta\|_1.$$

Using the Lasso ℓ_1 rate and the approximate-inverse rate, show why $s \log p/\sqrt{n} \to 0$ is sufficient for this term to be $o_p(1)$.

Exercise 11.8 (Compatibility condition). Work through the proof sketch of Theorem 11.2. Make the constants explicit under the event $\lambda \ge 2\|X^\top \varepsilon/n\|_\infty$, using the normalization in (5).

Exercise 11.9 (Precision rows from regressions). Let $W \sim N(0, \Sigma)$ and $\Gamma = \Sigma^{-1}$. Use the block inverse formula to prove (12). Interpret the zeros of $\Gamma_{j,-j}$ in terms of conditional linear regression.

EXERCISE 11.10 (Projection direction). For a fixed coordinate j, derive (15). Show how the two requirements $\|v_j\|_2^2/n = O_p(1)$ and $\|v_j^\top X_{-j}/n\|_\infty = o_p\{1/(s\sqrt{\log p})\}$ lead to an asymptotically normal projection estimator.

Computational.

EXERCISE 11.11 (Small debiasing simulation). Simulate $n = 150$, $p = 80$ Gaussian linear models with AR(1) feature correlation. Compare empirical coverage for three intervals for β_1: an interval centered at the Lasso estimate, a debiased interval using a nodewise residual direction, and an oracle interval using the true support. Report coverage and average interval length as the sparsity and correlation change. As a basic sanity check, the naive Lasso-centered intervals should undercover once shrinkage bias is non-negligible, while the debiased intervals should move substantially closer to nominal coverage; compare the setup in `fig09_coverage_sim()` in `scripts/make_figures.R`.

EXERCISE 11.12 (Nodewise Lasso diagnostics). In a simulated design, compute nodewise Lasso rows for several target coordinates. For each coordinate, report

$$\|\widehat{\Gamma}_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \quad \text{and} \quad \widehat{\Gamma}_j^\top \widehat{\Sigma} \widehat{\Gamma}_j.$$

Explain how these quantities correspond to estimated bias and variance.

EXERCISE 11.13 (JM row program). For a small high-dimensional design with $p > n$ (so that $\widehat{\Sigma}$ is singular), solve the Javanmard–Montanari program (16) for several values of η. Compare the resulting variance $m^\top \widehat{\Sigma} m$ and approximation error $\|m^\top \widehat{\Sigma} - e_j^\top\|_\infty$. What happens as η is made smaller? Why does the analogous program with $p < n$ and $\eta = 0$ miss the point of the exercise?

Advanced.

EXERCISE 11.14 (Uniform inference). Theorem 11.3 is stated for a fixed coordinate. Formulate additional conditions that would be needed to justify simultaneous inference over a set J whose size grows with n. Which parts of the proof need to become uniform over $j \in J$?

EXERCISE 11.15 (Post-selection versus coordinate targets). Construct a simple example in which the Lasso selects variable j only when its noisy estimate is unusually large. Explain why a debiased interval for the pre-specified coefficient β_j is not the same as a selective interval conditional on the event that j was selected.

EXERCISE 11.16 (Adaptivity limits). Let $\Theta(s) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le s\}$, with $s_0 < s_1$. Suppose an honest confidence interval for a coordinate must cover uniformly over $\Theta(s_1)$, while an oracle interval is designed only for $\Theta(s_0)$. Using the debiasing remainder rate

$$s \log p/\sqrt{n},$$

explain why a procedure that is honest over $\Theta(s_1)$ cannot in general have the same length as an interval that knows in advance the smaller sparsity level s_0. Relate this heuristic to the lower-bound message of Cai and Guo [26].

CHAPTER 12

Selective Inference and False Coverage Rate

Classical confidence intervals are usually derived for questions fixed before the data are inspected. Modern analyses often reverse the order. The data are collected, interesting variables or clusters are found, and then intervals or p-values are reported for the same discoveries. This changes both the reference distribution and sometimes the target parameter itself.

The issue is not that exploration is illegitimate. Exploration is often how scientific questions are found. The issue is that an interval with marginal coverage for every fixed parameter need not cover at the advertised rate after we restrict attention to the intervals that looked interesting. Sorić summarized this point in the language of "interesting" confidence intervals: coverage over all intervals does not transfer automatically to coverage over the subset selected for attention [111].

This chapter covers two broad responses. The first controls an average error rate over selected intervals, the false coverage rate. The second conditions on the selection event and changes the reference law accordingly. POSI gives a third, deliberately broad, option: protect against arbitrary model selection by using simultaneous intervals over all possible selected-model targets.

1. A Selected-Interval Problem

Let $\theta_1, \dots, \theta_m$ be parameters and let C_i be a marginal $1-\alpha$ confidence interval for θ_i:

$$\mathbb{P}_\theta(\theta_i \in C_i) \geq 1-\alpha \qquad \text{for each fixed } i.$$

Let $\widehat{S} \subseteq \{1, \dots, m\}$ be a data-dependent selected set. The selected intervals are

$$\{C_i : i \in \widehat{S}\}.$$

The marginal statement above does not imply

$$\mathbb{P}_\theta(\theta_i \in C_i \mid i \in \widehat{S}) \geq 1-\alpha.$$

Selection typically favors large noisy estimates, and large noisy estimates are exactly the estimates whose naive intervals are most likely to have missed their targets.

Figure 1 shows the phenomenon in the Gaussian sequence model. Draw

$$\theta_i \overset{\text{i.i.d.}}{\sim} N(0, 0.2^2), \qquad Z_i \mid \theta_i \sim N(\theta_i, 1), \qquad i = 1, \dots, 20,$$

and form 90 percent marginal intervals $Z_i \pm z_{0.95}$. If we report only intervals excluding zero, the selected intervals are biased toward extreme observations.

The same phenomenon appears in regression. If a model is chosen by AIC, BIC, stepwise search, looking at residual plots, or using the same data to screen variables, the usual t-statistic from the final model is no longer distributed as though the model had been fixed in advance. The distortion can be severe when many candidate variables were available.

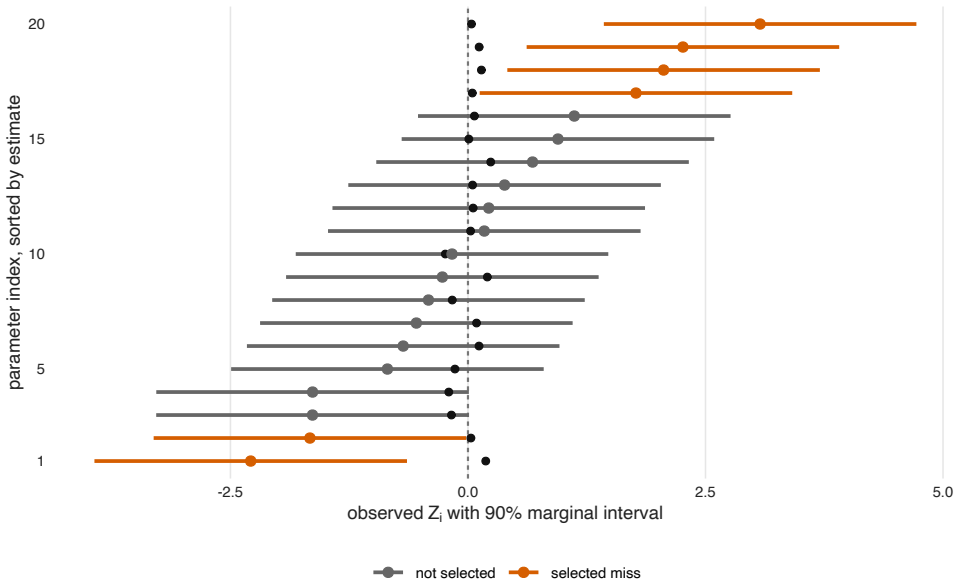

FIGURE 1. Marginal intervals after selection. Each horizontal segment is a 90 percent marginal interval; black dots mark the true θ_i's. Selecting intervals that exclude zero concentrates attention on noisy extremes, so selected intervals can miss far more often than their marginal confidence level suggests.

2. Targets After Selection

The first question in selective inference is: what parameter is the interval supposed to cover?

For a fixed family of parameters $\theta_1, \dots, \theta_m$, the selected set $\widehat{S}$ changes only which intervals are reported. This is the setting of false coverage rate control. The targets remain θ_i, but the error criterion is evaluated over selected indices.

In regression, selection can change the estimand. Suppose $X \in \mathbb{R}^{n \times p}$ is fixed and

$$y \sim N(\mu, \sigma^2 I_n).$$

For a selected model $M \subseteq \{1, \dots, p\}$, the selected-model target is

$$\beta_M = (X_M^\top X_M)^{-1} X_M^\top \mu, \tag{17}$$

assuming $X_M^\top X_M$ is invertible. The coordinate $\beta_{j \bullet M}$ is the coefficient of variable j in the population projection of μ onto the columns selected in M. This target differs from the full-model coefficient in general. Chapter 11 focused on pre-specified coordinates of a high-dimensional coefficient vector; this chapter focuses on targets created or chosen after selection.

3. Conditional Coverage and Its Limits

One natural goal is conditional coverage:

$$\mathbb{P}_\theta(\theta_i \in C_i \mid i \in \widehat{S}) \geq 1 - \alpha. \tag{18}$$

This is attractive but often impossible without changing the interval or the conditioning rule. Consider

$$Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$$

and the usual 95 percent intervals $Y_i \pm 1.96$. Select intervals that exclude zero:

$$\widehat{S} = \{i : 0 \notin [Y_i - 1.96, Y_i + 1.96]\}.$$

If $\mu = 0$, every selected interval excludes the true value. Thus

$$\mathbb{P}_0(\mu \in C_i \mid i \in \widehat{S}) = 0$$

whenever the conditioning event has positive probability. Bonferroni widening can control the probability of at least one error, but it does not solve this conditional problem when the selection rule is itself defined by excluding the null value.

This example is the interval analogue of conditioning on making a rejection under the global null. Conditional error rates can become degenerate because the conditioning event may select precisely the cases in which the nominal procedure failed.

4. False Coverage Rate

Following Benjamini and Yekutieli [17], let

$$R_{\mathrm{CI}} = |\widehat{S}|$$

be the number of selected intervals, and let

$$V_{\mathrm{CI}} = |\{i \in \widehat{S} : \theta_i \notin C_i\}|$$

be the number of selected intervals that fail to cover. The false coverage rate is

$$\mathrm{FCR} = \mathbb{E}\left[\frac{V_{\mathrm{CI}}}{R_{\mathrm{CI}} \vee 1}\right]. \tag{19}$$

This is the confidence-interval analogue of the false discovery rate. It does not promise that every selected interval has conditional coverage (18); it controls the expected fraction of noncovering intervals among the selected intervals.

Two simple procedures control FCR but can be conservative. If all m intervals are reported, marginal $1 - \alpha$ intervals give

$$\mathrm{FCR} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_\theta(\theta_i \notin C_i) \le \alpha.$$

If only selected intervals are reported, simultaneous coverage also suffices:

$$\mathbb{P}_\theta(\theta_i \in C_i \text{ for all } i = 1, \dots, m) \ge 1 - \alpha \quad \Longrightarrow \quad \mathrm{FCR} \le \alpha.$$

Bonferroni intervals are an example. They protect against any selected interval failing, so they also protect the average selected fraction. The price is width.

5. The Benjamini–Yekutieli FCR Adjustment

The same paper proposes an FCR-controlling adjustment that adapts the interval width to the number of selected parameters. Let

$$T = (T_1, \dots, T_m)$$

be statistics used by an arbitrary selection rule $\widehat{S} = \widehat{S}(T)$. For each i, define

$$T^{(i:t)} = (T_1, \dots, T_{i-1}, t, T_{i+1}, \dots, T_m),$$

and

$$R^{(i)} = \min\big(\{|\widehat{S}(T^{(i:t)})| : i \in \widehat{S}(T^{(i:t)}),\ t \in \mathbb{R}\} \cup \{\infty\}\big). \tag{20}$$

Thus $R^{(i)}$ is the smallest number of selections that could occur while holding all other statistics fixed and varying T_i in a way that still selects i. The set on the right of (20) is nonempty on

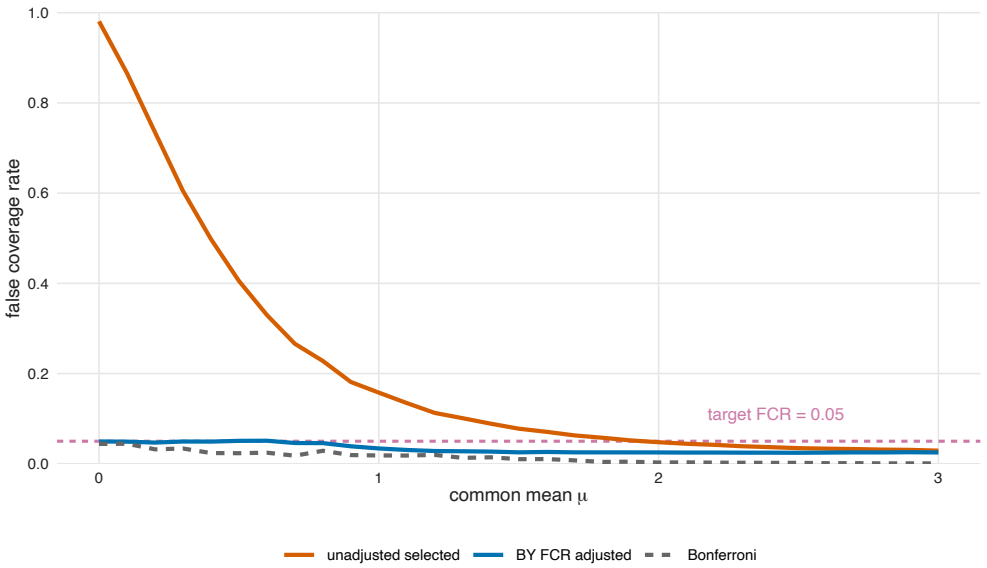

FIGURE 2. FCR in the Gaussian common-mean example with $m = 80$, $\alpha = 0.05$, and 2500 Monte Carlo repetitions at each μ. Intervals selected for excluding zero can have very large FCR when the true mean is near zero. The Benjamini–Yekutieli selected-interval adjustment controls FCR while being less conservative than always using Bonferroni intervals.

the event $i \in \widehat{S}(T)$, since the realized $t = T_i$ is feasible there; off this event the value is ∞. This makes $R^{(i)}$ a function of T_{-i} alone for all T_{-i}, while the reported interval below uses it only when i is selected. For many monotone selection rules, $R^{(i)}$ equals the realized number R_{CI}, but the definition above is the one used in the finite-sample proof.

