THE UNIVERSITY OF CHICAGO

TESTING ON THE EDGE: DETECTION AND LOCALIZATION OF SPARSE SIGNALS IN SEQUENCE DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
DANIEL XIANG

CHICAGO, ILLINOIS
AUGUST 2023

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Professors Peter McCullagh and Chao Gao, whose commitment to the statistical profession and passion for research and teaching has been an ongoing source of inspiration to me. Thank you for your patient guidance, for teaching me how to articulate my understanding and explore its boundaries, and for providing me with the intellectual freedom to do so at my own pace.

I would also like to thank Prof. Will Fithian for orienting me in fruitful directions and for nudging me towards making visually intuitive presentations. Thanks to my wonderful collaborator Jake Soloff for his clear-eyed advice and for pushing me to gain fresh insight into familiar ideas. Thanks to Prof. Rina Barber for her balanced perspective and attention to detail in reading this thesis. I often find myself impressed and inspired by her ability to create academic environments in which everyone's needs are met. Also thanks to my wonderful collaborator and academic sibling Micól Tresoldi for stimulating conversations about research and life. Thank you Adela DePavia for your support and encouragement throughout but especially towards the end of my graduate studies. A special shoutout to Abby Stevens, Quinn Underriner, and Kota Mizutani for looking out for me during challenging times. Many thanks also to the remaining faculty, students and staff within the statistics department for the fantastic environment which was both intellectual and friendly, and for technical support, as well as to students in the CAM department for organizing engaging student seminars. This includes Professors Mei Wang, Fei Liu, Yibi Huang, Yali Amit, Steve Lalley, Dan Nicolae, and Linda Collins, as well as Mitzi Nakatsuka, Keisha Prowoznik, Jonathan Rodriguez, Yandi Shen, Yier Lin, Chih-Hsuan Wu, Dongyue Xie, Wanrong Zhu, Pinhan Chen, Ahmed Bou-Rabee, Keshav Vemuri, Nathan Gill, Jasha Sommer-Simpson, Hai Tran Bach, Subhodh Kotekal, Melissa Adrian, Irina Cristali, Qing Yan, Zhan Lin, Lizhen Nie, Jiacheng Wang, Haoyang Liu, Misha Sohan, Sean O'Hagan, Cheuk To Tsui, Chris McKennan, Zhimei Ren, Solomon Quinn, Kimberly Liu, Abby Poteshman, Mark Olsen, Nathan Waniorek, Angela Wang, Phillip Lo, Peter Nekrasov, Jialiang Zhou, and Sue Parkinson. Thanks also to people outside of UChicago for engaging discussions and moral support including Professors Asaf Weinstein, Dani Yekutieli, Ruth Heller, Lihua Lei and Richard Samworth, as well as Tijana Zrnic, Drew Nguyen, Yixiang Luo, Asher Spector, Tyler Devlin, Marley Kirton, Daniel Kunin, Michelle Hoang, Zack Fey, Eric Cheng, Erik Lee, Antoine Malfroy-Camine, Mike Packard, Deni Lushi, Kirby Stoff, and Adam Zakaria.

Finally, my deepest gratitude is felt towards my parents, Hong Xiang and Li Zhu, and my sister Alice, for loving and supporting me unconditionally throughout my life. Thank you for everything.

ABSTRACT

Exchangeability and sparsity are fundamental concepts in statistical modeling. The former requires that the model or procedure satisfy a certain symmetry, either probabilistic or operational in nature. On the other hand, sparsity implies that underlying features are mostly null with a small fraction of exceptions, and can be difficult to reconcile with exchangeability in complex data structures.

Chapter 2 discusses joint work with Professor Peter McCullagh on modeling exchangeable random graphs. Compatible notions of exchangeability and sparsity are motivated by the Ewens process (see, e.g. Crane (2016), Tavaré (2021)), which is exchangeable and sparse when viewed as a sequence of distributions on a random permutation. The Permanental Graph Model (PGM) is proposed as a generalization of the random permutation to a directed random graph, which is characterized mathematically by its normalization constant and degree distribution. A negative result is found, implying that no setting of parameters in the PGM yields an exchangeable random graph process.

Chapters 3 and 4 discuss joint work with Professors Chao Gao and Peter McCullagh on sparse signal detection. Procedures are developed to detect sparse alternatives to a global null for matrix and sequence data with independent Gaussian errors. For a signal that is sparse in the sense of McCullagh and Polson (2018), an identification boundary determines the number of independent samples required in order for the signal to be identifiable in a sparse limiting sense. There is a close relationship with the detection-boundary literature, which studies the problem of discriminating between two distributions $F_0^{\otimes n}$ and $F_1^{\otimes n}$ in the large-sample limit.

Chapter 5 discusses joint work with Jake Soloff and Professor Will Fithian on boundary false discovery rate (bFDR) controlling methodology. Some theory for local false discovery rates is developed, providing a frequentist interpretation of Bayes motivated procedures operating in a model where effect sizes of individual studies are fixed and unknown. The local false discovery rate (lfdr) is traditionally defined as a posterior probability when effects are random, but can be more generally interpreted as the expected proportion of nulls among hypotheses with similar test-statistics. This concept inspires a new frequentist type 1 error criterion, the bFDR, which describes the rate of false discoveries near the rejection threshold and is a local analogue to the FDR concept of Benjamini and Hochberg (1995). We discuss bFDR control under relaxed null assumptions, and demonstrate our main ideas on a dataset of "nudges" from the behavioral psychology literature.

CHAPTER 1
INTRODUCTION

Modern datasets in scientific fields such as genetics, behavioral psychology, and astronomy, to name a few, often contain information regarding an aggregate of comparable problem instances. Instead of testing a single hypothesis, the scientist may be asking thousands of parallel questions at a time, where true signals in the data are sparse. This kind of repeated structure is a setting where frequentists and Bayesians can find common ground, agreeing for example, on how to answer each of the many questions in a controlled way.

Multiple hypothesis testing. For testing hypotheses $H_1, \dots, H_m \in \{0, 1\}$ based on p-values $p_1, \dots, p_m \in [0, 1]$, the FDR criterion of Benjamini and Hochberg (1995) has been widely adopted in large scale testing problems as the notion of type 1 error in an exploratory phase of testing, where the goal is to to reject as many hypotheses as possible while controlling the average quality of these rejections. It is often said that controlling FDR at say, 10%, expresses a willingness to take on 1 false discovery for every 9 true discoveries. However, this viewpoint applies equally well to the subset of rejections near the decision threshold, whose rate of false discoveries could be much higher than the FDR among all rejections. From a Bayesian point of view, the local false discovery rate (lfdr, Efron et al. (2001)) relates more directly to our willingness to trade off between making false positives and missing potential discoveries (Sun and Cai; 2007).

Rather than maximizing the number of rejections subject to a constraint on the FDR, a Bayesian would condition on the entire observation, rejecting hypothesis $H_i = 0$ when the posterior probability of $H_i = 0$ is small, ensuring that each rejection is of high quality based on what was seen. In contrast, a procedure controlling FDR may include arbitrarily low quality rejections, as long as the average quality of the rejections is maintained. Despite being intuitive to grasp, the FDR has led to the (perhaps passive) acceptance of wider implications of the BH idea, namely that by including a few extra strong bets in our rejection set, we become more willing to reject "free-riding" hypotheses for which the evidence is weak.

In large scale settings, one can justify individual rejections by comparison with hypotheses for which a similar level of evidence is observed. In a Bayesian model, the evidence about a hypothesis is summarized by the local fdr (Efron et al.; 2001),

$$\mathrm{lfdr}(t) = \mathbb{P}(H_i = 0 \mid p_i = t).$$

A frequentist might be skeptical of this quantity, since if each H_i is non-random, the probability definition above is vacuous. However, a more general interpretation of lfdr (Chapter 5) makes sense in the underlying frequentist model, instead of describing our epistemic uncertainty in an imagined Bayesian model. Proposition 5.2.1 in Chapter 5 states that under regular conditions,

$$\mathrm{FDP}(\{H_i : |\mathrm{lfdr}(p_i) - \alpha| \le \varepsilon_m\}) \xrightarrow{\mathbb{P}} \alpha,$$

as $\varepsilon_m \to 0$, for $\alpha \in \mathrm{range}(\mathrm{lfdr})$, where $\mathrm{lfdr}(t) := \bar{\pi}_0 / \bar{f}(t)$ is defined in the frequentist model in terms of the null proportion $\bar{\pi}_0 := \frac{|\mathcal{H}_0|}{m}$ and average density $\bar{f}$ of the p-values. In other words, the realized proportion of null hypotheses $H_i = 0$ for which $p_i \approx \text{lfdr}^{-1}(\alpha)$ tends to α as $m \to \infty$. Informally, letting $t = \text{lfdr}^{-1}(\alpha)$

$$\alpha \approx \frac{\#\{i : H_i = 0, \text{lfdr}(p_i) \approx \alpha\}}{\#\{i : \text{lfdr}(p_i) \approx \alpha\}} \iff \text{lfdr}(t) \approx \frac{\#\{i : H_i = 0, p_i \approx t\}}{\#\{i : p_i \approx t\}},$$

under mild assumptions, such as independent test statistics with continuous densities. The above interpretation of lfdr makes no reference to a Bayes two-groups model, and is purely an asymptotic statement about the null proportion among hypotheses with comparable test statistics. A main claim in this thesis is that the lfdr concept is useful beyond Bayesian multiple testing, and can be used to remedy logical inconsistencies in the FDR criterion, such as the "free-rider" problem described above.

For a rejection set $\mathcal{R}$ making R rejections, its boundary FDR is defined as the marginal rate at which the last rejection is null,

$$\text{bFDR}(\mathcal{R}) := \mathbb{P}(H_{(R)} = 0),$$

where $H_{(0)} := 1$ indicates no rejections, and $H_{(k)}$ denotes the hypothesis associated with the order statistic $p_{(k)}$. This quantity is again puzzling to a frequentist at first glance, since if each of $H_1, \ldots, H_m \in \{0, 1\}$ is fixed and non-random, then how can we speak of the probability that one of them is null? To resolve this confusion, note that the boundary FDR is a property of the procedure $\mathcal{R}$, whose last rejected hypothesis $H_{(R)} \in \{0, 1\}$ is a non-degenerate random variable. By contrast, the FDR of the procedure $\mathcal{R}$ is the probability that a uniformly selected rejection is null,

$$\text{FDR}(\mathcal{R}) = \mathbb{P}(H_{(I)} = 0) \quad I \sim \text{Uniform}\{1, \ldots, R\}. \tag{1.1}$$

Viewed as a corrective to the FDR, the bFDR remains a more liberal notion of type 1 error than the family wise error rate, and it is more robust to the free-rider problem.