Suppose $C_i(\gamma)$ denotes a marginal interval with noncoverage probability at most γ:

$$\mathbb{P}_\theta\{\theta_i \notin C_i(\gamma)\} \le \gamma.$$

For each $i \in \widehat{S}$, report

$$C_i\left(\frac{\alpha R^{(i)}}{m}\right). \tag{21}$$

If many parameters are selected, the adjustment is mild. If only one parameter is selected, the interval becomes Bonferroni-like.

THEOREM 12.1 (Benjamini–Yekutieli FCR control). *Assume $T_1, \dots, T_m$ are independent, and assume that for each i, $C_i(\gamma)$ is constructed from T_i and prespecified quantities only (in particular, not from T_{-i}) and has marginal noncoverage at most γ. Then the selected intervals in (21) satisfy*

$$\mathrm{FCR} \le \alpha.$$

PROOF. Let

$$X_i = \frac{\mathbf{1}\{i \in \widehat{S},\ \theta_i \notin C_i(\alpha R^{(i)}/m)\}}{R_{\mathrm{CI}} \vee 1}.$$

Then FCR $= \sum_{i=1}^{m} \mathbb{E}[X_i]$. On the event $\{i \in \widehat{S}, R^{(i)} = k\}$, the definition of $R^{(i)}$ implies $R_{\mathrm{CI}} \geq k$. Therefore

$$X_i \leq \sum_{k=1}^{m} \frac{\mathbf{1}\{\theta_i \notin C_i(\alpha k/m),\ R^{(i)} = k\}}{k}.$$

Condition on the vector of all statistics except T_i. The value $R^{(i)}$ is a function of T_{-i} alone, so it is fixed by this conditioning. The interval $C_i(\gamma)$ is a function of T_i and prespecified ingredients, and under the marginal coverage assumption, $\mathbb{P}_\theta\{\theta_i \notin C_i(\gamma)\} \leq \gamma$ when the law of T_i is the assumed marginal. Because the T_i's are independent, the conditional law of T_i given T_{-i} equals its marginal law, so

$$\mathbb{P}_\theta\{\theta_i \notin C_i(\alpha k/m) \,|\, T_{-i}\} \leq \frac{\alpha k}{m} \quad \text{on} \quad \{R^{(i)} = k\}.$$

Hence

$$\mathbb{E}[X_i \mid T_{-i}] \leq \sum_{k=1}^{m} \frac{\mathbf{1}\{R^{(i)} = k\}}{k} \frac{\alpha k}{m} = \frac{\alpha}{m}.$$

Taking expectations and summing over i gives FCR $\leq \alpha$. □

As with the Benjamini–Hochberg procedure, the independence assumption can be weakened. Benjamini and Yekutieli [17] establish the same FCR bound when the joint law of $(T_1, \ldots, T_m)$ and the coverage events are positive regression dependent in the PRDS sense reviewed in Chapter 6; under arbitrary dependence, replacing α by α / L_m with $L_m = \sum_{k=1}^{m} 1/k$ restores the control.

The FCR criterion is useful, but it is not a cure for every post-selection effect. It controls the frequency of noncoverage for the stated targets. It does not necessarily correct the direction of selection bias, the slope of empirical-Bayes shrinkage, or the fact that a regression target may have changed after model selection. In large screening problems, intervals that control FCR can still look visually odd because they widen symmetrically around selected estimates rather than directly modeling regression to the mean. This is one motivation for empirical-Bayes refinements and for target-specific selective methods.

6. POSI: Protection Against Arbitrary Selection

False coverage rate concerns a fixed family of targets. POSI, short for post-selection inference, addresses a different problem: after looking at the data, an analyst may choose a regression model by an unknown or unreported procedure. Berk et al. [19] protect against this by constructing simultaneous intervals over all selected-model targets.

Let $\mathcal{M} = \{M \subseteq \{1, \ldots, p\} : M \neq \emptyset,\ X_M^\top X_M \text{ invertible}\}$ be the family of full-rank submodels. For every $M \in \mathcal{M}$ and $j \in M$, let

$$\eta_{j \bullet M}^\top = e_j^\top (X_M^\top X_M)^{-1} X_M^\top,$$

where $e_j \in \mathbb{R}^{|M|}$ is the local basis vector picking out j within M, so that

$$\widehat{\beta}_{j \bullet M} = \eta_{j \bullet M}^\top y, \qquad \beta_{j \bullet M} = \eta_{j \bullet M}^\top \mu.$$

Define the standardized error

$$Z_{j \bullet M} = \frac{\eta_{j \bullet M}^\top (y - \mu)}{\sigma \|\eta_{j \bullet M}\|_2}.$$

The POSI constant $K_{1-\alpha}(X)$ is the $1 - \alpha$ quantile of

$$\max_{M \in \mathcal{M}} \max_{j \in M} |Z_{j \bullet M}|.$$

Then the intervals

$$\widehat{\beta}_{j\bullet\widehat{M}} \pm K_{1-\alpha}(X)\,\sigma\|\eta_{j\bullet\widehat{M}}\|_2 \tag{22}$$

cover all selected-model coefficients simultaneously, no matter how $\widehat{M}$ was chosen. In practice σ is unknown and is replaced by an estimate $\widehat{\sigma}$ independent of y (or based on a t-type pivot), which leads to a slightly different constant $K_{1-\alpha}(X, n-p)$; the simultaneous-coverage logic is the same.

THEOREM 12.2 (POSI simultaneous protection). *Under the fixed-design Gaussian linear model, if $K_{1-\alpha}(X)$ is chosen as above with known σ, then for every data-dependent model selection rule $\widehat{M} = \widehat{M}(y)$,*

$$\mathbb{P}\left\{\beta_{j\bullet\widehat{M}} \in C_{j\bullet\widehat{M}} \text{ for all } j \in \widehat{M}\right\} \geq 1-\alpha,$$

where $C_{j\bullet\widehat{M}}$ is the interval in (22).

PROOF. The event

$$\max_M \max_{j\in M} |Z_{j\bullet M}| \leq K_{1-\alpha}(X)$$

implies simultaneous coverage for every model and every coordinate in that model. It therefore implies coverage for the random subset chosen by any selection rule $\widehat{M}(y)$. The probability of the event is at least $1-\alpha$ by the definition of $K_{1-\alpha}(X)$. □

Computing $K_{1-\alpha}(X)$ requires considering many model-coordinate pairs, but it can be approximated by Monte Carlo simulation: draw $y^{(b)} - \mu^{(b)} \sim N(0, \sigma^2 I_n)$ for $b = 1, \ldots, B$, compute the $Z_{j\bullet M}$ for every M and $j \in M$, record the maximum absolute value, and take its empirical $1-\alpha$ quantile. The constant can range from roughly

$$\sqrt{2\log p} \quad \text{to} \quad \sqrt{p}$$

depending on the design and the selection geometry. Orthogonal designs are near the lower end; designs and selection strategies that hunt aggressively over model spaces can approach the upper end. POSI is attractive when the selection process is hard to formalize, but the intervals can be much wider than intervals tailored to a specific selection event.

Sample splitting is a simpler alternative: use one part of the data to select questions and an independent part to perform inference. Its validity relies on an exchangeability or independence structure that justifies the split. Designed experiments, time series, spatial data, and clustered observations may not allow arbitrary splitting without changing the estimand or the reference law.

7. Polyhedral Selective Inference for the Lasso

Chapter 11 treated inference for pre-specified coordinates in a high-dimensional linear model. Here the target is a selected-model coefficient after the Lasso has selected variables. The procedure in Lee et al. [77] assumes that the Lasso tuning parameter λ is fixed before looking at y. Cross-validation is itself a selection procedure; treating a cross-validated λ as fixed changes the conditioning event and is not covered by the basic theorem.

Consider

$$y \sim N(\mu, \sigma^2 I_n),$$

with fixed design X. For fixed $\lambda > 0$, the Lasso estimator is

$$\widehat{\beta}^{\lambda} \in \arg\min_b \left\{\frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1\right\}.$$

FIGURE 3. A schematic comparison of interval width multipliers. Ordinary fixed-model intervals are narrow because the target is pre-specified. POSI protects against arbitrary selection and can be much wider. FCR and selective intervals protect more specific selection problems and targets.

Let

$$\widehat{M} = \{j : \widehat{\beta}^{\lambda}_j \neq 0\}, \qquad \widehat{s} = \operatorname{sign}(\widehat{\beta}^{\lambda}_{\widehat{M}}).$$

The KKT conditions are

$$X_j^\top (y - X\widehat{\beta}^{\lambda}) = \lambda \operatorname{sign}(\widehat{\beta}^{\lambda}_j) \quad (j \in \widehat{M}),$$

and

$$|X_j^\top (y - X\widehat{\beta}^{\lambda})| \leq \lambda \quad (j \notin \widehat{M}).$$

For a fixed active set M and sign vector $s \in \{\pm 1\}^{|M|}$, three ingredients pin down the selection event: the active-set KKT equations expressed in y, the inactive-set inequalities expressed in y, and the strict sign constraint $\operatorname{diag}(s)\widehat{\beta}^{\lambda}_M > 0$ that ensures the active coefficients carry the prescribed signs. After eliminating $\widehat{\beta}^{\lambda}$ using the KKT identity, the equality and inequality parts together can be packaged, up to Gaussian-null boundary events that have probability zero under $y \sim N(\mu, \sigma^2 I_n)$, as the polyhedral event

$$\{\widehat{M} = M, \widehat{s} = s\} = \{Ay \leq b\}, \tag{23}$$

where A and b depend on X, λ, M, and s, but not on the unknown mean μ.

Fix a selected model M, a selected coordinate $j \in M$, and the contrast

$$\eta^\top = e_j^\top (X_M^\top X_M)^{-1} X_M^\top,$$

where, throughout this section, $e_j \in \mathbb{R}^{|M|}$ denotes the local basis vector picking out the position of j within M, not the global basis vector in $\mathbb{R}^p$. The selected-model target is

$$\eta^\top \mu = \beta_{j \bullet M}.$$

Without selection, $\eta^\top y$ is normal with mean $\eta^\top \mu$ and variance $\sigma^2 \|\eta\|_2^2$. After selection, we condition on $\{Ay \leq b\}$. To obtain a one-dimensional pivot, also condition on the nuisance

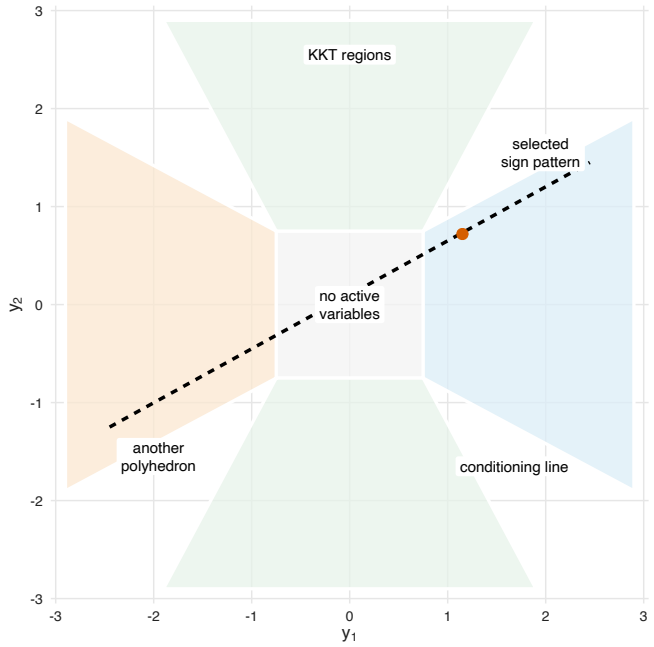

FIGURE 4. A schematic lasso selection partition in the response space. For a fixed active set and sign vector, the KKT conditions define a polyhedron $\{Ay \le b\}$. Conditioning further on the nuisance projection restricts inference to a line segment through the selected polyhedron.

projection

$$P_{\eta^\perp} y = \left(I_n - \frac{\eta\eta^\top}{\|\eta\|_2^2} \right) y.$$

Because y is Gaussian, $\eta^\top y$ is independent of this orthogonal projection.

Let $z = P_{\eta^\perp} y$ be the observed nuisance component. Along the line

$$y = z + \frac{\eta}{\|\eta\|_2^2} t, \qquad t = \eta^\top y,$$

the polyhedral constraint $Ay \le b$ becomes an interval constraint

$$t \in [V^-(z), V^+(z)].$$

Explicitly, for each row $a_\ell^\top$ of A,

$$a_\ell^\top z + \frac{a_\ell^\top \eta}{\|\eta\|_2^2} t \le b_\ell,$$

which gives an upper bound on t when $a_\ell^\top \eta > 0$, a lower bound when $a_\ell^\top \eta < 0$, and a feasibility check $a_\ell^\top z \le b_\ell$ when $a_\ell^\top \eta = 0$. Intersecting,

$$V^-(z) = \max_{\ell:\, a_\ell^\top \eta < 0} \frac{\|\eta\|_2^2 (b_\ell - a_\ell^\top z)}{a_\ell^\top \eta}, \qquad V^+(z) = \min_{\ell:\, a_\ell^\top \eta > 0} \frac{\|\eta\|_2^2 (b_\ell - a_\ell^\top z)}{a_\ell^\top \eta},$$

with $V^-(z) = -\infty$ and $V^+(z) = +\infty$ when the corresponding index sets are empty.