In a Bayesian setting, Soloff et al. (2022) proposed a method, called the Support Line (SL) procedure, which rejects the hypotheses corresponding to the R_α smallest p-values, where

$$R_\alpha := \text{argmax}_{k=0,\ldots,m} \left\{ \frac{\alpha k}{m} - p_{(k)} \right\}, \quad p_{(0)} := 0. \tag{1.2}$$

This procedure can be understood in relation to the BH procedure, as illustrated in Figure 1.1. Instead of finding the last time for which the process $\frac{\alpha k}{m} - p_{(k)}$ is positive, SL finds the maximizer $p_{(R_\alpha)}$, and rejects all hypotheses with p-values at or below $p_{(R_\alpha)}$. Whereas the BH procedure (Benjamini and Hochberg; 1995) controls the quality of rejections on average in the sense of (1.1), the SL procedure controls the quality of the last rejection (Soloff et al.; 2022); if p-values are independent and uniform under the null, then

$$\text{bFDR}(\mathcal{R}_\alpha) = \bar{\pi}_0 \alpha,$$

where $\mathcal{R}_\alpha \coloneqq \{i : p_i \leq p_{(R_\alpha)}\}$ is defined by (1.2).

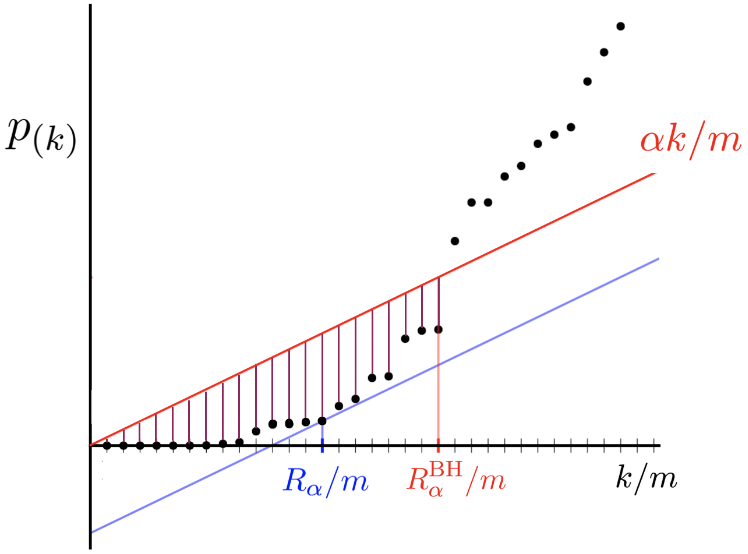


Figure 1.1: The order statistics $p_{(k)}$ of the p-values are plotted on the vertical axis as a function of the index k/m, shown by the black dots. The BH procedure, in red, finds the largest index R_α^{BH} such that $p_{(R_\alpha^{\mathrm{BH}})}$ falls below the ray of slope α; by contrast, SL computes where the length of the purple vertical bars is maximized, which is the boundary point $(R_\alpha/m, p_{(R_\alpha)})$ of the supporting line of slope α.

The strict uniform assumption for the null distribution can be relaxed, while preserving bFDR control in finite samples or in an asymptotic sense. This point is discussed further in Sections 5.3 and 5.3.3. Connections between the boundary FDR and the lfdr are discussed in Section 5.4.

The boundary FDR and lfdr are conceptually simple to grasp and are helpful for directing attention to a hypothesis test at the boundary, namely the rejection for which the evidence is weakest. In what comes next, we discuss boundary phenomena of a different nature having to do with the more basic problem of testing the global null. The main assumption in Chapters 3 and 4 is that the signals are sparse when the alternative hypothesis is true.

Sparse signal detection. Time series data arise in a variety of contexts where the question of interest is to determine whether and where there is a moment in time, called a changepoint, before which the data behave differently than after. In Chapter 3, we study a hypothesis testing problem in a stylized setting where our observation takes the form,

$$X = \theta + E \in \mathbb{R}^{p \times n},$$

where each $E_{ij} \overset{\text{iid}}{\sim} N(0, 1)$, and the mean matrix θ may contain a shift in mean at some column index. The global null scenario is that each row of θ is a constant vector, whereas the alternative scenario allows for $s \ll p$ rows of θ to be non-constant, sharing a common column index $t^* \in [n-1]$ at which there is a change in mean parametrized by a signal to noise ratio $\rho > 0$. The index t^* is called the changepoint location.

$$\Theta_0 := \big\{\theta \in \mathbb{R}^{p\times n} : \text{Every row in } \theta \text{ is a constant multiple of the all ones vector.}\big\}$$
$$\Theta_1(\rho, s) := \big\{\theta \in \mathbb{R}^{p\times n} : \text{At most } s \text{ rows share a common changepoint with signal size } \rho.\big\}$$

In the detection boundary literature (see e.g. Donoho and Jin (2004), Cai and Wu (2014)), the information theoretic limits on distinguishing the global null from a sparse alternative are characterized by a 'critical signal size'. In our setting, this is the smallest signal size ρ needed to distinguish an element $\theta \in \Theta_0$ from $\theta \in \Theta_1$ based on the data X. The following asymptotic regime relates the sparsity s to the dimensions of the matrix of observations,

$$\log\log\log n \sim a \log p, \quad s \sim p^{1-\beta}, \quad a > 0, \quad \beta \in (0,1).$$

The critical signal size is given by a formula called the detection boundary,

$$\rho^*(a,\beta)^2 := \begin{cases} p^{a-(1-2\beta)} & a \le 1-2\beta \\ (a-(1-2\beta))\log p & 1-2\beta < a \le 1-4\beta/3 \\ 2(\sqrt{1-a}-\sqrt{1-a-\beta})^2 \log p & 1-4\beta/3 < a \le 1-\beta \\ p^{a-(1-\beta)} & a > 1-\beta, \end{cases}$$

which determines when it is possible to distinguish between the global null and a sparse alternative with signal size ρ. Namely, ρ^* provides a sharp characterization of the global null testing problem in terms of asymptotic negligibility of the worst case sum of type 1 and 2 errors. A precise statement is given in terms of the minimax testing risk,

$$\mathcal{R}(\rho, s) := \inf_{\psi:\mathbb{R}^{p\times n}\to\{0,1\}} \left[\sup_{\theta\in\Theta_0} \mathbb{E}_\theta \psi(X) + \sup_{\theta\in\Theta_1(\rho,s)} \mathbb{E}_\theta(1-\psi(X)) \right].$$

Letting $\rho := \rho^*(a, \beta_1)$ for $a > 0$ and $\beta_1 \in (0,1)$,

1. (Lower Bound) if $1 > \beta > \beta_1$, then $\mathcal{R}(\rho, s) \not\to 0$, as $p \to \infty$,

2. (Upper Bound) if $0 < \beta < \beta_1$, then $\mathcal{R}(\rho, s) \to 0$, as $p \to \infty$.

Variants of this problem are discussed in Sections 3.7 and 3.2.2, where the changepoint location may be different between non-null rows in θ, or where the mean shift may be either positive or negative at the changepoint location.

In Chapter 4, we consider detectability of a signal distribution that is sparse in the sense of McCullagh and Polson (2018). The observations $Y_1, \ldots, Y_n \in \mathbb{R}$ are drawn independently from a convolutional model,

$$Y_i = X_i + \varepsilon_i \quad i = 1, \ldots, n$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, and $X_i \overset{\text{iid}}{\sim} P_\rho$, a distribution whose tails behave inverse-polynomially,

$$\begin{aligned} P_\rho(\mathrm{d}x) &\asymp |x|^{-(d+1)}\mathrm{d}x, \qquad \text{as } |x| \to \infty, \quad \rho > 0 \text{ fixed} \\ P_\rho(|X_i| > x) &\asymp \rho|x|^{-d} + o(\rho) \qquad \text{as } \rho \to 0, \end{aligned}$$

for an inverse-power index $d \in (0,2)$. A consequence of the sparsity assumption is that the marginal distribution of each Y_i depends only on ρ and the limiting exceedance measure $H_d(\mathrm{d}x)$ (defined in Section 4.2),

$$Y_i \overset{\text{iid}}{\sim} (1-\rho)\phi + \rho\psi_d, \quad i = 1, \dots, n, \tag{1.3}$$

where ϕ is the standard normal density and ψ_d depends only on the limiting tail behavior of P_ρ through H_d as $\rho \to 0$. Here we are abusing notation, since each Y_i is only approximately distributed according to the above two-groups model. However, the discrepancy is negligible to first order in the sparsity rate ρ, as discussed in Section 4.2.3 for a typical family of atom-and-slab distributions considered in the detection boundary literature.

This observation about the marginal density of Y_i essentially reduces the signal detection problem to estimation of the parameter ρ from $Y_1, \dots, Y_n$, assuming H_d is already specified[1]. To this point, the natural estimator is the maximum likelihood estimate for ρ based on the marginal approximation (1.3),

$$\hat{\rho}_n \coloneqq \operatorname{argmax}_{\rho \in [0,1]} \prod_{i=1}^{n} \left((1-\rho)\phi(Y_i) + \rho\psi_d(Y_i)\right).$$

If $\hat{\rho}_n = 0$, then either the global null is true, or there aren't enough independent samples n to distinguish (1.3) from pure noise. An identification boundary $n = b(\rho)$, defined in Section 4.2, is a function of ρ that determines the number of independent samples required in order for the event $\{\hat{\rho}_n > 0\}$ to have non-trivial probability in the sparse limit as $\rho \to 0$. There is a close relationship with the detection boundary literature, regarding detectability of a sparse signal distribution with inverse-power tails, which we discuss in Section 4.2.3.

Procedures considered in Chapters 4 and 5 are exchangeable in the sense that they do not crucially depend on the order in which observations arise. For example, the SL procedure calculates a rejection threshold $\hat{\tau}_\alpha \coloneqq p_{(R_\alpha)}$ based on the order statistics. Similarly, $\hat{\rho}_n$ depends on the sequence $(Y_i)_{i=1}^n$ in a way that is invariant to permuting the data. This type of invariance is a natural restriction to place on a procedure when the statistical task at hand doesn't depend crucially on the order in which test statistics are observed. Section 5.4 makes connections to the optimal procedures for permutation-invariant decision problems (Weinstein; 2021).