THEOREM 12.3 (Truncated-normal selective pivot). *Under the fixed-λ Gaussian lasso setting, condition on $\{\widehat{M} = M, \widehat{s} = s\}$ and on $P_{\eta^\perp} y = z$. Then*

$$\eta^\top y \sim \mathrm{TN}\left(\eta^\top \mu,\, \sigma^2 \|\eta\|_2^2,\, [V^-(z), V^+(z)]\right),$$

where TN *denotes a normal distribution truncated to the displayed interval. Therefore, if*

$$F^{\nu,\tau^2}_{[V^-,V^+]}$$

is the CDF of $N(\nu, \tau^2)$ truncated to $[V^-, V^+]$, then

$$F^{\eta^\top \mu, \sigma^2\|\eta\|_2^2}_{[V^-(z),V^+(z)]}(\eta^\top y) \sim \mathrm{Unif}(0,1)$$

under the conditional law.

PROOF. Decompose y into the scalar component $t = \eta^\top y$ and the nuisance component $z = P_{\eta^\perp} y$:

$$y = z + \frac{\eta}{\|\eta\|_2^2} t.$$

For a Gaussian vector with covariance $\sigma^2 I_n$, the scalar $\eta^\top y$ is independent of $P_{\eta^\perp} y$, with marginal distribution

$$\eta^\top y \sim N(\eta^\top \mu, \sigma^2 \|\eta\|_2^2).$$

After conditioning on $P_{\eta^\perp} y = z$, the only remaining randomness is therefore the one-dimensional variable t. The selection event $\{Ay \le b\}$, restricted to the line above, is exactly the interval constraint $t \in [V^-(z), V^+(z)]$, because each row of A contributes one linear upper bound, lower bound, or feasibility check. Conditioning a normal random variable on belonging to that interval gives the stated truncated normal law. Applying its conditional CDF to the observed value is uniform by the probability integral transform for the truncated distribution. □

Inverting the pivot gives a selective confidence interval for $\eta^\top \mu$:

$$C_j = \left\{\nu : \frac{\alpha}{2} \le F^{\nu,\sigma^2\|\eta\|_2^2}_{[V^-(z),V^+(z)]}(\eta^\top y) \le 1 - \frac{\alpha}{2}\right\}.$$

This interval targets the selected-model coefficient $\beta_{j\bullet M}$, conditional on the model and sign pattern selected by the fixed-λ Lasso. It is often shorter than POSI because the selection rule is much more specific. The cost is that the analyst must commit to the selection rule, and the computation can become unstable when the conditioning polyhedron is narrow.

8. Selective Inference for Clustering

Selection can also create groups. Suppose observations

$$W_i \sim N(\mu_i, \sigma^2 I_d), \qquad i = 1, \dots, n,$$

are clustered by an algorithm $\mathcal{C}$. After seeing the data, we may choose two clusters $\widehat{C}_1, \widehat{C}_2 \in \mathcal{C}(W)$ and test

$$H_{0,\widehat{C}_1,\widehat{C}_2} : \bar{\mu}_{\widehat{C}_1} = \bar{\mu}_{\widehat{C}_2}.$$

A naive Wald test treats the clusters as fixed. But under a global null, clustering will still tend to separate observations into groups with different sample means. The selected clusters are not fixed labels; they were produced by the same noise used for testing.

For fixed disjoint sets C_1, C_2, define

$$v_i(C_1, C_2) = \frac{\mathbf{1}\{i \in C_1\}}{|C_1|} - \frac{\mathbf{1}\{i \in C_2\}}{|C_2|}.$$

Then

$$\bar{W}_{C_1} - \bar{W}_{C_2} = W^\top v(C_1, C_2).$$

Let

$$P_v^\perp = I_n - \frac{vv^\top}{\|v\|_2^2}.$$

Under the null, the norm of the mean difference, its direction, and the orthogonal nuisance projection can be separated using Gaussian independence. The selective clustering p-value conditions on enough information to make the remaining one-dimensional law tractable:

$$p_{\text{sel}} = \mathbb{P}_{H_0}\{D(W) \geq D(w) \mid E_{\text{sel}}(W, w)\},$$

where $D(W) = \|\bar{W}_{C_1} - \bar{W}_{C_2}\|_2$, and $E_{\text{sel}}(W, w)$ is the event that $C_1, C_2 \in \mathcal{C}(W)$, $P_v^\perp W = P_v^\perp w$, and

$$\operatorname{dir}(\bar{W}_{C_1} - \bar{W}_{C_2}) = \operatorname{dir}(\bar{w}_{C_1} - \bar{w}_{C_2}).$$

After conditioning on the nuisance projection and the observed direction, the data vary along a scalar separation parameter ϕ. The event that C_1, C_2 appear in the clustering becomes

$$\phi \in \mathcal{S}(w, C_1, C_2),$$

a data-dependent truncation set. Thus the selective p-value can be written as

$$\frac{\mathbb{P}\{\phi \geq \phi_{\text{obs}},\ \phi \in \mathcal{S}(w, C_1, C_2)\}}{\mathbb{P}\{\phi \in \mathcal{S}(w, C_1, C_2)\}}, \tag{24}$$

where, under the global null $\mu_1 = \cdots = \mu_n$ (the pairwise null $\bar{\mu}_{C_1} = \bar{\mu}_{C_2}$ alone leaves residual mean structure in the nuisance projection),

$$\phi \sim \sigma\sqrt{|C_1|^{-1} + |C_2|^{-1}}\,\chi_d,$$

and χ_d denotes the chi distribution with d degrees of freedom, that is, the law of the Euclidean norm of a standard $N(0, I_d)$ vector. This is the clustering analogue of the lasso truncated-normal pivot: condition on the selection event and nuisance information until the remaining statistic has a truncated, computable reference distribution.

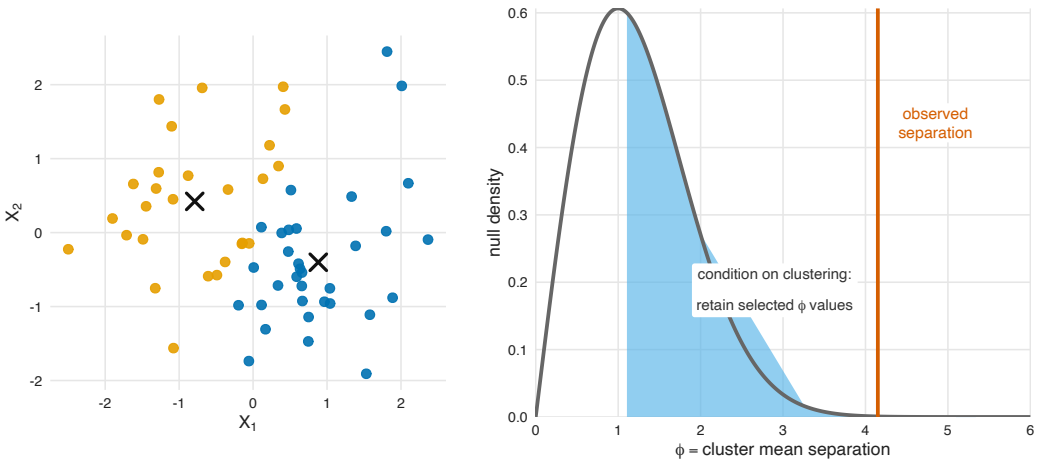

FIGURE 5. Selective inference for clustering. Even under a null model, clustering creates separated groups. Conditional inference varies the cluster separation ϕ along a line determined by the observed nuisance information and evaluates the null law truncated to values that reproduce the selected clustering event.

The practical computation may require Monte Carlo sampling over ϕ and rerunning the clustering algorithm to determine whether $\phi \in \mathcal{S}(w, C_1, C_2)$. The same principle applies to hierarchical clustering [51] and k-means [32], although the selection events differ by algorithm.

9. Data Thinning and Data Fission

Sample splitting is the simplest reusable device for selective inference: split the data into a discovery half and an inference half, choose a target on the discovery half, and conduct standard inference on the inference half. The construction is elegant, finite-sample valid, and immune to the polyhedral-conditioning bookkeeping of Section 7. It also has a fundamental limitation. When the inferential target is the *latent structure* of the data — clusters, principal components, change-points, or any structure jointly identified across observations — splitting observations into two halves changes the target. The clusters estimated on the training half may not map cleanly to the inference half, and the parameter we wished to test ("the mean of cluster 1 estimated by k-means on the full data") is not the same as the parameter we can test ("the mean of cluster 1 estimated by k-means on the training half").

Data thinning and *data fission* address this limitation by splitting *within* each observation rather than *across* observations.

Data thinning for convolution-closed distributions. Neufeld et al. [94] introduce data thinning for distributions closed under convolution. Suppose X has a distribution in a parametric family $\{P_\theta : \theta \in \Theta\}$ such that the convolution of two members of the family is again in the family. Examples include Gaussian, Poisson, negative binomial, and gamma families. Thinning chooses a parameter $\epsilon \in (0, 1)$ and constructs two independent variables $(X^{(1)}, X^{(2)})$ by

$$X^{(1)} \sim P_{\epsilon\theta}, \quad X^{(2)} \sim P_{(1-\epsilon)\theta}, \quad X = X^{(1)} + X^{(2)}.$$

Because the family is convolution-closed and the parameter is additive, the sum $X^{(1)} + X^{(2)}$ reproduces the distribution of X. More importantly, $X^{(1)}$ and $X^{(2)}$ are independent.

THEOREM 12.4 (Independence of thinned components). *Let $X \sim P_\theta$ belong to a convolution-closed family with additive parameter θ: that is, if $Y_1 \sim P_{\theta_1}$ and $Y_2 \sim P_{\theta_2}$ are independent then $Y_1 + Y_2 \sim P_{\theta_1+\theta_2}$. Fix $\epsilon \in (0, 1)$. Then there is a distributional split with $(X^{(1)}, X^{(2)})$ with $X^{(1)} \perp X^{(2)}$, $X^{(1)} + X^{(2)} \overset{d}{=} X$, $X^{(1)} \sim P_{\epsilon\theta}$, and $X^{(2)} \sim P_{(1-\epsilon)\theta}$. An implementable thinning mechanism from an observed X is obtained by sampling from the conditional law of (Y_1, Y_2) given $Y_1 + Y_2 = X$ when that conditional law is known and does not involve unknown nuisance parameters, or when those parameters are known.*

PROOF. For the abstract split, draw $Y_1 \sim P_{\epsilon\theta}$ and $Y_2 \sim P_{(1-\epsilon)\theta}$ independently. Convolution closure gives $Y_1 + Y_2 \sim P_\theta$, so the sum has the same distribution as X. If the conditional law of (Y_1, Y_2) given $Y_1 + Y_2 = x$ is available and parameter-free (or has known parameters), then sampling from that conditional law after observing $X = x$ reconstructs the same joint distribution as (Y_1, Y_2) conditional on its sum. Averaging over $X \sim P_\theta$ therefore gives unconditional marginals $P_{\epsilon\theta}$ and $P_{(1-\epsilon)\theta}$, and the components are independent because the reconstructed joint law is the original product law.

The cleanest concrete case is Poisson, where the conditional law is the classical binomial split. Let $X \sim \text{Poisson}(\lambda)$ and, conditional on $X = k$, draw $X^{(1)} \sim \text{Binomial}(k, \epsilon)$ and set $X^{(2)} = X - X^{(1)}$. The joint mass function factorizes:

$$\mathbb{P}(X^{(1)} = j, X^{(2)} = k) = e^{-\lambda}\frac{\lambda^{j+k}}{(j+k)!}\binom{j+k}{j}\epsilon^j(1-\epsilon)^k = \underbrace{e^{-\epsilon\lambda}\frac{(\epsilon\lambda)^j}{j!}}_{X^{(1)}\sim\text{Pois}(\epsilon\lambda)} \cdot \underbrace{e^{-(1-\epsilon)\lambda}\frac{((1-\epsilon)\lambda)^k}{k!}}_{X^{(2)}\sim\text{Pois}((1-\epsilon)\lambda)}.$$

Independence and the two marginal distributions fall out of this single algebraic identity. Convolution closure of the Poisson family gives $X^{(1)} + X^{(2)} \sim \text{Poisson}(\lambda)$, so the sum has the same distribution as X.

The same recipe works whenever the conditional split law given the sum is available in a form that can be sampled without estimating the target from the same data. For Gaussian data, the analogous independent views are usually treated as data fission via external Gaussian noise when the variance is known; negative binomial thinning uses a beta-binomial split; the abstract split is the conditional law of the sum $Y_1 + Y_2 = X$ for hypothetical independent draws $Y_1 \sim P_{\epsilon\theta}$, $Y_2 \sim P_{(1-\epsilon)\theta}$. The detailed verification and the parameter conditions for common families are in Neufeld et al. [94]. □

The operational consequence is that we can perform any data-driven step (clustering, PCA, change-point detection, feature selection) on $X^{(1)}$ and then conduct *exact* inference on $X^{(2)}$, because $X^{(2)}$ is independent of the choices made from $X^{(1)}$. The two matrices have the same dimension as the original data, so latent structures identified on $X^{(1)}$ map directly to $X^{(2)}$.

EXAMPLE 12.5 (Single-cell clustering). Let X_{ij} be the read count of gene j in cell i, modeled as Poisson with mean μ_{ij}. Compute the thinned counts $X^{(1)}_{ij} \sim \text{Poisson}(\epsilon\mu_{ij})$ by sampling $X^{(1)}_{ij} \mid X_{ij} \sim \text{Binomial}(X_{ij}, \epsilon)$ and setting $X^{(2)}_{ij} = X_{ij} - X^{(1)}_{ij}$. By Theorem 12.4 the matrices $X^{(1)}$ and $X^{(2)}$ are independent. Cluster the cells using $X^{(1)}$ — for instance, by Louvain on the cosine-similarity graph of normalized $X^{(1)}$ — to obtain a partition $\widehat{\mathcal{C}} = (\widehat{C}_1, \ldots, \widehat{C}_g)$. Now test, on $X^{(2)}$, the hypothesis that two estimated clusters $\widehat{C}_a$ and $\widehat{C}_b$ share the same per-gene Poisson mean. Because $\widehat{\mathcal{C}}$ is a function of $X^{(1)}$ alone and $X^{(2)}$ is independent of $X^{(1)}$, the conditional distribution of any test statistic computed from $X^{(2)}$ given $\widehat{\mathcal{C}}$ is the unconditional distribution under the null, and the resulting two-sample Poisson test has nominal type I error.

REMARK 12.6 (Why sample splitting fails here). A natural alternative is to split cells into two halves, cluster on one half, and test on the other. But the clusters identified on the training half are defined by the cells in that half; they may not exist as identifiable structures on the test half, and any rule for projecting them onto the test half (e.g., assign each test cell to the nearest training centroid) introduces a selection event in the projection that data thinning sidesteps by keeping all observations in both $X^{(1)}$ and $X^{(2)}$.

Dharamshi et al. [36] extend the construction to a broader class of distributions by replacing the additive parameter requirement with a sufficient-statistic decomposition. This generalization covers uniform, beta, and Wishart distributions and recovers the original thinning construction for natural exponential families.

Data fission. Leiner et al. [81] propose a closely related but more permissive framework called *data fission*. Rather than requiring an exact convolution decomposition, data fission decomposes a single observation into two parts $(f(X), g(X))$ such that either (i) $f(X)$ and $g(X)$ are independent with known distributions, or (ii) the conditional distribution of $g(X)$ given $f(X)$ is known in closed form. The second condition weakens the independence requirement of data thinning, in exchange for a mandatory "debiasing" step that uses the conditional distribution to compute the test statistic on the inference part.

EXAMPLE 12.7 (Gaussian fission). Let $X \sim N(\mu, \tau^2)$ with τ^2 known. Generate an external $Z \sim N(0, \sigma^2)$ independent of X and define

$$f(X) = X + Z, \qquad g(X) = X - \frac{\tau^2}{\sigma^2} Z.$$

We claim $f(X) \perp g(X)$ and compute their marginals.

PROOF OF THE CLAIM IN EXAMPLE 12.7. The pair $(f(X), g(X))$ is a linear transformation of the jointly Gaussian pair (X, Z), hence jointly Gaussian. Independence then reduces to zero covariance:

$$\mathrm{Cov}(f(X), g(X)) = \mathrm{Cov}(X + Z,\, X - \tau^2 Z/\sigma^2) = \mathrm{Var}(X) - \frac{\tau^2}{\sigma^2}\mathrm{Var}(Z) = \tau^2 - \tau^2 = 0.$$

For the marginals, $f(X) = X + Z$ has mean μ and variance $\tau^2 + \sigma^2$; similarly,

$$\mathbb{E}[g(X)] = \mu, \qquad \mathrm{Var}(g(X)) = \mathrm{Var}(X) + \frac{\tau^4}{\sigma^4}\mathrm{Var}(Z) = \tau^2 + \frac{\tau^4}{\sigma^2}.$$

Selecting hypotheses based on $f(X)$ and conducting inference on $g(X)$ is valid because any function of $f(X)$ is independent of $g(X)$. □

The Gaussian construction illustrates the central trade-off: as $\sigma^2 \to \infty$, $f(X)$ becomes very noisy (less informative for selection) while $g(X)$ approaches X (more informative for inference); as $\sigma^2 \to 0$, the reverse. Choosing $\sigma^2 = \tau^2$ makes f and g symmetric: both views have variance $2\tau^2$, so they are equally informative after the noise injection. The general data fission framework generalizes the Gaussian construction by replacing the τ^2/σ^2 coefficient with the appropriate analogue from the known conditional distribution of $g(X)$ given $f(X)$; for example, when X is Poisson, the role of σ^2 is played by the expected-information ratio between the two components.

REMARK 12.8. Data thinning and data fission both retain the full sample size for inference, unlike sample splitting. The cost is a noise inflation: each component carries only a fraction of the original signal. In practice the inflation is modest for moderate ϵ, and the gain in target stability (latent clusters and structures identified on $X^{(1)}$ apply unchanged to $X^{(2)}$) is decisive in unsupervised settings.

The relationship to the polyhedral and clustering selective inference of the earlier sections is illuminating. Polyhedral selective inference computes the conditional distribution of a pivot given the selection event, which can be combinatorially intricate. Data thinning bypasses the combinatorics by constructing an independent "fresh" dataset for inference, at the cost of distributional assumptions on the data-generating process. Each approach has its place: polyhedral SI is distribution-free in the design matrix and gives exact pivots for fixed-λ Lasso under Gaussian noise; data thinning is parametric in the data distribution but applies to arbitrary selection rules.

10. Selective Inference Beyond Fixed-λ Models

Polyhedral selective inference for the Lasso, as developed in Section 7, conditions on a particular fixed λ and a particular selected active set and sign pattern. In practice, λ is rarely fixed: it is chosen by cross-validation, selected by an information criterion, or chosen adaptively from the data. The selection rule, once it includes the tuning step, is no longer a single polyhedral event, and the truncated-normal pivot of Section 7 does not apply.

Cross-validated Lasso. The work of Loftus and Taylor [88] treats selective inference after cross-validation by characterizing the entire CV path as a union of polyhedral events: each fold's optimal λ corresponds to a sequence of model-selection events, and the CV-selected λ is the value that minimizes the combined CV criterion. Conditioning on the CV-selected λ (and the resulting active set and signs) gives a conditional distribution that is truncated normal on a union of polyhedral regions. The pivot calculation generalizes the single-polyhedron case but requires bookkeeping for each constituent polyhedron. In practice, the truncation set becomes complex, and importance sampling or randomized selective inference is often used to compute the pivot.

Randomized selective inference and data carving. A different approach is to add small randomization to the selection rule itself. Fithian et al. [48] show that adding independent noise to the selection step produces a smoother conditional distribution and tighter selective intervals. When the randomization is small relative to the information content, the selective pivot approaches the unconditional pivot in width, and "data carving" — using both the discovery half and a small fraction of the inference half — recovers most of the lost power.

Selective inference for non-linear models. Taylor and Tibshirani [120] extend selective inference to ℓ_1-penalized likelihood models such as logistic and Poisson regression. The polyhedral characterization no longer applies exactly, but a local quadratic approximation around the selected solution yields an asymptotically valid selective pivot.

Selective inference for change-points and graph saliency. Recent extensions apply selective inference principles to selection events that are not polyhedral at all. Shiraishi et al. [107] develop selective inference for change-point detection by recurrent neural networks: the selection event is the non-linear, temporally constrained sequence of states visited by the RNN, and the conditional distribution under the null is characterized by a sampling scheme. Nishino et al. [95] apply selective inference to saliency maps of graph neural networks: the selected nodes and edges are those flagged by a GNN explainability method, and the conditional distribution under the null accounts for the spatial dependence in the underlying graph. In both cases, the selective pivot is computed by simulating the algorithmic selection process under the null and using the empirical distribution as the reference law.

These extensions all share a common structure: identify the selection event, characterize the conditional distribution of the test statistic given that event, and invert. The polyhedral case of Section 7 is the cleanest instance; the algorithmic cases of change-point and graph saliency analysis are the most complex. Data thinning, by contrast, sidesteps conditioning entirely: it constructs an independent dataset on which any selection rule can be applied without post-selection adjustment.

11. Assumptions in Plain Language

Selective inference is target-specific. A valid interval must name the target and the selection rule. FCR controls an average over selected intervals, not conditional coverage for every selected interval. POSI protects against arbitrary model selection by paying a simultaneous-inference price. Lasso polyhedral inference is sharper because it conditions on a particular, fixed-λ selection rule. Clustering selective inference is valid for the clusters generated by the specified clustering algorithm and conditioning scheme.

The most common mistake is to combine a data-adaptive workflow with a reference law derived for a fixed workflow. If the selection rule is changed, the selective distribution changes as well.

12. Failure Modes and Diagnostics

- FCR intervals may cover the stated fixed targets on average while failing to represent regression-to-the-mean effects in a scientifically useful way.
- POSI intervals may be too wide to answer the practical question if the selection process was actually much more restricted than arbitrary search.
- Lasso selective intervals require the selection event to be explicit. Cross-validation, screening before the Lasso, or tuning by visual inspection adds selection layers that must be included or justified separately.

- Conditioning on very rare selection events can produce unstable truncated-law calculations and very wide intervals.
- Clustering selective inference depends on the exact clustering algorithm, distance, number of clusters, and any preprocessing used before clustering.

Useful diagnostics include reporting the selection rule in enough detail to be replicated, identifying whether the target is fixed or selected-model, checking the number of selected intervals R_{CI}, and reporting when conditioning events are rare enough to make selective intervals unstable.

13. Bibliographic Notes

Sorić's warning about selected confidence intervals is an early statement of the problem [111]. False coverage rate and selected confidence interval adjustments were introduced by Benjamini and Yekutieli [17]. POSI was developed by Berk et al. [19]. The general selective-inference framework that organizes polyhedral and other conditional pivots, including the optimality viewpoint, is due to Fithian et al. [48]. Polyhedral selective inference for the Lasso is due to Lee et al. [77]; related sequential regression procedures are developed in Tibshirani et al. [122]. Selective inference after clustering is developed for hierarchical clustering in Gao et al. [51] and for k-means in Chen and Witten [32].

Data thinning for convolution-closed distributions is due to Neufeld et al. [94]; the sufficient-statistic generalization that covers uniform, beta, and Wishart families is in Dharamshi et al. [36]. Data fission, which decomposes a single observation into independent components or components with known conditional distribution, is from Leiner et al. [81]. Selective inference for cross-validation-selected models was developed by Loftus and Taylor [88]; the related randomization and data-carving techniques are in Fithian et al. [48]. Post-selection inference for ℓ_1-penalized likelihood models is from Taylor and Tibshirani [120]. Recent extensions to RNN change-point detection and graph neural network saliency are from Shiraishi et al. [107] and Nishino et al. [95].

14. Exercises

Basic.

Exercise 12.9 (Selected marginal intervals). Let $Z_i \sim N(\theta_i, 1)$ independently and let

$$C_i = Z_i \pm z_{1-\alpha/2}.$$

Select $\widehat{S} = \{i : 0 \notin C_i\}$. Show that if $\theta_i = 0$, then a selected interval for i never covers θ_i.

Exercise 12.10 (FCR definition). For selected intervals C_i, $i \in \widehat{S}$, define R_{CI}, V_{CI}, and FCR. Letting $\mathrm{FWER}_{\mathrm{CI}} = \mathbb{P}(V_{\mathrm{CI}} \geq 1)$ be the probability that at least one selected interval fails to cover, show that $\mathrm{FCR} \leq \mathrm{FWER}_{\mathrm{CI}}$.

Exercise 12.11 (No-selection case). Suppose all m intervals are reported and each has marginal noncoverage at most α. Show directly from (19) that $\mathrm{FCR} \leq \alpha$, without assuming independence.

Intermediate.

Exercise 12.12 (The $R^{(i)}$ construction). Let $\widehat{S} = \{i : |T_i| > c\}$. For a selected i, compute $R^{(i)}$ in (20). Then repeat the calculation when $\widehat{S}$ consists of the indices of the k largest $|T_i|$'s.

Exercise 12.13 (FCR proof). Fill in the conditioning step in Theorem 12.1. Where is independence of the T_i's used? Construct a short explanation of why the same proof does not automatically apply under arbitrary dependence.

EXERCISE 12.14 (POSI target). For a fixed design with three predictors, write the selected-model target $\beta_{j\bullet M}$ for each two-variable model M. Give an example in which $\beta_{j\bullet M}$ differs from the coefficient of j in the full three-variable model.

EXERCISE 12.15 (KKT polyhedron). Consider the Lasso with fixed λ, two predictors, and selected active set $M = \{1\}$ with positive sign. Write the KKT conditions and express them as linear inequalities in y, treating X as fixed.

Computational.

EXERCISE 12.16 (Selected-interval simulation). Reproduce the common-mean simulation in Figure 2. Compare the FCR of unadjusted selected intervals, Bonferroni intervals, and the Benjamini–Yekutieli selected-interval adjustment as m changes. A successful reproduction should show the unadjusted selected intervals exceeding the target FCR, with Bonferroni conservative and the Benjamini–Yekutieli adjustment closer to the target; see `fig10_fcr_intervals()` in `scripts/make_figures.R`.

EXERCISE 12.17 (POSI constants by simulation). For a small fixed design with $p = 5$, enumerate all nonempty models M and simulate the POSI statistic

$$\max_M \max_{j\in M} |Z_{j\bullet M}|.$$

Estimate its 95 percent quantile and compare it with the Bonferroni cutoff and the ordinary 1.96 cutoff.

EXERCISE 12.18 (Truncated-normal pivot). Simulate $y \sim N(\mu, I_n)$ and a simple polyhedral event $Ay \le b$. For a chosen contrast η, compute $V^-(z)$ and $V^+(z)$ after conditioning on $P_{\eta^\perp} y = z$. Check by simulation that the truncated-normal CDF pivot is approximately uniform conditional on the event.

Advanced.

EXERCISE 12.19 (Cross-validation as selection). Explain why lasso selective inference with fixed λ does not automatically cover the workflow in which λ is chosen by cross-validation. What additional information would need to be included in the selection event?

EXERCISE 12.20 (Clustering truncation set). Let $W = (W_1, W_2, W_3) \sim N(\mu, \sigma^2 I_3)$ and let the clustering rule be the fixed threshold split

$$C_1(W) = \{i : W_i \le 0\}, \qquad C_2(W) = \{i : W_i > 0\}.$$

For the observed vector $w = (-1.0, -0.4, 0.8)$, the selected clusters are $C_1 = \{1, 2\}$ and $C_2 = \{3\}$. Let

$$v_i = \mathbf{1}\{i \in C_1\}/|C_1| - \mathbf{1}\{i \in C_2\}/|C_2|.$$

Condition on $P_v^\perp W = P_v^\perp w$, and parametrize the remaining line by $W(\phi) = P_v^\perp w + \phi v/\|v\|_2^2$. Derive the interval of ϕ values for which the threshold split still returns $C_1 = \{1, 2\}$ and $C_2 = \{3\}$. Compare the naive Gaussian p-value for $v^\top W$ with the selective p-value obtained from the normal law truncated to this interval.

EXERCISE 12.21 (Choosing the criterion). Give a data-analysis scenario where FCR control is the appropriate goal, one where POSI is more defensible, and one where a fully conditional selective interval is necessary. For each case, state the target parameter precisely.