In simply structured data such as an independent sequence of observations, exchangeability of the observations $Y_1, \dots, Y_n$ and sparsity of underlying signals $X_1, \dots, X_n$ are compat-

[1] One may also consider jointly estimating the power-index d and sparsity rate ρ by maximum likelihood, as demonstrated in Section 4.3.2 on a genomic dataset.

ible distributional assumptions, in the sense that they do not lead to obvious inconsistencies in the model specification. This is less true in more complex settings such as with network data, which we discuss in Chapter 2.

Sparse exchangeable graphs. Exchangeability plays a key role in statistical modeling, and is relatively straightfoward to define for an infinite sequence of real-valued random variables. The two components for infinite exchangeability of a sequence of distributions (P_n) are (i) consistency, and (ii) finite exchangeability for each n. Specifying both of these properties for graph-valued data while allowing for sparsity in the number of edges in the graph turns out to be a non-trivial task. A consequence of the Aldous-Hoover theorem (see e.g. Chapter 7.5 in Kallenberg (2005)) is that any node-exchangeable, subselection-consistent sequence of distributions on graphs yields an expected number of edges that grows quadratically in the number of nodes. When infinite node-exchangeability is defined with respect to the subselection notion of consistency, the resulting graphs will not exhibit the sparse behavior observed in many modern network datasets.

In Chapter 2, we consider another notion of consistency, namely, delete-and-repair consistency, and it is motivated by the sense in which the infinitely exchangeable permutations defined by the Chinese restaurant process (CRP, Aldous (1985)) are consistent. Delete-and-repair consistency can be used to obtain a nontrivial sequence of distributions on graphs (P_n) that is sparse and exchangeable, a well known example being the Ewens permutations (see e.g. Crane (2016)). This notion of consistency requires that to go from a graph on vertex set $[n]$ to the subgraph on vertex set $[n-1]$, the node n is deleted, and edges directing into and out of the removed vertex are repaired, i.e. connected. In contrast, subselection consistency requires that all edges going into and out of the deleted vertex, also are deleted.

A generalization of the CRP(α) can be defined using the α-weighted permanent,

$$\mathrm{per}_\alpha(G) \coloneqq \sum_{\sigma \in \mathcal{S}_n} \alpha^{\#\sigma} \prod_{i=1}^{n} G_{i,\sigma(i)}, \quad G \in \mathbb{R}^{n \times n}$$

where the sum runs over all permutations $\sigma : [n] \to [n]$, and $\#\sigma$ is the number of cycles. If G has Boolean entries, the product $\prod G_{i,\sigma(i)}$ is equal to one if σ is contained as a sub-graph in G, and zero otherwise. Consider the graph distribution supported on $\mathcal{G}_n \subseteq \{0,1\}^{n\times n}$ a subset of all adjacency matrices,

$$P_n(G) \propto 1_{\{G\in\mathcal{G}_n\}} \cdot \mathrm{per}_\alpha(G) = 1_{\{G\in\mathcal{G}_n\}} \cdot \sum_{\sigma \in \mathcal{S}_n} \alpha^{\#\sigma} 1_{\{\sigma \subseteq G\}}. \tag{1.4}$$

The sequence (P_n) coincides with the CRP(α) on permutations when $\mathcal{G}_n$ is taken to be the set of permutation matrices. Letting $\mathcal{G}_n = \{0,1\}^{n\times n}$ have unrestricted support, and parametrizing the edge density by a constant $\beta > 0$, we obtain the Permanental Graph Model (PGM),

$$P_n(G) \propto \beta^{\#G} \mathrm{per}_\alpha(G), \quad G \in \{0,1\}^{n\times n}$$

where $\alpha, \beta > 0$ and $\#G = \sum_{i,j \leq n} G_{ij}$ is the number of directed edges in G. This distribution is finitely exchangeable for each n, and detailed information, such as the normalization constant and the degree distribution, can be calculated exactly (see Theorem 2.2.1). However, a negative result is shown, stating that any sequence of pairs $(\alpha_n, \beta_n)_{n=1}^{\infty}$ defines a sequence (P_n) satisfying neither subselection nor delete-and-repair consistency. Thus an exchangeable graph process fully supported on $\{0,1\}^{n \times n}$ cannot be obtained using the finite dimensional distributions given by the PGM.

CHAPTER 2
PERMANENTAL GRAPHS

2.1 Introduction

Two components for infinite exchangeability of a sequence of distributions (P_n) are (i) consistency, and (ii) finite exchangeability for each n. From a modeling perspective, exchangeability is an assumption that is natural in a setting where the statistical units are labelled in an arbitrary manner. If the process being studied is a record of the relationships $(i, j) \mapsto X_{ij}$ between ordered pairs of units (i, j), the process X^σ after label permutation has components $(i, j) \mapsto X_{\sigma(i),\sigma(j)}$. As a matrix, $X^\sigma = \sigma X \sigma^{-1}$ is obtained from X by permuting rows and columns, i.e., by conjuation by $\sigma \in \mathcal{S}_n$. In this setting, finite exchangeability means that all of the permuted matrices have the same joint distribution.

Subselection consistency is not specific to Boolean matrices, but applies to real-valued matrices and to more general arrays. It requires that for $X \in \{0, 1\}^{n\times n}$ distributed according to P_n, the top left $(n - 1) \times (n - 1)$ submatrix of X is distributed according to P_{n-1}. The operation defined by deleting the last row and column of the adjacency matrix does not rely on the fact that the entries of an adjacency matrix are boolean valued. Sampling a sub-network according to subselection amounts to picking a subset of vertices and including only the edges between pairs of vertices in the selected subset. The following calculation, replicated from Crane (2018), demonstrates how edge sparsity, node-exchangeability, and subselection-sampling are at odds with each other.

Calculation. $X \in \{0, 1\}^{n\times n}$ is a simple directed graph possibly containing self-loops. X is assumed to be “sparse”, i.e.

$$\sum_{i,j} X_{ij} = \varepsilon n = o(n^2),$$

for some $\varepsilon > 0$ independent of n. Let $\sigma \in \mathcal{S}_n$ be drawn uniformly at random, and put $Y := \sigma X \sigma^{-1}$. In this model, we observe the top left $m \times m$ sub-matrix of Y, denoted $Y_{1:m,1:m}$, which is exchangeable according to this construction. Assume $m \ll n$, at the order $m = o(\sqrt{n})$. By the union bound,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i,j\le m} \{Y_{ij} = 1\} \right) &\le \sum_{i,j\le m} \mathbb{P}(Y_{ij} = 1) \\ &\asymp m^2 \mathbb{P}(Y_{12} = 1). && \text{(exchangeability)} \end{aligned}$$

The event $Y_{12} = 1$ corresponds to having picked two vertices uniformly at random from the $\binom{n}{2} \asymp n^2$ possible pairs, and observing an edge between them. Hence

$$\mathbb{P}(Y_{12} = 1) \asymp \frac{\varepsilon n}{n^2} = \frac{\varepsilon}{n}.$$

Plugging this back into the union bound, we find that

$$\mathbb{P}\left(\bigcup_{i,j\leq m}\{Y_{ij}=1\}\right) \leq \frac{\varepsilon m^2}{n} \approx 0. \qquad (m = o(\sqrt{n}))$$

The network we “observe” contains no edges with high probability. Put more plainly, we observe no network at all!

□

In order to resolve the contradiction suggested by the calculation above, at least one of sparsity, node-exchangeability, or subselection-consistency must be modified. In this note, an alternative notion of consistency is considered, namely, delete-and-repair consistency. It is motivated by the sense in which infinitely exchangeable permutations defined via the Chinese restaurant process (CRP) are consistent.

The α-weighted permanent is used to generalize the probability function associated to the partitions and permutations generated from the one parameter CRP(α) to probabilities on general directed graphs. Prescribing a probability to a directed graph according to the α permanent of its adjacency matrix automatically yields an exchangeable distribution with tractable calculations for the normalization constant and degree distribution similar to that of the Erdos-Renyi(n,p) model. The negative result we obtain is that any setting of the parameters for the permanental graphs allows for neither delete-and-repair nor subselection consistency. All proofs are deferred to Section 2.4.

2.2 Main Results

The α-weighted matrix permanent $\mathrm{per}_\alpha : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a matrix functional defined by,

$$\mathrm{per}_\alpha(G) \coloneqq \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma} \prod_{i=1}^{n} G_{i,\sigma(i)},$$

where the sum runs over all permutations $\sigma : [n] \to [n]$, and $\#\sigma$ is the number of cycles. If G is Boolean, the product $\prod G_{i,\sigma(i)}$ is equal to one if σ is contained as a sub-graph in G, and zero otherwise. Thus $\mathrm{per}_1(G)$ is the number of permutations contained as sub-graphs in G, and $\mathrm{per}_\alpha(G)$ is the cycle-weighted count. The matrix permanent is recovered by setting $\alpha = 1$, whereas the determinant is obtained as $\mathrm{per}_{-1}(G) = (-1)^n \det(G)$. In this note, the word graph or n-graph means a simple directed graph, with no multiple edges, but possibly containing self-loops. In other words, each n-graph is a Boolean matrix of order n.

For each $n \geq 1$, let $\mathcal{G}_n \subset \{0,1\}^{n\times n}$ be any subset of n-graphs having the following properties:

1. $\mathcal{G}_n$ is closed under conjugation:

$$\sigma\mathcal{G}_n\sigma^{-1} = \mathcal{G}_n \text{ for } \sigma \in \mathcal{S}_n$$

2. There exists $G \in \mathcal{G}_n$ and $\sigma \in \mathcal{S}_n$ such that $\sigma \subset G$.

Examples that we have in mind include the whole space, $\mathcal{G}_n = \{0,1\}^{n\times n}$, the permutations $\mathcal{G}_n = \mathcal{S}_n$, permutations having no fixed points for $n \geq 2$, single-cycle permutations, equivalence relations or set partitions as graphs, and so on. Condition 1 means that $\mathcal{G}_n$ is a union of group orbits, while condition 2 excludes trivialities such as graphs having fewer edges than vertices.