EXERCISE 12.22 (Gaussian data fission). Let $X \sim N(\mu, \sigma^2)$ with σ^2 known. Let $Z \sim N(0, \sigma^2)$ independent of X and define $X^{(1)} = \sqrt{\epsilon} X + \sqrt{1-\epsilon} Z$ and $X^{(2)} = \sqrt{1-\epsilon} X - \sqrt{\epsilon} Z$. Show that $X^{(1)} \perp X^{(2)}$, and that $X^{(1)} \sim N(\sqrt{\epsilon}\mu, \sigma^2)$ and $X^{(2)} \sim N(\sqrt{1-\epsilon}\mu, \sigma^2)$. Explain how

this construction allows clustering on $X^{(1)}$ and inference on $X^{(2)}$ without any post-selection adjustment.

Exercise 12.23 (Poisson thinning). Let $X \sim \text{Poisson}(\lambda)$ and, conditional on X, draw $X^{(1)} \sim \text{Binomial}(X, \epsilon)$; set $X^{(2)} = X - X^{(1)}$. Show that $X^{(1)}$ and $X^{(2)}$ are independent and marginally Poisson with rates $\epsilon\lambda$ and $(1-\epsilon)\lambda$. Apply this construction to a count-matrix simulation of single-cell RNA-seq: cluster cells on $X^{(1)}$ using k-means and test, on $X^{(2)}$, whether two clusters differ in gene-expression mean. Verify the type I error rate matches the nominal level.

Exercise 12.24 (Data fission for Gaussian). For $X \sim N(\mu, \tau^2)$ and $Z \sim N(0, \sigma^2)$ independent, define $f(X) = X + Z$ and $g(X) = X - (\tau^2/\sigma^2)Z$. Verify that $f(X)$ and $g(X)$ are independent, compute their marginal distributions, and discuss how the choice of σ^2 trades off between the noise added to $f(X)$ and the noise added to $g(X)$. How does this compare with sample splitting in terms of effective sample size?

Exercise 12.25 (CV-selected lambda as a compound selection event). For Lasso regression with cross-validation choosing λ from a finite grid $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$, explain why the event "CV selects λ_k" is not a single fixed-λ Lasso polyhedron. After conditioning on the active sets and signs along the foldwise Lasso paths, describe how the event is represented as a union of many selection regions. How does the number of pieces grow with K and the possible foldwise active sets? Discuss why importance sampling or randomization is typically needed to compute the corresponding selective pivot.

CHAPTER 13

Applications in Genomics and Large Language Models

The methods in this book were developed for statistical inference, not for a single scientific area. They become most useful when an application has three features at the same time: many questions, strong dependence, and a sharp distinction between a calibrated claim and an attractive story. Genomics is the classical example. A single experiment may measure expression levels for thousands of genes or test millions of variants. Large language models give a new example. A model may be evaluated on many benchmarks, audited for watermarks or contamination, and asked to make factual claims whose reliability must be calibrated.

The two domains look different on the surface. The statistical grammar is the same. We need to define the null hypothesis carefully, choose the family of claims before inspecting the most interesting cases, account for dependence, and report a statement whose error rate is the one we actually want. This chapter is therefore a capstone rather than a collection of unrelated case studies. Genomics shows how FDR, local FDR, knockoffs, and high-dimensional regression work in a mature scientific workflow. LLM applications show how exchangeability, p-values, multiple testing, and conformal prediction reappear in modern model evaluation and safety auditing.

1. A Shared Workflow

Figure 1 shows a stylized genomic discovery pipeline. The important point is not the biology in the boxes; it is the placement of inference after preprocessing and before scientific follow-up. Quality control and normalization are allowed to use scientific knowledge, but the error rate should be attached to a family of hypotheses that has been defined before individual discoveries are interpreted.

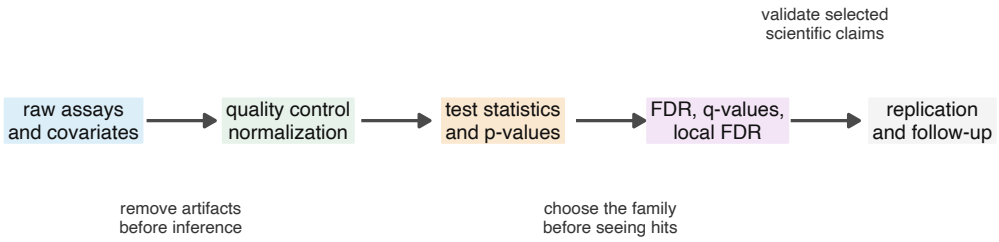

FIGURE 1. A large-scale genomic analysis pipeline. Multiple-testing methods enter after the measurement scale and test family have been fixed. The selected discoveries then need replication, biological validation, or a second-stage inferential analysis.

An LLM evaluation pipeline has the same structure. Raw model outputs are converted into scores, the scores are compared across many tasks or prompts, and the analyst must decide whether a reported improvement, watermark signal, or contamination signal is strong enough to survive the relevant multiplicity. The rest of the chapter uses this analogy repeatedly.

2. Genomic Screening

Suppose m genes or variants are tested. For each feature i there is a null hypothesis H_{0i}, a test statistic Z_i, and a p-value p_i. In a differential-expression study, H_{0i} might state that gene i has the same mean expression in two conditions. In a genome-wide association study, H_{0i} might state that a variant has no marginal association with a phenotype after adjustment for pre-specified covariates.

The first temptation is to sort the p-values and report the smallest ones. That is not an inferential procedure. If $m = 20{,}000$, then even perfectly null data will produce many small-looking p-values. The question is not whether the best gene looks interesting; it is how many false discoveries are expected among the genes declared interesting. This is why the Benjamini–Hochberg procedure became central in genomics [15, 40].

For a target FDR level q, BH rejects

$$H_{0(1)}, \dots, H_{0(\widehat{k})}, \qquad \widehat{k} = \max\left\{k : p_{(k)} \le \frac{qk}{m}\right\},$$

where $p_{(1)} \le \cdots \le p_{(m)}$. The selected set $\mathcal{R}$ is a discovery list, not a list of guaranteed truths:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] \le q$$

under the assumptions of the procedure. This average statement is often more scientifically useful than familywise error control, because the goal is to produce a stable set of candidates for follow-up rather than to certify that no false positive exists anywhere.

Storey's q-value and Efron's local FDR refine this view [42, 43, 112]. A Storey q-value Q_i can be read as the smallest FDR level at which feature i would enter the discovery set. The local FDR, defined under a two-groups empirical-Bayes model in which each feature is null with prior probability π_0 and nonnull with probability $1 - \pi_0$, is

$$\mathrm{lfdr}(z) = \mathbb{P}(H_{0i} \text{ is true} \mid Z_i = z)$$

and can be read as a posterior null probability under that mixture model. BH, q-values, and local FDR answer related but distinct questions. BH controls an error rate over a selected set; Q_i records where a feature enters such a set; $\mathrm{lfdr}(z)$ summarizes feature-level evidence under a model for the mixture of null and nonnull effects.

3. Dependence, LD, and Structured Discoveries

Genomic p-values are rarely independent. Nearby variants are correlated by linkage disequilibrium, expression measurements share batch and pathway effects, and sample-level covariates can create broad shifts across many features. Dependence does not merely change a proof condition. It changes the shape of the evidence. A region with many correlated variants may produce many small p-values even when there is one underlying signal.

Several responses are possible.

- If the dependence is positive and satisfies the PRDS-type conditions from Chapter 6, BH remains valid.
- If the analyst wants protection under arbitrary dependence, the Benjamini–Yekutieli correction is valid but can be conservative.
- If the goal is to identify regions or pathways, the hypotheses should be defined at the region or pathway level rather than pretending that every correlated variant is a separate biological discovery.
- If the target is conditional variable importance, knockoffs or CRTs are more appropriate than marginal p-values.

The last point is particularly important in genetics. Let

$$X \in \mathbb{R}^{n \times p}$$

be a genotype or feature matrix and let y be a phenotype. A marginal test of X_j asks whether X_j is associated with y. A conditional null asks whether

$$Y \perp X_j \mid X_{-j}.$$

These are different null hypotheses when features are correlated. Concretely, suppose variant X_j is in tight linkage disequilibrium with a causal variant X_k: X_j carries no independent biological signal once X_k is conditioned on, yet a marginal test of X_j versus y will reject because of the correlation. The conditional null $Y \perp X_j \mid X_{-j}$ is the one a model-X knockoff procedure targets, and X_j is null under that null even though it is highly significant marginally. Model-X knockoffs construct artificial variables $\widetilde{X}_j$ that act as negative controls for the original variables X_j. Feature statistics W_j are built so that large positive values favor the original variable, large negative values favor its knockoff, and null signs are symmetric. The knockoff threshold estimates how many false positives are entering the selected set [6, 30].

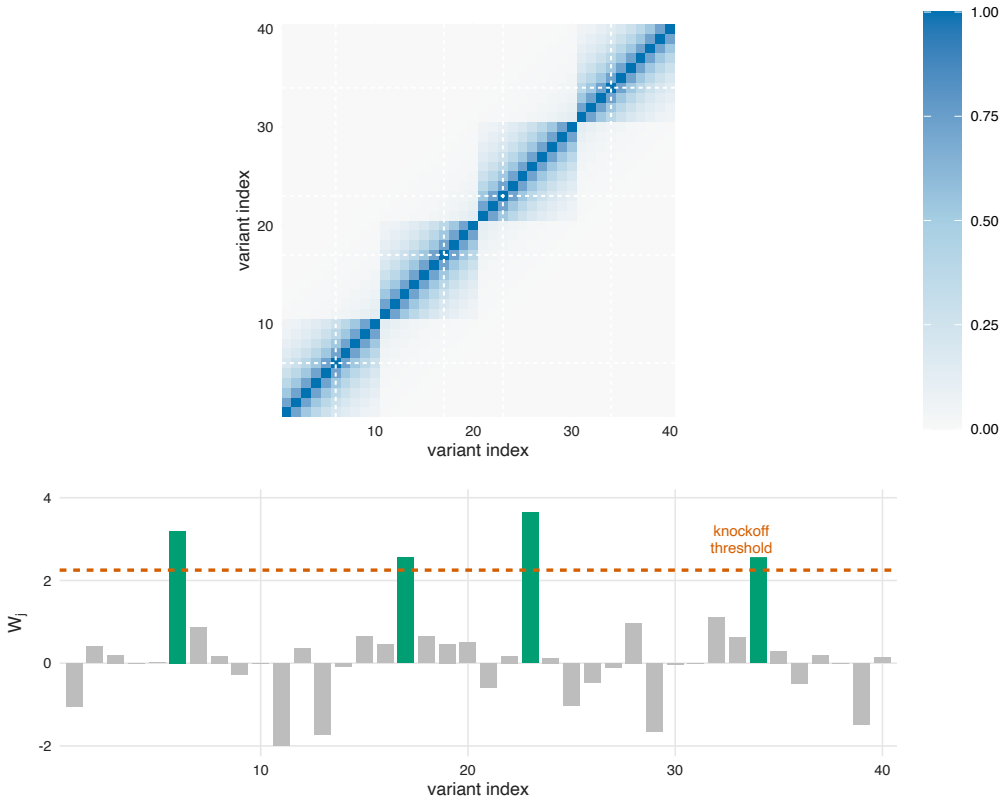

FIGURE 2. A schematic GWAS dependence problem. The heatmap displays an LD-like correlation structure among variants. The lower panel shows knockoff feature statistics W_j, where large positive values indicate evidence for conditional importance and the threshold is chosen to control FDR.

Debiased Lasso intervals from Chapter 11 can be used for pre-specified variants or low-dimensional contrasts. Knockoffs are more natural when the primary objective is discovery among many correlated features. Selective-inference ideas from Chapter 12 enter after a set of

genes or variants has already been selected and the analyst wants intervals for selected targets. The methods are complementary because they attach error control to different inferential targets.

4. Benchmark Multiplicity for Language Models

Large language models are usually compared across many tasks, datasets, metrics, prompts, and decoding settings. Let Δ_{ab} denote the performance difference between a model and a baseline on task b, or between two models a and a' on task b. Each comparison may have a standard error from paired examples, bootstrap resampling, or a binomial model for success/failure outcomes. If many comparisons are inspected, then the largest improvements will be noisy in the same way that the smallest genomic p-values are noisy.

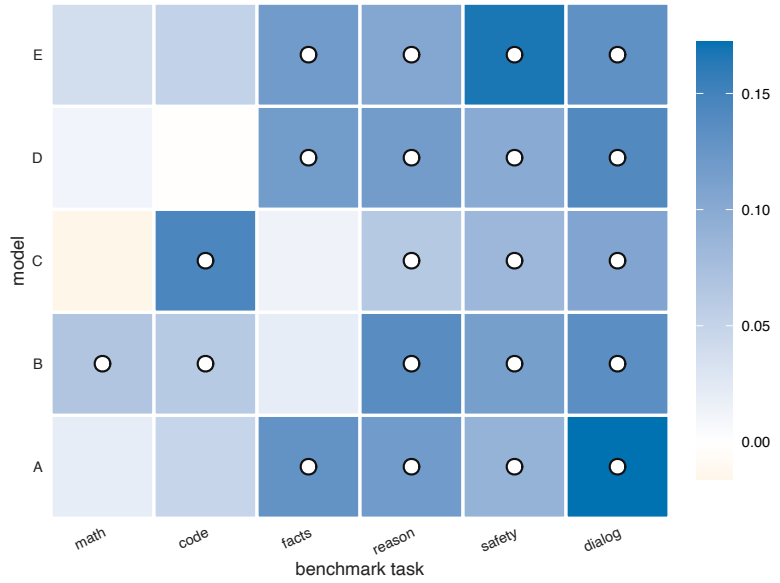

FIGURE 3. Benchmark multiplicity for model evaluation. Each cell is a model-by-task comparison, and circled cells are discoveries after a BH-style adjustment. Reporting only the best cells without multiplicity control overstates the strength of evidence.

There is no universal error rate for leaderboards. If the scientific claim is "model A improves this one pre-specified metric," then an ordinary paired test or confidence interval may be appropriate. If the claim is "model A improves many tasks," the family is the collection of task-level tests. If the claim is "we found the best model among many models and tasks," the selection process must be included in the uncertainty statement. The same distinction appeared in Chapters 4 and 12: the reported claim determines the family.