Consider the graph distribution

$$P_n(G) \propto \mathbf{1}\{G \in \mathcal{G}_n\}\mathrm{per}_\alpha(G) = \mathbf{1}\{G \in \mathcal{G}_n\} \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma}\mathbf{1}\{\sigma \subset G\}, \tag{2.1}$$

which is proportional to the α-permanent restricted to $\mathcal{G}_n$. Assumption 2 guarantees that

$$\sum_{G\in\mathcal{G}_n} \mathrm{per}_\alpha(G) > 0 \text{ for all } \alpha > 0,$$

so the normalizing constant is strictly positive. Note that P_n is automatically exchangeable, because for any $\tau \in \mathcal{S}_n$,

$$\begin{aligned}
P_n(G^\tau) &\propto \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma} \prod_{i=1}^n G_{\tau(i),\tau(\sigma(i))} \\
&= \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma} \prod_{j=1}^n G_{j,\tau\sigma\tau^{-1}(j)} && (i = \tau^{-1}(j) \text{ for some } j) \\
&= \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\tau\sigma\tau^{-1}} \prod_{j=1}^n G_{j,\tau\sigma\tau^{-1}(j)} && (\#\sigma = \#\tau\sigma\tau^{-1}) \\
&= \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma} \prod_{i=1}^n G_{i,\sigma(i)}. && (\text{sum ranges over all } \sigma \in \mathcal{S}_n)
\end{aligned}$$

Consistency in any sense is not immediately clear. When $\mathcal{G}_n$ is taken to be the set of adjacency matrices corresponding to partitions ($G_{ij} = 1 \iff i \sim j$), we have

$$P_n(\pi) \propto \sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma}\mathbf{1}\{\sigma \subset \pi\},$$

where $\sigma \subset \pi$ means the graph induced by σ is a subgraph of the graph induced by π, i.e. the cycles of the permutation σ coincide with the blocks of the partition π. Further simplification gives

$$P_n(\pi) \propto \alpha^{\#\pi} \cdot \#\{\sigma \in \mathcal{S}_n : \sigma \subset \pi\} = \alpha^{\#\pi} \prod_{j=1}^{\#\pi} (n_j - 1)!,$$

where n_j are the block sizes of π, and $\#\pi$ is the number of blocks in π. It follows from the above formula that the sequence (P_n) coincides with the CRP(α) on partitions when $\mathcal{G}_n$ is taken to be the set of adjacency matrices corresponding to partitions. When $\mathcal{G}_n$ is the set of permutations, similar reasoning shows that (P_n) is the same as the CRP(α) for permutations. For an introduction to the CRP for partitions and permutations, see Section 3.1 of Pitman (2006). Letting $\mathcal{G}_n = \{0,1\}^{n\times n}$ have unrestricted support, and including an additional "odds" parameter $\beta > 0$, the following collection of distributions, called the Permanental Graph Model, is obtained.

Theorem 2.2.1 (Permanental Graph Model)**.** *Let $\mathcal{G}_n = \{0,1\}^{n\times n}$ be the whole space, and put*

$$P_n(G) \propto \beta^{\#G}\mathrm{per}_\alpha(G), \tag{2.2}$$

for $G \in \mathcal{G}_n$, $\alpha, \beta > 0$, where $\#G = \sum_{i,j\le n} G_{ij}$ is the number of edges in G. Then the normalization constant is

$$z_n(\alpha,\beta) := \sum_{G\in\mathcal{G}_n} \beta^{\#G}\mathrm{per}_\alpha(G) = \alpha_{n\uparrow 1}\left(\frac{\beta}{1+\beta}\right)^n (1+\beta)^{n^2},$$

where $\alpha_{n\uparrow 1} := \alpha(\alpha+1)\cdots(\alpha+n-1)$ is the rising factorial starting at α. The degree distribution is given by

$$\sum_{j=1}^n G_{1j} - 1 \sim \mathit{Binom}\left(n-1, \frac{\beta}{1+\beta}\right).$$

A consequence of the above proposition is that the expected number of edges in this model grows as

$$\mathbb{E}\left[\sum_{i,j} G_{ij}\right] \sim \frac{\beta n^2}{1+\beta}, \tag{2.3}$$

in the sense of $a_n \sim b_n \iff a_n/b_n \to 1$.

2.2.1 Two Notions of Consistency

Before stating the negative result, we present two notions of projection on a graph.

Definition 2.2.1. *The subselection map $\varphi_n^{ss} : \{0,1\}^{(n+1)\times(n+1)} \to \{0,1\}^{n\times n}$ is defined by*

$$(\varphi_n^{ss}(G))_{ij} = G_{ij} \quad \text{for } i,j \in [n].$$

Definition 2.2.2. *The delete-and-repair map $\varphi_n^{dr} : \{0,1\}^{(n+1)\times(n+1)} \to \{0,1\}^{n\times n}$ is defined by*

$$\left(\varphi_n^{dr}(G)\right)_{ij} = G_{ij} \vee (G_{i(n+1)} \wedge G_{(n+1)j}) \quad \textit{for } i,j \in [n], \tag{2.4}$$

where $\vee$ *and* $\wedge$ *represent boolean "or" and "and" respectively.*

Note that the definition of the delete-and-repair projection mapping is specific to matrices with boolean valued entries, whereas the definition of the subselection projection mapping applies equally well to matrices whose entries are real valued. In words, given a graph on $n+1$ vertices, the delete-and-repair projection (2.4) deletes all edges connecting to node $n+1$, and repairs edges for pairs of nodes (including self pairs, (v,v)) between which there was a length 2 path going through node $n+1$. These notions of projection are illustrated in Figure 2.1.

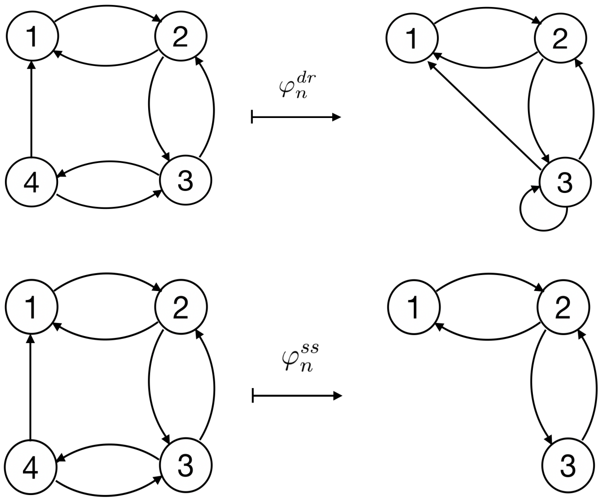


Figure 2.1: Pictured above (top) is an example of a directed graph on vertex set [4] projected down to a directed graph on [3] according to the delete-and-repair operation (2.4), and (bottom) the same graph projected down according to subselection.

The CRP(α) for partitions is recovered from (2.1) by setting $\mathcal{G}_n$ equal to the set of partition matrices, while the CRP(α) for permutations is recovered by setting $\mathcal{G}_n$ equal to the set of permutation matrices. It is straightforward to check that CRP(α) on partitions is consistent with respect to both subselection and delete-and-repair, which in this case are equivalent due to the transitivity property of equivalence relations. However, when viewed as a distribution on permutations, the CRP(α) is consistent only with respect to delete-and-repair.

Corollary 2.2.1. *Put* $\mathcal{G}_n = \{\Pi_\sigma \in \{0,1\}^{n\times n} : \sigma \in \mathcal{S}_n\}$, *where* $(\Pi_\sigma)_{ij} = 1 \iff \sigma(i) = j$, *and* $\beta = 1$ *in (2.2). Then the following probabilities,*

$$P_n(\sigma) = \frac{\alpha^{\#\sigma}}{\alpha_{n\uparrow 1}} \quad \textit{for } \sigma \in \mathcal{S}_n, \tag{2.5}$$

define a valid probability distribution on $\mathcal{S}_n$ that is delete-and-repair consistent. Here, $\alpha_{n\uparrow 1} := \alpha(\alpha+1)\cdots(\alpha+n-1)$ is the rising factorial starting at α. The above distribution is known as the CRP(α) for permutations.

From Corollary 2.2.1, it follows that sparsity and node-exchangeability are not mutually exclusive properties. Indeed, the Ewens permutations described by (2.5) are delete-and-repair consistent, node-exchangeable, and sparse, as they contain exactly $n = o(n^2)$ edges for each n.

To see why the CRP(α) for permutations is not subselection consistent, consider the permutation (123). It has adjacency matrix satisfying,

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \xmapsto{\varphi_n^{\text{ss}}} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

under subselection of the first two vertices. A permutation matrix must have a single “1” in each row and column, so the adjacency matrix on the right does not correspond to a permutation. Thus, these distributions cannot be consistent in the sense of subselection. However, the distribution can be specified by a generative a process (CRP seating plan), meaning that the law of total probability is satisfied; the distributions are consistent in some sense. Indeed, a more natural notion of projection in this example is delete-and-repair, for which we would instead obtain,

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \xmapsto{\varphi_n^{\text{dr}}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

according to the formula (2.4). For a permutation $\sigma \in \mathcal{S}_{n+1}$, the delete-and-repair operation deletes node $n+1$ from its cycle, while repairing an edge from the preimage of $n+1$ to the image of $n+1$ under σ. If $n+1$ is contained in its own cycle, then the node and the cycle are entirely removed, and no edges are repaired.

2.2.2 A Negative Result

As a graph is projected down to a smaller graph according to the boolean operation (2.4), edges may be repaired and thus added to the graph. It follows from this observation and (2.3) that, for the distributions (2.2) to be delete and repair consistent, it is natural to expect the β parameter to be decreasing in n. One may then suspect that (2.3) is $o(n^2)$ for a delete and repair consistent sequence (P_n). But, the next result states that the PGM(α_n, β_n) defined in (2.2), which has unrestricted support, i.e. $\mathcal{G}_n = \{0,1\}^{n\times n}$, does not admit a consistent sequence (P_n) in either sense described above, for any sequence of pairs (α_n, β_n).

Proposition 2.2.1. *For no sequence of pairs $(\alpha_n, \beta_n)_{n\in\mathbb{N}} > 0$ are the distributions*

$$P_n(G) \propto \beta_n^{\#G} \operatorname{per}_{\alpha_n}(G) \quad \textit{for } G \in \mathcal{G}_n := \{0,1\}^{n\times n}$$

delete-and-repair or subselection consistent. Equivalently, the statement

$$P_n(G) = \sum_{G' \in \mathcal{G}_{n+1} : \varphi_n^{\bullet}(G') = G} P_{n+1}(G') \quad \textit{for any } G \in \mathcal{G}_n, \textit{ for all } n \in \mathbb{N} \tag{2.6}$$

is not true, where $\bullet$ is either dr (delete-and-repair) or ss (subselection).

A sketch of the proof of Proposition 2.2.1 is provided below, while the full proof is presented in Section 2.4.

Proof sketch. When $\bullet$ is dr, the equation in (2.6) can be written

$$\frac{z_{n+1}(\alpha_{n+1}, \beta_{n+1})}{z_n(\alpha_n, \beta_n)} = \frac{\sum_{\sigma' \in \mathcal{S}_{n+1}} \alpha_{n+1}^{\sigma'} \sum_{G' : \varphi_n^{\mathrm{dr}}(G') = G} \beta_{n+1}^{\#G'} \mathbf{1}\{\sigma' \subset G'\}}{\beta_n^{\#G} \sum_{\sigma \in \mathcal{S}_n} \alpha_n^{\#\sigma} \mathbf{1}\{\sigma \subset G\}}. \tag{2.7}$$

One consequence is that the right hand side is constant in $G \in \{0,1\}^{4\times 4}$. The following two graphs,

$$G_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

have the same number of edges. G_1 contains the permutations (1234) and (12)(34). G_2 contains the permutations (1234) and (123)(4). These graphs are visualized in Figure 2.2. Hence the denominators in (2.7) are equal,

$$\beta_4^{\#G_1} \sum_{\sigma \in \mathcal{S}_4} \alpha_4^{\#\sigma} \mathbf{1}\{\sigma \subset G_1\} = \beta_4^{\#G_2} \sum_{\sigma \in \mathcal{S}_4} \alpha_4^{\#\sigma} \mathbf{1}\{\sigma \subset G_2\}.$$

The key observation used to exhibit a contradiction is that the numerator of the right hand side of (2.7) depends on the set of graphs which project down to G according to the delete and repair operation (2.4). It is shown in Section 2.4 that the set of graphs in $\mathcal{G}_5$ which project down to G_2 has greater cardinality than the corresponding set of graphs for G_1. Upon computing the right hand side of (2.7) for G_1 and G_2, it becomes clear that they are not equal for any pair $(\alpha_4, \beta_4), (\alpha_5, \beta_5) > 0$, hence contradicting the statement (2.6). This computation, along with the proof for subselection inconsistency, can be found in Section 2.4.

□

Note that the above proposition implies that no constant (in n) pair (α, β) allows for a consistent sequence of distributions (P_n) in either sense discussed above.

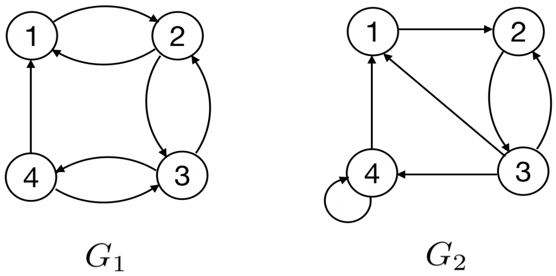


Figure 2.2: Graphs $G_1 = (1234) \vee (12)(34) \vee (23)$ and $G_2 = (1234) \vee (123)(4) \vee (23)$ for which the right hand sides of (2.7) are not the equal.

2.3 Discussion

We have investigated a collection of exchangeable distributions on graphs, defined via the α permanent. Setting $\mathcal{G}_n$ to be the set of all directed graphs allows for tractable calculations regarding the normalizing constant and degree distribution. A negative result was obtained; no choice of the parameters (α_n, β_n) yields a consistent collection of graph distributions in the sense of subselection or delete-and-repair. The following questions remain unaddressed: Can we distinguish between the permanental graph model and the Erdos-Renyi$(n, \frac{\beta}{1+\beta})$ graph as $n \to \infty$? Can we find a different delete-and-repair projective system $(\mathcal{G}_n)_{n \in \mathbb{N}}$ (meaning $\varphi_n^{\mathrm{dr}}(\mathcal{G}_{n+1}) \subset \mathcal{G}_n$) for which similar calculations are tractable? It would be surprising if the only delete-and-repair consistent graphs were exchangeable partitions and permutations.

2.4 Proofs

Proof of Theorem 2.2.1. The sum over all graphs $G \in \mathcal{G}_n$ is

$$\begin{aligned}
z_n(\alpha,\beta) &\coloneqq \sum_{G\in\mathcal{G}_n} \beta^{\#G} \sum_{\sigma} \alpha^{\#\sigma} \prod_{i=1}^{n} G_{i,\sigma(i)} \\
&= \sum_{\sigma} \alpha^{\#\sigma} \sum_{G\in\mathcal{G}_n} \beta^{\#G} \mathbf{1}\{\sigma \subset G\} \\
&= \underbrace{\sum_{\sigma} \alpha^{\#\sigma}}_{\alpha_{n\uparrow 1}} \sum_{m=n}^{n^2} \beta^m \underbrace{\sum_{G\in\mathcal{G}_n:\#G=m} \mathbf{1}\{\sigma \subset G\}}_{=\binom{n^2-n}{m-n}} \\
&= \alpha_{n\uparrow 1}\beta^n \sum_{m=0}^{n^2-n} \beta^m \binom{n^2-n}{m} \\
&= \alpha_{n\uparrow 1}\beta^n (1+\beta)^{n^2-n},
\end{aligned}$$

where the first underbrace in the third line is due to Lemma 2.2.1, and last equality is by the binomial formula. Hence the normalization constant is

$$z_n(\alpha,\beta) = \alpha_{n\uparrow 1}\left(\frac{\beta}{1+\beta}\right)^n (1+\beta)^{n^2}.$$

Put $G_{1\bullet} \coloneqq \sum_{j=1}^{n} G_{1j}$, whose distribution is computed below.

$$\begin{aligned}
\mathbb{P}(G_{1\bullet} = k+1) &= \sum_{G\in\mathcal{G}_n:G_{1\bullet}=k+1} P_n(G) \\
&= \frac{1}{z_n(\alpha,\beta)} \sum_{G\in\mathcal{G}_n:G_{1\bullet}=k+1} \beta^{\#G} \sum_{\sigma} \alpha^{\#\sigma} \mathbf{1}\{\sigma \subset G\} \\
&= \frac{1}{z_n(\alpha,\beta)} \sum_{\sigma} \alpha^{\#\sigma} \sum_{G\in\mathcal{G}_n:G_{1\bullet}=k+1} \beta^{\#G} \mathbf{1}\{\sigma \subset G\} \\
&= \frac{1}{z_n(\alpha,\beta)} \sum_{\sigma} \alpha^{\#\sigma} \sum_{m=n+k}^{n^2-(n-(k+1))} \beta^m \sum_{G\in\mathcal{G}_n:G_{1\bullet}=k+1,\#G=m} \mathbf{1}\{\sigma \subset G\}.
\end{aligned}$$

In the last equality, we have partitioned the terms in the sum over all $G_{1\bullet} = k+1$ into groups of graphs with m edges. Clearly any graph G in the event $\{G_{1\bullet} = k+1\}$, for which $\sigma \subset G$, has at least $n+k$ edges (n for the permutation, k for the additional edges in the first row), and at most $n^2 - (n-(k+1))$ edges (there must be exactly $n-(k+1)$ zeros in the first row). Continuing on, the above is equal to

$$= \frac{1}{z_n(\alpha,\beta)} \sum_\sigma \alpha^{\#\sigma} \sum_{m=n+k}^{n^2-(n-(k+1))} \beta^m \cdot \binom{n-1}{k} \cdot \binom{n^2-n-(n-1)}{m-(n+k)}.$$

The combinatorial factor counts the number of ways to pick the additional k edges in the first row, and the number of ways to pick the other $m-(n+k)$ edges from the $n^2-n-(n-1)$ possibilities, since n entries are fixed due to the permutation σ, and the remaining $n-1$ entries of the first row are fixed by the (exactly) k ones in the first row. The product of these two represents the number of graphs G with m total edges with $k+1$ in the first row, such that $\sigma \subset G$.

Factoring out β^{n+k} and shifting the summation index, the above becomes

$$\begin{aligned}
&= \frac{1}{z_n(\alpha,\beta)} \sum_\sigma \alpha^{\#\sigma} \beta^{n+k} \binom{n-1}{k} \sum_{m=0}^{n^2-(n-(k+1))-(n+k)} \beta^m \binom{n^2-n-(n-1)}{m} \\
&= \frac{1}{\alpha_{n\uparrow 1} \left(\frac{\beta}{1+\beta}\right)^n (1+\beta)^{n^2}} \alpha_{n\uparrow 1} \beta^{n+k} \binom{n-1}{k} \underbrace{\sum_{m=0}^{n^2-n-(n-1)} \beta^m \binom{n^2-n-(n-1)}{m}}_{=(1+\beta)^{n^2-n-(n-1)} \text{ by the binomial formula}} \\
&= \binom{n-1}{k} \beta^k (1+\beta)^{n-n^2} (1+\beta)^{n^2-n-(n-1)} \\
&= \binom{n-1}{k} \beta^k (1+\beta)^{-(n-1)}.
\end{aligned}$$

The above is the pmf of a Binom$(n-1, \frac{\beta}{1+\beta})$ variable at k.