Dependence is substantial. Benchmark examples overlap in topic and difficulty; models share pretraining data and architectures; metrics can be correlated within a task. This does not make inference impossible, but it discourages naive independence calculations. Paired designs, bootstrap procedures that resample examples rather than cells, hierarchical testing plans, and FDR-controlling procedures are often more defensible than a table of unadjusted p-values. Hierarchical FDR procedures from Chapter 7 are a natural fit when benchmarks are nested inside capability domains. AI governance documents such as National Institute of Standards and Technology [93] and benchmark-design work such as Liang et al. [85] provide the policy framing for how these multiplicity-adjusted reports should be presented to decision makers.

5. LLM-as-a-Judge and Latent Entanglement

A particularly subtle multiplicity problem arises when language models are themselves used to evaluate the outputs of other language models. The LLM-as-a-Judge paradigm has become standard for evaluating tasks where ground-truth answers are expensive or impossible to collect: ask a strong LLM to score the output of a candidate model, treat the score as a measurement, and average across many examples.

The statistical question is whether the judge LLM provides an independent measurement signal. The answer is generally no. Models that share pretraining data, tokenizers, distillation pipelines, or alignment via Reinforcement Learning from Human Feedback exhibit correlated reasoning patterns and synchronized failure modes. When the judged model and the judging model belong to the same pretraining family — or even when they were trained on related corpora — the judge's score for a given prompt is not statistically independent of the candidate model's output.

Mavrogiannis et al. [90] formalize this as a problem of *behavioral entanglement*: ensembles of judges that look superficially independent share correlated error modes that produce *confabulation consensus* — apparent agreement that reflects shared bias rather than independent validation. Statistically, the analyst is using an estimator that under-reports its variance because the contributing measurements are positively correlated. Sun et al. [116] propose multi-agent reasoning-tree auditing as an alternative to flat majority voting, demonstrating that conventional LLM-as-Judge ensembles can be Pareto-dominated by procedures that account for reasoning-trace correlations.

For our purposes the import is methodological: an LLM-as-Judge evaluation is not a flat multiple-testing family of independent measurements. It is at best a clustered family in which the cluster structure is partially observable through pretraining-data and architecture metadata. The hierarchical and group-aware procedures of Chapter 7 therefore suggest an analysis plan once cluster-level p-values or e-values and their dependence assumptions have been specified: control error at the cluster (model-family) level, and treat within-cluster variance as nuisance. Cumulative Information Gain and similar dependence-corrected metrics offer one operationalization [90]. Selective inference (Chapter 12) gives a second planning principle: when an analyst picks the "best" judging ensemble from multiple candidates, the post-selection inference target should be explicitly the selected ensemble's performance, not the population performance.

6. Watermarked Generation

Watermarking asks whether a piece of text was generated using a known hidden randomization scheme. The typical setting has a trusted model provider, a secret key sequence, and a public text that may have been edited. A detector uses the key to test whether the text carries the statistical signature of the watermark. Modern watermarking methods differ in their engineering details, but the inferential core is a hypothesis test [72, 102].

Let $\mathcal{A}$ be the vocabulary and $K = |\mathcal{A}|$. Write $Y_{<t}$ for the history of tokens generated before step t (truncated to the context window when one is in force). At time t, the autoregressive model assigns probabilities

$$p_t(k) = \mathbb{P}(Y_t = k \mid Y_{<t} = y_{<t}), \qquad k \in \mathcal{A}.$$

A keyed decoder receives p_t and a key ξ_t drawn from a specified key distribution that is independent of $Y_{<t}$, and outputs

$$Y_t = \Gamma(\xi_t, p_t).$$

In practice the ξ_t across positions are produced by a pseudorandom function of a master secret and the position, so each token uses fresh keyed randomness but the entire sequence is reproducible

by anyone holding the master secret. The watermark should not degrade the distribution of model outputs in a way that is visible to ordinary users. This motivates the following definition.

Definition 13.1 (Distortion-free keyed generation). A keyed decoder Γ is distortion-free for a key distribution under which ξ_t is independent of $Y_{<t}$ if

$$\mathbb{P}\{\Gamma(\xi_t, p_t) = k \mid Y_{<t} = y_{<t}\} = p_t(k), \qquad k \in \mathcal{A}.$$

Thus the secret key changes the coupling between the text and the key, but not the marginal distribution of the next token.

For a binary vocabulary $\mathcal{A} = \{0, 1\}$, a simple distortion-free decoder is

$$Y_t = \mathbf{1}\{U_t \le p_t(1)\}, \qquad U_t \sim \mathrm{Unif}[0, 1].$$

Another distortion-free decoder used in practice is exponential minimum sampling. Assume for the moment that $p_1, \dots, p_K > 0$ and $\sum_{k=1}^{K} p_k = 1$. Draw independent

$$U_k \sim \mathrm{Unif}[0, 1], \qquad E_k = \frac{-\log U_k}{p_k}.$$

Then $E_k \sim \mathrm{Exp}(p_k)$, where the parameter is the rate, and the decoder chooses the token with the smallest exponential variable:

$$Y = \operatorname*{arg\,min}_{1 \le k \le K} E_k.$$

Proposition 13.2 (Exponential minimum sampling is distortion-free). *With $E_1, \dots, E_K$ as above,*

$$\mathbb{P}(Y = k) = p_k, \qquad k = 1, \dots, K.$$

Proof. The minimum of independent exponential variables with rates $p_1, \dots, p_K$ has total rate $\sum_j p_j = 1$. For a fixed k,

$$\mathbb{P}(Y = k) = \int_0^\infty \mathbb{P}(E_j > t \text{ for all } j \ne k)\, p_k e^{-p_k t}\, dt.$$

Independence gives

$$\mathbb{P}(E_j > t \text{ for all } j \ne k) = \exp\left(-t \sum_{j \ne k} p_j\right).$$

Therefore

$$\mathbb{P}(Y = k) = \int_0^\infty p_k e^{-t \sum_j p_j}\, dt = p_k.$$

Tokens with $p_k = 0$ can be excluded or assigned $E_k = \infty$. □

7. Watermark Detection and Scanning

Let $\widetilde{y}_{1:\ell}$ be a public text segment and let ξ denote the secret key sequence. A detector computes a statistic

$$\phi(\xi, \widetilde{y}_{1:\ell}),$$

where larger values mean stronger agreement between the text and the key. The null hypothesis is that the text was not generated using this watermark key. The alternative is that the text was generated, or partly generated, by the keyed decoder.

The cleanest p-value is obtained by key resampling. Draw independent fake keys $\xi^{(1)}, \dots, \xi^{(B)}$ from the same key distribution and compute their scores against the same text. If large scores are evidence for the watermark, use

$$p_{\mathrm{wm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\phi(\xi^{(b)}, \widetilde{y}_{1:\ell}) \ge \phi(\xi, \widetilde{y}_{1:\ell})\}}{B + 1}. \tag{25}$$

The direction of the inequality is part of the method. If the statistic were defined so that smaller values were more suspicious, the inequality would have to be reversed.

PROPOSITION 13.3 (Key-resampling p-value). *Under the null hypothesis that $\widetilde{y}_{1:\ell}$ is independent of the true key and the resampled keys, the p-value in (25) is super-uniform:*

$$\mathbb{P}(p_{\mathrm{wm}} \le \alpha) \le \alpha.$$

PROOF. Conditional on the text, the $B+1$ scores

$$\phi(\xi, \widetilde{y}_{1:\ell}), \phi(\xi^{(1)}, \widetilde{y}_{1:\ell}), \ldots, \phi(\xi^{(B)}, \widetilde{y}_{1:\ell})$$

are exchangeable under the null. Therefore the rank of the true-key score among these $B+1$ scores is uniform up to ties. Counting the number of resampled scores at least as large as the observed score gives the usual randomization p-value, with the added 1 in numerator and denominator ensuring finite-sample validity. □

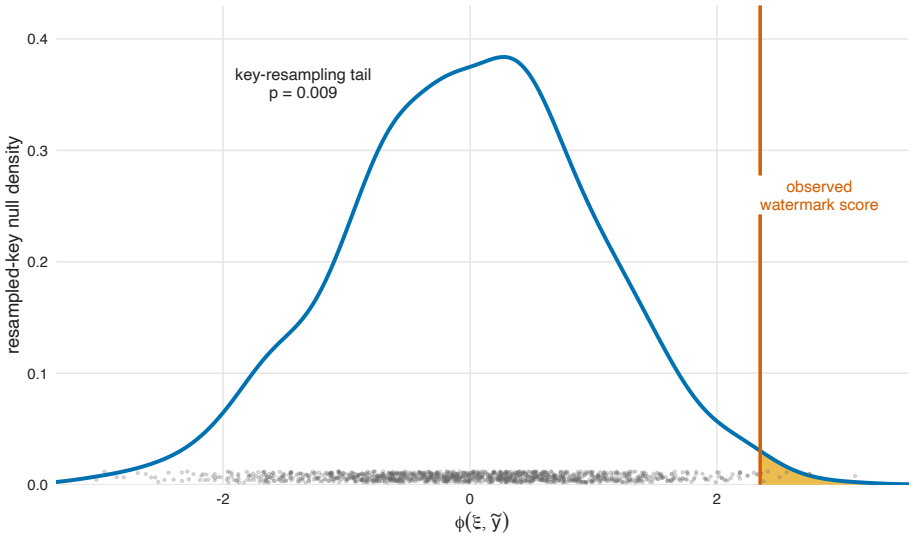

FIGURE 4. A watermark test calibrated by resampling keys. When larger statistics mean stronger watermark evidence, the p-value is the upper-tail fraction of resampled-key scores that exceed the observed-key score.

Editing creates an additional search problem. A long document may contain a short watermarked passage, or a watermarked passage may have been moved. A substring detector might scan over key locations a, text locations c, and a fixed block length L:

$$\phi_{\max} = \max_{a,c} \mathcal{M}\big(\xi_{a:a+L-1}, \widetilde{y}_{c:c+L-1}\big),$$

where $\mathcal{M}$ is a local match score. This is a multiple-testing problem in disguise. Calibrating a single window and then taking the largest window score would be anti-conservative. One should either use a Bonferroni correction over the scan or, more powerfully, resample the entire maximum statistic:

$$\phi^{(b)}_{\max} = \max_{a,c} \mathcal{M}\big(\xi^{(b)}_{a:a+L-1}, \widetilde{y}_{c:c+L-1}\big), \qquad b = 1, \ldots, B,$$

using resampled keys $\xi^{(b)}$ indexed by b. The reference law must include the same search that was used to find the reported match.

Tokenization multiplicity. A subtle complication of any token-level watermark or auditing scheme is *tokenization multiplicity*. With a fixed deterministic tokenizer, the encoding of a displayed string is usually unique. The statistical issue arises when the testing protocol accepts token sequences, byte strings, provider-side encodings, or other representations that are not canonically tied to the displayed text. In that broader protocol, an identical semantic string can be represented by multiple distinct token sequences. For example, "Multilingual" might decode as (`Multi`, `ling`, `ual`), (`Mu`, `lti`, `lingual`), or (`M`, `ulti`, `lin`, `gual`), each carrying a different probability under the model.

Artola Velasco et al. [5] show that this multiplicity is not merely a theoretical curiosity: an LLM service provider can exploit alternative tokenizations to inflate token counts (and therefore pay-per-token charges) by arbitrary amounts *without altering the displayed text*. For statistical auditing, the same multiplicity can attack watermark detectors if the protocol lets a service provider choose, after the fact, a representation that minimizes the detection statistic while preserving the displayed output.

The statistical remedy is *canonical generation*. Define a deterministic mapping from semantic strings to a unique canonical token sequence, and compute the watermark detection statistic only on the canonical sequence. This restores exchangeability of the test statistic across alternative provider behaviors, at the cost of constraining the tokenization to a fixed rule. The constraint can be enforced cryptographically by hashing the canonical token sequence and including the hash in the watermark verification protocol.

Verifiable oblivious watermark detection. A second cryptographic refinement targets the user-privacy implications of watermark detection. In the basic detection protocol, the user submits the generated text to the provider for verification; the provider learns the text content and can use it for training or auditing. When the underlying text is sensitive — legal documents, medical communications, or private correspondence — this raises a privacy concern.

Fairoze et al. [45] propose *verifiable oblivious watermarking* (VOW): the user can verify watermark presence with the provider's cooperation, but the provider learns nothing about the text being verified. The construction combines a watermark embedding scheme with a verifiable oblivious pseudorandom function. The detection statistic is computed inside a secure two-party computation between the user and the provider, with the user contributing the text and the provider contributing the watermark key; the output of the computation is a p-value, but neither party learns the other's input.

For our statistical purposes, the key point is conditional: if the VOW implementation exactly evaluates the same detection statistic with the same key distribution and canonicalization rule as the public detector, then the resulting p-value has the same null distribution as the public detection p-value. Under that protocol-level equivalence, verifiable oblivious detection inherits the validity statements developed in this chapter without modification.

8. Contamination Auditing by Exchangeability

Benchmark contamination asks whether a model had access to evaluation data during training or tuning. String matching can be useful evidence, but it is not a statistical proof by itself. A strong audit states a null hypothesis under which a test statistic has a known or resampling-calibrated reference distribution.

The exchangeability test of Oren et al. [96] uses a simple idea. Let

$$X = (X_1, \dots, X_n)$$

be a benchmark dataset in its canonical order, and let $\mathcal{P}$ be the language model being audited. Write $T(X)$ for a score that should be large if the model assigns unusually high likelihood to

the canonical sequence. For the test to have power, $T(X)$ must actually depend on the order: scoring rules that evaluate each example independently and then sum or average produce a score that is invariant under permutations, giving zero power. A sequence-level model evaluation that conditions on previously seen examples, or a score that explicitly uses adjacent-example structure, is needed. A canonical choice with this property is

$$T(X) = \log \mathcal{P}(X)$$

when $\mathcal{P}$ scores the whole sequence with cross-example conditioning. The null hypothesis is not "the model is generally clean." A sharper null is that $\mathcal{P}$ is independent of the benchmark dataset and that the benchmark examples are exchangeable under permutations relevant to the audit. Under this null, the canonical ordering is not special.