□

Proof of Lemma 2.2.1. The validity of the probabilities

$$P_n(\sigma) = \frac{\alpha^{\#\sigma}}{\alpha_{n\uparrow 1}} \quad \text{for } \sigma \in \mathcal{S}_n$$

follows from the recursion

$$\begin{aligned}
\sum_{\sigma\in\mathcal{S}_n} \alpha^{\#\sigma} &= \sum_{\sigma\in\mathcal{S}_{n-1}} \Big((n-1)\alpha^{\#\sigma} + \alpha \cdot \alpha^{\#\sigma} \Big) \\
&= (\alpha + n - 1) \sum_{\sigma\in\mathcal{S}_{n-1}} \alpha^{\#\sigma}.
\end{aligned}$$

Letting $\varphi_n^{\mathrm{dr}} : \{0,1\}^{(n+1)\times(n+1)} \to \{0,1\}^{n\times n}$ denote the delete-and-repair mapping defined by (2.4), consistency amounts to showing

$$P_n(\sigma) = \sum_{\sigma' \in \mathcal{S}_{n+1} : \varphi_n^{\mathrm{dr}}(\sigma') = \sigma} P_{n+1}(\sigma').$$

The right hand side is a sum over the $n+1$ permutations $\sigma' \in \mathcal{S}_{n+1}$ that delete and repair down to σ. Exactly one of these permutations has $\#\sigma' = \#\sigma + 1$, namely the permutation obtained by placing $n+1$ in its own cycle. The other σ' have the same number of cycles as σ.

$$\begin{aligned}\sum_{\sigma' \in \mathcal{S}_{n+1} : \varphi_n^{\mathrm{dr}}(\sigma') = \sigma} P_{n+1}(\sigma') &= \frac{1}{\alpha_{(n+1)\uparrow 1}} \cdot \left(n\alpha^{\#\sigma} + \alpha^{\#\sigma+1}\right) \\ &= \alpha^{\#\sigma} \cdot \frac{\alpha+n}{\alpha_{(n+1)\uparrow 1}} \\ &= \frac{\alpha^{\#\sigma}}{\alpha_{n\uparrow 1}},\end{aligned}$$

as desired. A similar calculation yields delete and repair consistency for the partitions generated by the CRP(α).

□

Proof of Proposition 2.2.1. It must be the case that $\alpha_n, \beta_n > 0$ in order for the probabilities defined by

$$P_n(G) \propto \beta_n^{\#G} \mathrm{per}_{\alpha_n}(G)$$

to be valid for all $G \in \mathcal{G}_n = \{0,1\}^{n \times n}$. Keeping this in mind, we compute the right hand side of (2.7) for G_1 and G_2.

The set of $G' \in \mathcal{G}_5$ for which $\varphi_4^{\mathrm{dr}}(G') = G_1$ (and also contain at least one permutation) is the inverse image $(\varphi_4^{\mathrm{dr}})^{-1}(G_1)$, which is equal to

$$\left\{\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\1&0&1&0&0\\\ast&\ast&\ast&\ast&1\end{bmatrix},\begin{bmatrix}0&1&0&0&\ast\\1&0&1&0&\ast\\0&1&0&1&\ast\\1&0&1&0&\ast\\0&0&0&0&1\end{bmatrix},\begin{bmatrix}0&\ast&0&0&1\\1&0&1&0&0\\0&1&0&1&0\\1&0&1&0&0\\0&1&0&0&1\end{bmatrix},\begin{bmatrix}0&\ast&0&0&1\\1&0&1&0&0\\0&\ast&0&1&1\\1&0&1&0&0\\0&1&0&0&1\end{bmatrix},\right.$$

$$\begin{bmatrix}0&1&0&0&0\\\ast&0&1&0&1\\0&1&0&1&0\\1&0&1&0&0\\1&0&0&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\\ast&0&\ast&0&1\\0&1&0&1&0\\1&0&1&0&0\\1&0&1&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\\ast&0&1&0&1\\0&1&0&1&0\\\ast&0&1&0&1\\1&0&0&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\\ast&0&\ast&0&1\\1&0&1&0&1\end{bmatrix},$$

$$\begin{bmatrix}0&1&0&0&0\\1&0&\ast&0&1\\0&1&0&1&0\\1&0&\ast&0&1\\0&0&1&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\\ast&0&\ast&0&1\\0&1&0&1&0\\\ast&0&\ast&0&1\\1&0&1&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&\ast&0&1\\0&1&0&1&0\\1&0&1&0&0\\0&0&1&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&\ast&0&1&1\\1&0&1&0&0\\0&1&0&0&1\end{bmatrix},$$

$$\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&\ast&0&\ast&1\\1&0&1&0&0\\0&1&0&1&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&\ast&1\\1&0&1&0&0\\0&0&0&1&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\\ast&0&1&0&1\\1&0&0&0&1\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\1&0&\ast&0&1\\0&0&1&0&1\end{bmatrix},$$

$$\begin{bmatrix}0&\ast&0&0&1\\1&0&1&0&0\\0&1&0&1&0\\1&0&1&0&0\\0&1&0&0&0\end{bmatrix},\begin{bmatrix}0&\ast&0&0&1\\1&0&1&0&0\\0&\ast&0&1&1\\1&0&1&0&0\\0&1&0&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\\ast&0&1&0&1\\0&1&0&1&0\\1&0&1&0&0\\1&0&0&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\\ast&0&\ast&0&1\\0&1&0&1&0\\1&0&1&0&0\\1&0&1&0&0\end{bmatrix},$$

$$\begin{bmatrix}0&1&0&0&0\\\ast&0&1&0&1\\0&1&0&1&0\\\ast&0&1&0&1\\1&0&0&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&\ast&0&1\\0&1&0&1&0\\1&0&\ast&0&1\\0&0&1&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\\ast&0&\ast&0&1\\1&0&1&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\\ast&0&\ast&0&1\\0&1&0&1&0\\\ast&0&\ast&0&1\\1&0&1&0&0\end{bmatrix},$$

$$\begin{bmatrix}0&1&0&0&0\\1&0&\ast&0&1\\0&1&0&1&0\\1&0&1&0&0\\0&0&1&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&\ast&0&1&1\\1&0&1&0&0\\0&1&0&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&\ast&0&\ast&1\\1&0&1&0&0\\0&1&0&1&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&\ast&1\\1&0&1&0&0\\0&0&0&1&0\end{bmatrix},$$

$$\left.\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\\ast&0&1&0&1\\1&0&0&0&0\end{bmatrix},\begin{bmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\1&0&\ast&0&1\\0&0&1&0&0\end{bmatrix}\right\}$$

where $*$ means either 0 or 1. In total there are 139 graphs in $\mathcal{G}_5$ which delete and repair down to G_1 (that contain at least one permutation). Listed below are the permutations in $\mathcal{S}_5$ that project down to (1234) or (12)(34),

$$\begin{aligned}\{&(12345),(12354),(12534),(15234),(1234)(5),(125)(34),(152)(34),(12)(345),(12)(354),\\ &(12)(34)(5)\}.\end{aligned}$$

Going through each σ' in the above set in the order listed above and computing the sum

$$\sum_{G'\in(\varphi_4^{\mathrm{dr}})^{-1}(G_1)} \beta^{\#G'}\mathbf{1}\{\sigma'\subset G'\},$$

will show that in this case, the right hand side of (2.7) becomes

$$\begin{aligned}
&= \frac{1}{\beta_4^7\cdot(\alpha_4+\alpha_4^2)}\Big[\alpha_5(\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+3\beta_5^{10}+3\beta_5^{11}+\beta_5^{12}\\
&+\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+3\beta_5^9+3\beta_5^{10}+\beta_5^{11}+\beta_5^8+\beta_5^9)\\
&+\alpha_5(\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+\beta_5^9)+\alpha_5(\beta_5^9+2\beta_5^{10}+\beta_5^{11}\\
&+\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+3\beta_5^{10}+3\beta_5^{11}+\beta_5^{12}+\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9\\
&+\beta_5^{10}+\beta_5^8+3\beta_5^9+3\beta_5^{10}+\beta_5^{11}+\beta_5^8+\beta_5^9)+\alpha_5(\beta_5^9+\beta_5^{10}+\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^8+\beta_5^9\\
&+\beta_5^8+2\beta_5^9+\beta_5^{10})+\alpha_5^2(\beta_5^8+4\beta_5^9+6\beta_5^{10}+4\beta_5^{11}+\beta_5^{12}+4\beta_5^9+6\beta_5^{10}+4\beta_5^{11}+\beta_5^{12}+\beta_5^{10}\\
&+\beta_5^{10}+\beta_5^{11}+\beta_5^9+\beta_5^{10}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+2\beta_5^{11}\\
&+\beta_5^{12}+\beta_5^{10}+\beta_5^9+\beta_5^{10}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{10}+\beta_5^9+\beta_5^{10})+\alpha_5^2(\beta_5^9+\beta_5^{10}+\beta_5^9+2\beta_5^{10}\\
&+\beta_5^{11}+\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+3\beta_5^{10}+3\beta_5^{11}+\beta_5^{12}+\beta_5^8+\beta_5^9+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8\\
&+2\beta_5^9+\beta_5^{10}+\beta_5^8+3\beta_5^9+3\beta_5^{10}+\beta_5^{11})+\alpha_5^2(\beta_5^9+\beta_5^{10}+\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^8+\beta_5^9\\
&+\beta_5^8+2\beta_5^9+\beta_5^{10})+\alpha_5^2(\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+3\beta_5^{10}+3\beta_5^{11}+\beta_5^{12}\\
&+\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+3\beta_5^9+3\beta_5^{10}+\beta_5^{11}+\beta_5^8+\beta_5^9)\\
&+\alpha_5^2(\beta_5^9+2\beta_5^{10}+\beta_5^{11}+\beta_5^9+\beta_5^{10}+\beta_5^8+2\beta_5^9+\beta_5^{10}+\beta_5^8+\beta_5^9)+\alpha_5^3(\beta_5^8+4\beta_5^9\\
&+6\beta_5^{10}+4\beta_5^{11}+\beta_5^{12}+4\beta_5^9+6\beta_5^{10}+4\beta_5^{11}+\beta_5^{12}+\beta_5^{10}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{10}+\beta_5^{11}\\
&+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+2\beta_5^{11}+\beta_5^{12}+\beta_5^9+\beta_5^{10}+\beta_5^9\\
&+\beta_5^{10}+\beta_5^{10}+\beta_5^{11}+\beta_5^{10}+\beta_5^9+\beta_5^{10}+\beta_5^{10})\Big].
\end{aligned}$$

Simplifying, the above becomes

$$
\begin{aligned}
&= \frac{1}{\beta_4^7 \cdot (\alpha_4 + \alpha_4^2)} \Big[\alpha_5 (12\beta_5^8 + 34\beta_5^9 + 34\beta_5^{10} + 14\beta_5^{11} + 2\beta_5^{12}) \qquad (2.8) \\
&+ \alpha_5^2 (13\beta_5^8 + 45\beta_5^9 + 60\beta_5^{10} + 30\beta_5^{11} + 5\beta_5^{12}) + \alpha_5^3 (\beta_5^8 + 11\beta_5^9 + 26\beta_5^{10} + 16\beta_5^{11} + 3\beta_5^{12}) \Big].
\end{aligned}
$$

The set of $G' \in \mathcal{G}_5$ for which $\varphi_4^{\mathrm{dr}}(G') = G_2$ (and also contain at least one permutation) is the inverse image $(\varphi_4^{\mathrm{dr}})^{-1}(G_2)$, which is equal to the set

$$\left\{ \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ 1&0&0&1&0\\ \ast&\ast&\ast&\ast&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&\ast\\ 0&0&1&0&\ast\\ 1&1&0&1&\ast\\ 1&0&0&1&\ast\\ 0&0&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&\ast&0&0&1\\ 0&0&1&0&0\\ 1&1&0&1&0\\ 1&0&0&1&0\\ 0&1&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&\ast&0&0&1\\ 0&0&1&0&0\\ 1&\ast&0&1&1\\ 1&0&0&1&0\\ 0&1&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&\ast&0&1\\ 1&1&0&1&0\\ 1&0&0&1&0\\ 0&0&1&0&1 \end{bmatrix} \right.$$

$$\begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&1&1\\ 1&0&0&1&0\\ 1&0&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&\ast&0&1&1\\ 1&0&0&1&0\\ 1&1&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&\ast&0&\ast&1\\ 1&0&0&1&0\\ 0&1&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&\ast&1\\ 1&0&0&1&0\\ 1&0&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&\ast&0&\ast&1\\ 1&0&0&1&0\\ 1&1&0&1&1 \end{bmatrix}$$

$$\begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&1&1\\ \ast&0&0&1&1\\ 1&0&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&\ast&1\\ 1&0&0&\ast&1\\ 0&0&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ \ast&0&0&\ast&1\\ 1&0&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&\ast&1\\ \ast&0&0&\ast&1\\ 1&0&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&\ast&0&1&1\\ 1&0&0&1&0\\ 0&1&0&0&1 \end{bmatrix}$$

$$\begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&\ast&1\\ 1&0&0&1&0\\ 0&0&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ \ast&0&0&1&1\\ 1&0&0&0&1 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ 1&0&0&\ast&1\\ 0&0&0&1&1 \end{bmatrix}, \begin{bmatrix} 0&\ast&0&0&1\\ 0&0&1&0&0\\ 1&1&0&1&0\\ 1&0&0&1&0\\ 0&1&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&\ast&0&0&1\\ 0&0&1&0&0\\ 1&\ast&0&1&1\\ 1&0&0&1&0\\ 0&1&0&0&0 \end{bmatrix}$$

$$\begin{bmatrix} 0&1&0&0&0\\ 0&0&\ast&0&1\\ 1&1&0&1&0\\ 1&0&0&1&0\\ 0&0&1&0&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&1&1\\ 1&0&0&1&0\\ 1&0&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&\ast&0&1&1\\ 1&0&0&1&0\\ 1&1&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&\ast&0&\ast&1\\ 1&0&0&1&0\\ 0&1&0&1&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&\ast&1\\ 1&0&0&1&0\\ 1&0&0&1&0 \end{bmatrix}$$

$$\begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&\ast&0&\ast&1\\ 1&0&0&1&0\\ 1&1&0&1&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&1&1\\ \ast&0&0&1&1\\ 1&0&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ \ast&0&0&\ast&1\\ 1&0&0&1&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&\ast&1\\ 1&0&0&\ast&1\\ 0&0&0&1&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ \ast&1&0&\ast&1\\ \ast&0&0&\ast&1\\ 1&0&0&1&0 \end{bmatrix}$$

$$\left. \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&\ast&0&1&1\\ 1&0&0&1&0\\ 0&1&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&\ast&1\\ 1&0&0&1&0\\ 0&0&0&1&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ \ast&0&0&1&1\\ 1&0&0&0&0 \end{bmatrix}, \begin{bmatrix} 0&1&0&0&0\\ 0&0&1&0&0\\ 1&1&0&1&0\\ 1&0&0&\ast&1\\ 0&0&0&1&0 \end{bmatrix} \right\}$$

where $\ast$ means either 0 or 1. In total there are 163 graphs in $\mathcal{G}_5$ which delete and repair down to G_2 (that contain at least one permutation). Listed below are the permutations in

$\mathcal{S}_5$ that project down to (1234) or $(123)(4)$,

$$\begin{aligned}&\{(12345),(12354),(12534),(15234),(1234)(5),(1235)(4),(1253)(4),(1523)(4),(123)(45),\\ &\qquad (123)(4)(5)\}.\end{aligned}$$

Going through each σ' in the above set in the order listed above and computing the sum

$$\sum_{G' \in (\varphi_4^{\mathrm{dr}})^{-1}(G_1)} \beta^{\#G'} \mathbf{1}\{\sigma' \subset G'\},$$

will show that in this case, the right hand side of (2.7) becomes

$$\begin{aligned}
&= \frac{1}{\beta_4^7 \cdot (\alpha_4 + \alpha_4^2)} \Big[\alpha_5 (\beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12} \\
&+ \beta_5^9 + \beta_5^{10} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 3\beta_5^9 + 3\beta_5^{10} + \beta_5^{11} + \beta_5^8 + \beta_5^9) \\
&+ \alpha_5 (\beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} \\
&+ \beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12} + \beta_5^9 + \beta_5^{10} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 3\beta_5^9 \\
&+ 3\beta_5^{10} + \beta_5^{11} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 3\beta_5^9 + 3\beta_5^{10} + \beta_5^{11} + \beta_5^8 + \beta_5^9) + \alpha_5 (\beta_5^9 + \beta_5^{10} \\
&+ \beta_5^8 + \beta_5^9) + \alpha_5 (\beta_5^9 + \beta_5^{10} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^8 + \beta_5^9 + \beta_5^8 + 2\beta_5^9 + \beta_5^{10}) + \alpha_5^2 (\beta_5^8 \\
&+ 4\beta_5^9 + 6\beta_5^{10} + 4\beta_5^{11} + \beta_5^{12} + 4\beta_5^9 + 6\beta_5^{10} + 4\beta_5^{11} + \beta_5^{12} + \beta_5^{10} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + \beta_5^9 \\
&+ \beta_5^{10} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + 2\beta_5^{11} + \beta_5^{12} + \beta_5^{10} + \beta_5^{11} \\
&+ \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + 2\beta_5^{11} + \beta_5^{12} + \beta_5^9 + \beta_5^{10} + \beta_5^{10} + \beta_5^{10} + \beta_5^9 + \beta_5^{10}) \\
&+ \alpha_5^2 (\beta_5^9 + \beta_5^{10} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12} \\
&+ \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12} + \beta_5^8 + \beta_5^9 + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 \\
&+ 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 3\beta_5^9 + 3\beta_5^{10} + \beta_5^{11} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 3\beta_5^9 + 3\beta_5^{10} + \beta_5^{11}) \\
&+ \alpha_5^2 (\beta_5^9 + \beta_5^{10} + \beta_5^8 + \beta_5^9) + \alpha_5^2 (\beta_5^9 + \beta_5^{10} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^8 + \beta_5^9 + \beta_5^8 + 2\beta_5^9 \\
&+ \beta_5^{10}) + \alpha_5^2 (\beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12} + \beta_5^9 \\
&+ \beta_5^{10} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 2\beta_5^9 + \beta_5^{10} + \beta_5^8 + 3\beta_5^9 + 3\beta_5^{10} + \beta_5^{11} + \beta_5^8 + \beta_5^9) \\
&+ \alpha_5^3 (\beta_5^8 + 4\beta_5^9 + 6\beta_5^{10} + 4\beta_5^{11} + \beta_5^{12} + 4\beta_5^9 + 6\beta_5^{10} + 4\beta_5^{11} + \beta_5^{12} + \beta_5^{10} + \beta_5^{10} + \beta_5^{11} \\
&+ \beta_5^{10} + \beta_5^{10} + \beta_5^{10} + \beta_5^{11} + \beta_5^9 + 2\beta_5^{10} + \beta_5^{11} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + 2\beta_5^{11} + \beta_5^{12} + \beta_5^{10} \\
&+ \beta_5^{11} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + \beta_5^{11} + \beta_5^{10} + 2\beta_5^{11} + \beta_5^{12} + \beta_5^9 + \beta_5^{10} + \beta_5^9 + \beta_5^{10} + \beta_5^9 \\
&+ \beta_5^{10} + \beta_5^{10}) \Big].
\end{aligned}$$

Simplifying, the above becomes

$$= \frac{1}{\beta_4^7 \cdot (\alpha_4 + \alpha_4^2)} \Bigg[\alpha_5 (13\beta_5^8 + 38\beta_5^9 + 40\beta_5^{10} + 18\beta_5^{11} + 3\beta_5^{12}) + \alpha_5^2 (14\beta_5^8 + 50\beta_5^9 \tag{2.9}$$
$$69\beta_5^{10} + 37\beta_5^{11} + 7\beta_5^{12}) + \alpha_5^3 (\beta_5^8 + 12\beta_5^9 + 29\beta_5^{10} + 19\beta_5^{11} + 4\beta_5^{12}) \Bigg].$$

Setting expressions (2.8) and (2.9) equal to each other, we have

$$\begin{aligned} 0 &= \alpha_5 (\beta_5^8 + 4\beta_5^9 + 6\beta_5^{10} + 4\beta_5^{11} + \beta_5^{12}) + \alpha_5^2 (\beta_5^8 + 5\beta_5^9 + 9\beta_5^{10} + 7\beta_5^{11} + 2\beta_5^{12}) \\ &\quad + \alpha_5^3 (\beta_5^9 + 3\beta_5^{10} + 3\beta_5^{11} + \beta_5^{12}). \end{aligned}$$

Since $\alpha_5, \beta_5 > 0$, the right hand side of the above equality is greater than 0, a contradiction.