PROPOSITION 13.4 (Permutation p-value for canonical-order contamination). *Let $\pi_1, \ldots, \pi_B$ be independent random permutations of $\{1, \ldots, n\}$. If X is exchangeable and independent of the audited model under the null, then*

$$p_{\text{contam}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T(X_{\pi_b}) \geq T(X)\}}{B + 1}$$

is a valid Monte Carlo permutation p-value when large T is evidence of contamination.

PROOF. Under the null, the canonical ordering and the randomly permuted orderings are exchangeable. Conditional on the unordered collection of examples, the scores

$$T(X), T(X_{\pi_1}), \ldots, T(X_{\pi_B})$$

therefore have an exchangeable joint distribution. The rank argument used in Proposition 13.3 applies verbatim. □

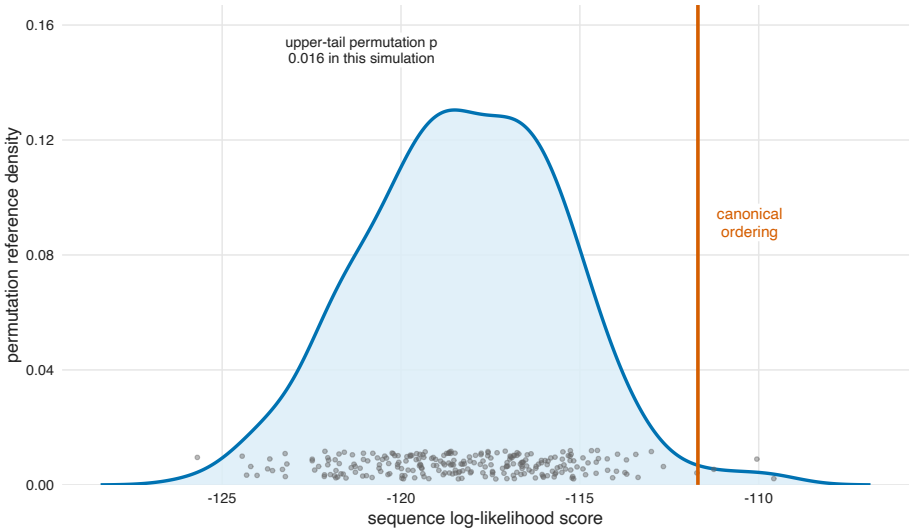

FIGURE 5. A canonical-versus-permuted contamination audit. If the canonical ordering receives much larger likelihood than random orderings, the upper-tail permutation p-value is small. The test supports a specific contamination claim only under the exchangeability and scoring assumptions.

The canonical ordering matters. If the benchmark has a natural order that could have appeared in training, a contaminated model may assign high probability to the transitions

between consecutive examples. Shuffling the examples breaks those transitions. Conversely, if examples are not exchangeable because the benchmark is deliberately ordered by difficulty or topic, the null reference distribution needs to preserve that structure.

Watermark-based contamination tests are another route [102]. If a benchmark or synthetic dataset is released with a known watermark, then a later model can be audited for the statistical signature of that watermark. The logic is again a hypothesis test. The null must specify how the watermark statistic behaves when the model was not trained on the watermarked data, and the p-value must be calibrated with the same statistic that will be reported.

When many datasets, models, or watermark keys are audited, the resulting p-values form a family. A single small p-value among thousands of audits is not the same claim as a pre-specified audit. FWER, FDR, or e-value methods from earlier chapters should be chosen according to the intended conclusion: one certified violation, a list of suspicious benchmarks, or an ongoing audit that may continue as new models are released.

9. Conformal Factuality

Watermarking and contamination auditing are defensive evaluations of a model. Conformal factuality uses the same exchangeability logic to modify model outputs so that they satisfy a coverage guarantee. The construction below follows the statistical structure of Mohri and Hashimoto [92].

Let $x \in \mathcal{X}$ be a user input and let a language model produce a base answer $A_0(x)$. Let $y^\star \in \mathcal{Y}$ denote a reference answer or reference knowledge object. We write

$$y^\star \Rightarrow a$$

to mean that the reference entails the answer a. For an answer a, define its entailment set

$$E(a) = \{y \in \mathcal{Y} : y \Rightarrow a\}.$$

The answer a is factual for reference $y^\star$ exactly when

$$y^\star \in E(a).$$

The challenge is that the base answer may be too specific. A back-off family $\{F_t : t \in \mathcal{T}\}$ indexed by a totally ordered threshold set $\mathcal{T} \subseteq \mathbb{R}$ (so the infimum below is well-defined) turns the base answer into less specific answers as t increases. For example, $F_t(x)$ may remove claims whose verifier scores are below a threshold, replace a precise numerical claim by a range, or eventually abstain from making any factual claim. The induced entailment sets should be nested:

$$E(F_t(x)) \subseteq E(F_{t'}(x)) \qquad \text{for } t \le t'.$$

Larger t means a safer but less informative answer.

For a calibration pair $(X_i, Y_i^\star)$, define the strict-safe score

$$S(X_i, Y_i^\star) = \inf \{t \in \mathcal{T} : Y_i^\star \in E(F_s(X_i)) \text{ for every } s \ge t\}. \tag{26}$$

The phrase "strict-safe" is useful. It is not enough that one threshold happens to be safe; every more conservative threshold should also be safe. This avoids pathologies when the back-off path is not perfectly monotone in its surface form. The infimum is over a possibly empty set: if no level of back-off entails the reference (for example, when the prompt is unanswerable in the chosen family), the score is set to ∞, and S takes values in $\mathcal{T} \cup \{\infty\}$. In that case the conformal quantile may also equal ∞, and the conformalized output abstains from a factual claim. For the theorem below, assume the upper-safe set in (26) is closed in $\mathcal{T}$ for every $(x, y^\star)$; this is automatic when $\mathcal{T}$ is a finite grid.

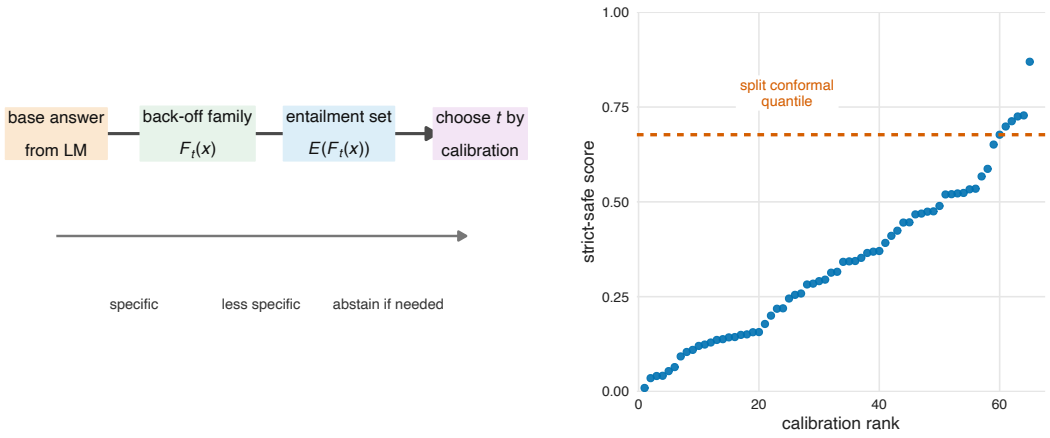

FIGURE 6. Conformal factuality. A back-off family produces less specific answers whose entailment sets are easier for the reference to satisfy. Split conformal calibration chooses the back-off threshold from strict-safe scores.

Let $(X_1, Y_1^\star), \dots, (X_n, Y_n^\star)$ be calibration examples and let $S_i = S(X_i, Y_i^\star)$. For $\alpha \geq 1/(n+1)$, set

$$k = \lceil (n+1)(1-\alpha) \rceil$$

and let $\widehat{q}_\alpha$ be the kth order statistic of $S_1, \dots, S_n$. The conformalized output is

$$\widetilde{\mathcal{P}}(x) = F_{\widehat{q}_\alpha}(x).$$

THEOREM 13.5 (Split conformal factuality). *Assume*

$$(X_1, Y_1^\star), \dots, (X_n, Y_n^\star), (X_{n+1}, Y_{n+1}^\star)$$

are exchangeable. Assume also that the back-off family is sound in the sense that a sufficiently large threshold abstains or makes a claim entailed by every reference in the target class, and that each upper-safe set defining (26) is closed so its infimum is safe. Then

$$\mathbb{P}\left\{ Y_{n+1}^\star \in E(\widetilde{\mathcal{P}}(X_{n+1})) \right\} \geq 1 - \alpha.$$

PROOF. The scores

$$S_1, \dots, S_n, S_{n+1} = S(X_{n+1}, Y_{n+1}^\star)$$

are exchangeable because they are deterministic functions of exchangeable pairs. By the split conformal rank argument from Chapter 10,

$$\mathbb{P}(S_{n+1} \leq \widehat{q}_\alpha) \geq 1 - \alpha.$$

On the event $S_{n+1} \leq \widehat{q}_\alpha$, the upper-safe set property and the closedness assumption imply that $\widehat{q}_\alpha$ itself is a safe threshold, so

$$Y_{n+1}^\star \in E(F_{\widehat{q}_\alpha}(X_{n+1})).$$

This is exactly the desired factuality event. □

The guarantee is marginal over future prompts drawn like the calibration prompts. It is not a pointwise guarantee for every possible prompt. It also inherits the definition of entailment used to score the calibration examples. If entailment is judged by humans, the target is human-labeled factuality. If entailment is judged by an automated verifier, the guarantee is only as meaningful as that verifier and its calibration labels.

10. What the Assumptions Are Doing

The applications in this chapter are useful precisely because the assumptions are visible.

In genomics, the main threats are measurement artifacts, population structure, dependence among features, and target confusion. A marginal association test does not become a conditional importance statement just because the p-value is small. A selected gene list controlled at FDR q does not imply that every gene has posterior null probability at most q. A debiased Lasso interval for a pre-specified coefficient is not the same object as an interval after choosing a model.

In LLM watermarking, the p-value is only valid for the statistic and search procedure that were calibrated. If the analyst scans many substrings, edits the statistic after seeing the text, or reports the best key among many keys, the reference law must include that search. In contamination auditing, exchangeability is the engine of the test. If the benchmark ordering is not exchangeable under the null, the permutation p-value is not justified without modification. In conformal factuality, the coverage statement is marginal and distributional; it does not remove the need to define the deployment population and the entailment target.

11. Bibliographic Notes

For genomic multiple testing, see Dudoit and van der Laan [40] for a broad treatment of multiple testing in genomics and Efron [43] for empirical Bayes and local FDR. Knockoffs were introduced for fixed designs by Barber and Candès [6] and extended to the model-X setting by Candès et al. [30]; hidden Markov model knockoffs for genome-wide association studies are developed by Sesia et al. [104]. Standard data resources for human genetic association studies include the NHGRI-EBI GWAS Catalog [110] and the GTEx tissue-expression atlas [55], which provide vetted summary statistics and annotation links underlying many of the multiple-testing analyses in this chapter.

For statistical significance testing in natural-language-processing benchmark evaluation, see the survey of Dror et al. [39]. The Holistic Evaluation of Language Models framework of Liang et al. [85] provides the operational template for benchmarking AI systems across many metrics; the NIST AI Risk Management Framework [93] gives the policy framing for how these evaluations should be reported. Green-token watermarking for LLM outputs is due to Kirchenbauer et al. [72]; distortion-free watermarks based on exponential minimum sampling are developed by Kuditipudi et al. [74], and benchmark-level watermarking is studied by Sander et al. [102]. Tokenization multiplicity and its consequences for pricing and watermark auditing are documented by Artola Velasco et al. [5]. Verifiable oblivious watermark detection, which preserves user privacy through a secure two-party computation, is from Fairoze et al. [45]. The exchangeability-based contamination test is developed by Oren et al. [96]. The conformal factuality framework is due to Mohri and Hashimoto [92].

LLM-as-a-Judge latent entanglement is analyzed by Mavrogiannis et al. [90], who propose a cumulative-information-gain metric to quantify behavioral correlation between judges. Multi-agent reasoning-tree auditing as a corrective procedure is from Sun et al. [116]. Replicability across multiple studies, which is increasingly relevant for AI-evaluation reproducibility, is treated in Bogomolov and Heller [22] and Heller and Bogomolov [58].

12. Exercises

Basic.

Exercise 13.6 (A genomic discovery list). An expression study tests $m = 18{,}000$ genes and reports $R = 240$ discoveries using BH at $q = 0.05$. State the FDR guarantee in words. Does the

guarantee imply that every reported gene has probability at most 0.05 of being null? Explain the difference between FDR and local FDR in this example.

EXERCISE 13.7 (Watermark null and alternative). Write a null and an alternative hypothesis for a watermark detector whose statistic $\phi(\xi, \widetilde{y})$ is larger when the text agrees more strongly with the secret key. Then write the key-resampling p-value and state which inequality direction should be used.

EXERCISE 13.8 (Exponential minimum sampling). Prove Proposition 13.2 directly for $K = 2$ by integrating over the value of E_1. Then explain how the same calculation extends to general K.

EXERCISE 13.9 (Benchmark families). A team compares four language models on ten tasks. For each model-task pair, it tests whether the model improves over a fixed baseline. What is the natural family if the team wants to report all improved cells? What is the natural family if it pre-specified one task as primary and treats all others as exploratory?

Intermediate.

EXERCISE 13.10 (Marginal versus conditional genomics). Let X_j be a genetic variant correlated with X_k, and suppose X_k is causal for a phenotype Y while X_j is not. Explain why a marginal test of X_j may reject. State the conditional null targeted by model-X knockoffs.

EXERCISE 13.11 (Scanning correction). A watermark detector scans 500 possible substrings and computes a single-window p-value for each substring. Describe a Bonferroni correction. Then describe a permutation or key-resampling procedure that calibrates the maximum statistic directly. Which approach is likely to be less conservative, and why?

EXERCISE 13.12 (Contamination p-value orientation). In a contamination audit, $T(X)$ is the log-likelihood of the canonical benchmark ordering. Large values are suspicious. Write the permutation p-value using B random permutations. How would the formula change if the statistic were a loss for which smaller values were suspicious?

EXERCISE 13.13 (Conformal factuality score). Suppose $F_0(x)$ is the original answer, $F_1(x)$ removes unsupported numbers, $F_2(x)$ removes unsupported named entities, and $F_3(x)$ abstains from factual claims. For a calibration reference $y^\star$, define the strict-safe score in this four-level family. Why is it important to require all more conservative thresholds to be safe?

Advanced.

EXERCISE 13.14 (Multiple contamination audits). An organization audits one model against 1,000 benchmark datasets and obtains one permutation p-value per dataset. Propose an analysis plan if the goal is to identify a list of suspicious datasets while controlling FDR. What dependence concerns arise, and how might e-values or hierarchical testing help if audits are repeated over time?

EXERCISE 13.15 (Design a conformal factuality study). Design a calibration study for conformal factuality in a medical question answering system. Specify the population of prompts, the reference answers, the entailment rule, the back-off family, and the target value of α. State clearly what the resulting guarantee does and does not say.

EXERCISE 13.16 (Knockoff interpretation). In a genotype study with strong LD, a knockoff procedure selects one variant from a correlated block. Explain why this should not automatically be interpreted as identifying the unique causal variant. What additional analyses or study designs would strengthen the biological interpretation?

Exercise 13.17 (From a p-value to a claim). For each of the following, state the null hypothesis, the test statistic, the reference distribution, and the claim justified by rejection: a gene-level BH analysis, a watermark key-resampling test, a canonical-order contamination test, and a conformal factuality calibration procedure.

Bibliography

[1] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

[2] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052, 2021.

[3] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proceedings of the 12th International Conference on Learning Representations*. ICLR, 2024.

[4] Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011. doi: 10.1214/11-AOS910.

[5] Ander Artola Velasco, Ivi Chatzi, Renato Vieira, Stratis Tsirtsis, and Manuel Gomez Rodriguez. Tokenization multiplicity leads to arbitrary price variation in LLM-as-a-service. arXiv preprint arXiv:2506.06446, 2025. Preprint.

[6] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

[7] Rina Foygel Barber and Aaditya Ramdas. The p-filter: Multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B*, 79(4): 1247–1268, 2017. doi: 10.1111/rssb.12218.

[8] Rina Foygel Barber, Emmanuel J. Candès, and Richard J. Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431, 2020. doi: 10.1214/19-AOS1852.

[9] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. doi: 10.1214/20-AOS1965.

[10] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

[11] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

[12] Stephen Bates, Emmanuel Candès, Lucas Janson, and Wenshuo Wang. Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535):1413–1427, 2021. doi: 10.1080/01621459.2020.1729163.

[13] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244.

[14] Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008. doi: 10.1111/j.1541-0420.2007.00984.x.

[15] Yoav Benjamini and Yosef 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.

[16] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[17] Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469): 71–81, 2005. doi: 10.1198/016214504000001907.

[18] Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006. doi: 10.1093/biomet/93.3.491.

[19] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

[20] Robert H. Berk and Douglas H. Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47:47–59, 1979. doi: 10.1007/BF00533250.

[21] Gilles Blanchard and Etienne Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871, 2009.

[22] Marina Bogomolov and Ruth Heller. Assessing replicability of findings across two studies of multiple features. *Biometrika*, 105(3):505–516, 2018. doi: 10.1093/biomet/asy029.

[23] Marina Bogomolov, Christine B. Peterson, Yoav Benjamini, and Chiara Sabatti. Hypotheses on a tree: New error rates and testing strategies. *Biometrika*, 108(3):575–590, 2021. doi: 10.1093/biomet/asaa086.

[24] Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009. doi: 10.1002/sim.3495.

[25] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Berlin, 2011. doi: 10.1007/978-3-642-20192-9.

[26] T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017. doi: 10.1214/16-AOS1461.

[27] T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B*, 73(5):629–662, 2011. doi: 10.1111/j.1467-9868.2011.00778.x.

[28] T. Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011. doi: 10.1198/jasa.2011.tm10155.

[29] T. Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B*, 76(2):349–372, 2014. doi: 10.1111/rssb.12034.

[30] Emmanuel J. Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

[31] Hongyuan Cao, Jun Chen, and Xianyang Zhang. Optimal false discovery rate control for large-scale multiple testing with auxiliary information. *The Annals of Statistics*, 50(2): 807–857, 2022. doi: 10.1214/21-AOS2128.

[32] Yiqun T. Chen and Daniela M. Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

[33] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013. doi: 10.1214/13-AOS1161.

[34] D. A. Darling and P. Erdős. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Mathematical Journal*, 23(1):143–155, 1956. doi: 10.1215/S0012-7094-56-02313-4.

[35] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4): 533–558, 2015. doi: 10.1214/15-STS527.

[36] Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L. Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, 2024. doi: 10.1080/01621459.2024.2353948. Forthcoming.

[37] Alex Dmitrienko, Ajit C. Tamhane, and Brian L. Wiens. General multistage gatekeeping procedures. *Biometrical Journal*, 50(5):667–677, 2008. doi: 10.1002/bimj.200710464.

[38] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004. doi: 10.1214/009053604000000265.

[39] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128.

[40] Sandrine Dudoit and Mark J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, 2008.

[41] Cynthia Dwork, Weijie J. Su, and Li Zhang. Differentially private false discovery rate control. *Journal of Privacy and Confidentiality*, 11(2), 2021. doi: 10.29012/jpc.755.

[42] Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004. doi: 10.1198/016214504000000089.

[43] Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Number 1 in Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge, 2012.

[44] European Medicines Agency. Guideline on multiplicity issues in clinical trials. EMA/CHMP/44762/2017, 2017. URL https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials-scientific-guideline.

[45] Jaiden Fairoze, Jiwon Kim, Lichao Yang, Sanjam Garg, Mingyuan Ma, and Dawn Song. Verifiable and oblivious watermark detection for large language models. arXiv preprint arXiv:2509.27666, 2025. Preprint.

[46] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 4 edition, 1932.

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

[48] William Fithian, Dennis L. Sun, and Jonathan Taylor. Optimal inference after model selection, 2017.

[49] Dean P. Foster and Robert A. Stine. Alpha-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B*, 70(2):429–444, 2008. doi: 10.1111/j.1467-9868.2007.00643.x.

[50] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045.

[51] Lucy L. Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024. doi: 10.1080/01621459.2022.2116331.

[52] Etienne Gauthier, Francis Bach, and Michael I. Jordan. E-values expand the scope of conformal prediction. arXiv preprint arXiv:2503.13050, 2025. Preprint.

[53] Jelle J. Goeman, Rosa J. Meijer, Thijmen J. P. Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106 (4):841–856, 2019. doi: 10.1093/biomet/asz041.

[54] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

[55] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

[56] Peter Hall and Jiashun Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732, 2010. doi: 10.1214/09-AOS764.

[57] Ruijian Han, Lan Luo, Yiming Lin, and Jian Huang. Adaptive debiased Lasso in high-dimensional GLMs with streaming data. arXiv preprint arXiv:2405.18284, 2024. Preprint.

[58] Ruth Heller and Marina Bogomolov. Replicability across multiple studies. *Statistical Science*, 38(4):602–620, 2023. doi: 10.1214/23-STS890.

[59] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. doi: 10.1093/biomet/75.4.800.

[60] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[61] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 1988. doi: 10.1093/biomet/75.2.383.

[62] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2): 1055–1080, 2021. doi: 10.1214/20-AOS1991.

[63] Nikolaos Ignatiadis, Bernd Klaus, Judith B. Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580, 2016. doi: 10.1038/nmeth.3885.

[64] Yu. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[65] International Council for Harmonisation. ICH harmonised tripartite guideline E9: Statistical principles for clinical trials. ICH guideline, 1998. URL https://www.ich.org/page/efficacy-guidelines.

[66] International Council for Harmonisation. ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials. ICH guideline addendum, 2020. URL https://www.ich.org/page/efficacy-guidelines.

[67] Leah Jager and Jon A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053, 2007. doi: 10.1214/0009053607000000244.

[68] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[69] Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 46(2):526–554, 2018. doi: 10.1214/17-AOS1559.

[70] Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980. doi: 10.1016/0047-259X(80)90065-2.

[71] Eugene Katsevich and Chiara Sabatti. Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1–33, 2019. doi: 10.1214/18-AOAS1185.

[72] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 2023.

[73] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000. doi: 10.1214/aos/1015957397.

[74] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=FpaCL1MO2C.

[75] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

[76] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2 edition, 2000.

[77] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

[78] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[79] Lihua Lei and William Fithian. AdaPT: An interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B*, 80(4):649–679, 2018. doi: 10.1111/rssb.12274.

[80] Lihua Lei, Aaditya Ramdas, and William Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267, 2021. doi: 10.1093/biomet/asaa064.

[81] James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025. doi: 10.1080/01621459.2023.2270148.

[82] Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B*, 81(1):45–74, 2019. doi: 10.1111/rssb.12298.

[83] Guanxun Li and Xianyang Zhang. A general framework for multiple testing via e-value aggregation and data-dependent weighting. arXiv preprint arXiv:2312.02905, 2023.

[84] Guanxun Li and Xianyang Zhang. A note on e-values and multiple testing. *Biometrika*, 112(1):asae050, 2025. doi: 10.1093/biomet/asae050.

[85] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

[86] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(1):145–156, 2022. doi:

10.1093/biomet/asab017.

[87] Yaowu Liu and Jun Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[88] Joshua R. Loftus and Jonathan E. Taylor. Selective inference in regression models with groups of variables. arXiv preprint arXiv:1511.01478, 2015. Preprint; foundational for CV-selected models.

[89] Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. doi: 10.1093/biomet/63.3.655.

[90] Christoforos Mavrogiannis, Angelos Mavrogiannis, and Constantine Dovrolis. How independent are large language models? A statistical framework for auditing behavioral entanglement and reweighting verifier ensembles. arXiv preprint arXiv:2410.07650, 2024. Preprint.

[91] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. doi: 10.1214/009053606000000281.

[92] Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36029–36047. PMLR, 2024.

[93] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, 2023.

[94] Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

[95] Yuki Nishino, Tomohiro Shiraishi, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for saliency maps of graph neural networks via selective inference. arXiv preprint arXiv:2501.18452, 2025. Preprint.

[96] Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models, 2023.

[97] Mehrdad Pournaderi and Yu Xiang. Differentially private model-X knockoffs via johnson–lindenstrauss transform. arXiv preprint arXiv:2508.04800, 2025. Preprint.

[98] Aaditya Ramdas, Tijana Zrnic, Martin J. Wainwright, and Michael I. Jordan. SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4286–4294. PMLR, 2018.

[99] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

[100] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[101] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020. doi: 10.1080/01621459.2019.1660174.

[102] Tom Sander, Pierre Fernandez, Saeed Mahloujifar, Alain Durmus, and Chuan Guo. Detecting benchmark contamination through watermarking, 2025.

[103] Sanat 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.

[104] Matteo Sesia, Chiara Sabatti, and Emmanuel J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 2019. doi: 10.1093/biomet/asy033.

[105] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647.

[106] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[107] Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Selective inference for change point detection by recurrent neural network. *Neural Computation*, 37 (6):1213–1252, 2025. doi: 10.1162/neco_a_01756.

[108] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi: 10.1080/01621459.1967.10482935.

[109] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986. doi: 10.1093/biomet/73.3.751.

[110] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The NHGRI–EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1): D977–D985, 2023. doi: 10.1093/nar/gkac1010.

[111] Branko Sorić. Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):608–610, 1989. doi: 10.1080/01621459.1989.10478811.

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

[113] John D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. doi: 10.1214/aos/1074290335.

[114] John D. Storey, Jonathan E. Taylor, and David 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.

[115] Samuel A. Stouffer, Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams. *The American Soldier: Adjustment During Army Life*, volume 1 of *Studies in Social Psychology in World War II*. Princeton University Press, Princeton, NJ, 1949.

[116] Hanqi Sun, Xiaoyu Wu, Jiansheng Maddu, Yuping Yu, Wenliang Wang, and Sarath Ramnath. Auditing multi-agent LLM reasoning trees outperforms majority vote and LLM-as-judge. arXiv preprint arXiv:2410.09341, 2024. Preprint.

[117] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. doi: 10.1093/biomet/ass043.

[118] Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

[119] Wenguang Sun, Brian J. Reich, T. Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale multiple testing. *Journal of the Royal Statistical Society: Series B*, 77(1):59–83, 2015.

[120] Jonathan Taylor and Robert Tibshirani. Post-selection inference for l1-penalized likelihood models. *The Canadian Journal of Statistics*, 46(1):41–61, 2018. doi: 10.1002/cjs.11313.

[121] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[122] Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. doi: 10.1080/01621459.2015.1108848.

[123] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2530–2540, 2019.

[124] U.S. Food and Drug Administration. Multiple endpoints in clinical trials: Guidance for industry. FDA guidance document, 2022. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry.

[125] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014. doi: 10.1214/14-AOS1221.

[126] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[127] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

[128] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[129] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022. doi: 10.1111/rssb.12489.

[130] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

[131] Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3997–4005. PMLR, 2024.

[132] Daniel Yekutieli. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008. doi: 10.1198/016214507000001373.

[133] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014.

[134] Xianyang Zhang. Testing High Dimensional Mean Under Sparsity, 2015. URL https://arxiv.org/abs/1509.08444. arXiv:1509.08444v2.

[135] Xianyang Zhang and Jun Chen. Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *Journal of the American Statistical Association*, 117(537):411–427, 2022. doi: 10.1080/01621459.2020.1783273.

[136] Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017. doi: 10.1080/01621459.2016.1166114.

[137] Lasse Fischer and Aaditya Ramdas. Admissible online closed testing must employ e-values, 2024. URL https://arxiv.org/abs/2407.15733.

[138] Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011. doi: 10.1214/11-STS356.

[139] Jelle J. Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2): 1218–1238, 2021. doi: 10.1214/20-AOS1999.

[140] Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas, and Jelle Goeman. Bringing closure to false discovery rate control: A general principle for multiple testing, 2026. URL https://arxiv.org/abs/2509.02517.