Next, we prove inconsistency with respect to subselection. The equation (2.6) can be rearranged as

$$\frac{z_{n+1}(\alpha_{n+1}, \beta_{n+1})}{z_n(\alpha_n, \beta_n)} \beta_n^{\#G} \sum_{\sigma} \alpha_n^{\#\sigma} \mathbf{1}\{\sigma \subset G\} = \sum_{\sigma'} \alpha_{n+1}^{\#\sigma'} \sum_{G' : \varphi_n^{\mathrm{ss}}(G') = G} \beta_{n+1}^{G'} \mathbf{1}\{\sigma' \subset G'\}. \tag{2.10}$$

Grouping the inner summands on the right hand side according to the number of edges in G', the right hand side is equal to

$$= \sum_{\sigma'} \alpha_{n+1}^{\#\sigma'} \sum_{m=\#G+1}^{\#G+2n+1} \beta_{n+1}^{m} \sum_{G' : \varphi_n^{\mathrm{ss}}(G') = G, \#G' = m} \mathbf{1}\{\sigma' \subset G'\},$$

since any G' for which $\varphi_n^{\mathrm{ss}}(G') = G$ has at most $\#G + 2n + 1$ edges (if node $n+1$ is connected to itself and all other nodes), and at least $\#G + 1$ edges for some $\sigma' \in \mathcal{S}_{n+1}$ to be contained in G'. The above can be split into two sums,

$$\begin{aligned} &= \sum_{\sigma' : \sigma'(n+1) = n+1} \alpha_{n+1}^{\#\sigma'} \sum_{m=\#G+1}^{\#G+2n+1} \beta_{n+1}^{m} \binom{2n}{m - (\#G+1)} \\ &\quad + \sum_{\sigma' : \sigma'(n+1) \neq n+1} \alpha_{n+1}^{\#\sigma'} \sum_{m=\#G+2}^{\#G+2n+1} \beta_{n+1}^{m} \binom{2n-1}{m - (\#G+2)}, \end{aligned}$$

since when $\sigma'(n+1) = n+1$ and $\sigma' \subset G'$, there are $2n$ unconstrained entries in G' which can be either zero or one. When $\sigma'(n+1) \neq n+1$ and $\sigma' \subset G'$, there are $2n-1$ unconstrained entries, because $G'_{\sigma^{-1}(n+1), n+1} = G'_{n+1, \sigma(n+1)} = 1$ are fixed. In both cases, the top left $n \times n$ submatrix is constrained to be exactly equal to G in order to have $\varphi_n^{\mathrm{ss}}(G') = G$. By the binomial formula, the above becomes

$$
\begin{aligned}
&= \beta_{n+1}^{\#G+1}(1+\beta_{n+1})^{2n} \sum_{\sigma':\sigma'(n+1)=n+1} \alpha_{n+1}^{\#\sigma'} + \beta_{n+1}^{\#G+2}(1+\beta_{n+1})^{2n-1} \sum_{\sigma':\sigma'(n+1)\neq n+1} \alpha_{n+1}^{\#\sigma'} \\
&= \beta_{n+1}^{\#G+1}(1+\beta_{n+1})^{2n}\alpha_{n+1}\,(\alpha_{n+1})_{n\uparrow 1} + \beta_{n+1}^{\#G+2}(1+\beta_{n+1})^{2n-1} n(\alpha_{n+1})_{n\uparrow 1},
\end{aligned}
$$

where in the last equality we have used Lemma 2.2.1. Combined with (2.10), we have shown

$$
\begin{aligned}
&\frac{z_{n+1}(\alpha_{n+1},\beta_{n+1})}{z_n(\alpha_n,\beta_n)}\left(\frac{\beta_n}{\beta_{n+1}}\right)^{\#G}\sum_\sigma \alpha_n^{\#\sigma}\mathbf{1}\{\sigma\subset G\} \\
&= \beta_{n+1}(1+\beta_{n+1})^{2n}(\alpha_{n+1})_{n\uparrow 1}\left[\alpha_{n+1}+\frac{n\beta_{n+1}}{1+\beta_{n+1}}\right]. \qquad (2.11)
\end{aligned}
$$

In particular, the above equality implies that the function $f:\{0,1\}^{n\times n}\to\mathbb{R}$ defined by

$$
f(G) \coloneqq \left(\frac{\beta_n}{\beta_{n+1}}\right)^{\#G}\sum_\sigma \alpha_n^{\#\sigma}\mathbf{1}\{\sigma\subset G\},
$$

is constant in G. Taking $G_1=(12\ldots n)$ and $G_2=(1)(23\ldots n)$, the requirement $f(G_1)=f(G_2)$ implies that $\alpha_n=\alpha_n^2$ for every n. Since we must have $\alpha_n>0$ for the probabilities (2.1) to be valid, this implies that $\alpha_n\equiv 1$, so that

$$
f(G) = \left(\frac{\beta_n}{\beta_{n+1}}\right)^{\#G}\sum_\sigma \mathbf{1}\{\sigma\subset G\}.
$$

Taking $G_1=(12\ldots n)$ and $G_2=(12\ldots n)\vee(n)$, the equation $f(G_1)=f(G_2)$ becomes

$$
\left(\frac{\beta_n}{\beta_{n+1}}\right)^{n} = \left(\frac{\beta_n}{\beta_{n+1}}\right)^{n+1} \Rightarrow \beta_n\equiv\beta
$$

is constant in n. Since $\alpha_n\equiv 1$ and $\beta_n\equiv\beta>0$, (2.11) becomes

$$
\frac{z_{n+1}(1,\beta)}{z_n(1,\beta)}\sum_\sigma \mathbf{1}\{\sigma\subset G\} = \beta(1+\beta)^{2n}\left[1+\frac{n\beta}{1+\beta}\right],
$$

for every G. Plugging in the formula for the normalization constant gives

$$
\frac{\beta}{1+\beta}(1+\beta)^{2n+1}\sum_\sigma \mathbf{1}\{\sigma\subset G\} = \beta(1+\beta)^{2n}\frac{\beta(n+1)+1}{1+\beta}.
$$

Simplifying, the above implies that a necessary condition of subselection consistency is that for every $G \in \{0,1\}^{n \times n}$,

$$\sum_{\sigma} \mathbf{1}\{\sigma \subset G\} = 1 + \frac{n\beta}{1+\beta}.$$

Since the left hand side is not constant in G, we reach a contradiction.

□

CHAPTER 3
CHANGEPOINT DETECTION BOUNDARIES

3.1 Introduction

Detecting a changepoint that occurs somewhere along a single sequence of numbers is a problem that has been studied since at least the middle of the 20$^{\text{th}}$ century (Page; 1955), and interest in its extension to multiple sequences has grown especially over the past two decades. Some of the recent interest can be attributed to technological advances in devices that provide measurements in the form of multidimensional data. An event of scientific relevance may be coded within a data set as one or more changepoints along a particular dimension of measurement.

1. *Genetics.* Abnormalities in an individual's DNA copy number measurements may reveal phenomena relating human genetics and disease. Array comparative genomic hybridization (aCGH) is a tool used to measure abnormalities with respect to a reference sample (Shah et al.; 2007). For each individual, a sequence of fluorescence intensities is measured; each element of the sequence corresponds to a distinct location on the individual's genome. The data can be represented by a $p \times n$ matrix of intensity measurements, where n is the number of individuals, and p is the number of locations on the genome where a measurement was made. Some regions along the genome may exhibit similar behavior among a sub-collection of the individuals, possibly related by common presence of a disease. It is then of interest to determine where, along the genome location dimension p, the average intensity of the signal becomes elevated or diminished. An approach using local fdr estimates to analyze copy number variation (CNV) data is developed by Efron and Zhang (2011), and Higher Criticism type test statistics for CNV analysis are developed by Jeng et al. (2013). Wang and Samworth (2018) project each column of a CNV dataset from Bleakley and Vert (2011) in the direction of the mean intensity change, and apply a univariate changepoint estimation procedure to the resulting single time series.

2. *Neuroscience.* A "resting state scan" involves a subject whose brain at rest is imaged for a period of time, and these data are used to understand how processes of the brain behave in the absence of external stimuli (Aston et al.; 2012). Certain analytical techniques require the assumption of stationarity of these measurements in time, but the extent to which this assumption is justified is not understood precisely (Cole et al.; 2010); changepoint detection methods provide an opportunity to assess deviations from stationarity. Functional magnetic resonance image (fMRI) is a tool used to measure a signal at each voxel (volume pixel) at a discretized set of points in time. In this setting, the data can be represented as a $p \times n$ matrix, where n is the number of discrete time points, and p is the number of voxels at which a sequence of measurements was taken. Aston and Kirch (2012) develop changepoint detection methods using ideas from principle components analysis to detect departures from stationarity in fMRI data.

We formulate the high dimensional changepoint testing problem as follows. The data is

$$X = \theta + E \in \mathbb{R}^{p \times n},$$

where E has independent $N(0,1)$ entries, and $\theta \in \mathbb{R}^{p \times n}$ is a mean matrix, in which a subset $S \subset [p]$ of the rows may potentially have a single changepoint, i.e. there is an index $t^* \in [n-1]$, before which each entry of $\theta_{j:}$, the j^{th} row of θ, is constant, and after which each entry of the row $\theta_{j:}$ is equal to a different constant, where $j \in S$. The null hypothesis is that none of the rows of θ have a changepoint. Informally, the testing problem can be written:

H_0 : For each $j \in [p]$, $\theta_{j:} \in \mathbb{R}^n$ is a constant vector.

H_1 : $s \in [p]$ of the rows in θ share a changepoint location $t^* \in [n]$ with signal strength ρ.

For a given level of sparsity $s := |S| \in [p]$ of non-null rows, separation of the hypotheses H_0 and H_1 is monotone in ρ. This suggests a critical level of signal strength ρ^* for which reliably distinguishing between H_0 and H_1 is impossible when $\rho < \rho^*$ and becomes just possible as ρ exceeds ρ^*. Liu et al. (2019) derive the formula for ρ^* up to a multiplicative constant, and parameterize the minimum signal strength in the alternative as a function of the true changepoint location t^*. In our formulation (3.10) of the alternative parameter space, the signal strength is also parametrized by t^* in a way that yields a formula for ρ^* that is sharp with respect to the multiplicative constant.

In this chapter, an interplay is shown between three features of the changepoint detection problem: the intensity of the signal change, the sparsity of the changes along the row indices $j \in [p]$, and the unknown location of the change along the column index $t^* \in [n-1]$. Here, the statistical difficulty of the changepoint detection problem is formulated in terms of the minimax testing risk,

$$\mathcal{R}(p, n, \rho, s) = \inf_{\text{test functions } \psi} \left[\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \psi + \sup_{\theta \in \Theta_1(\rho, s)} \mathbb{P}_\theta (1 - \psi) \right],$$

where Θ_0 and $\Theta_1(\rho, s) \subset \mathbb{R}^{p \times n}$ denote the sets of parameters satisfying the descriptions H_0 and H_1, and $\psi : \mathbb{R}^{p \times n} \to \{0, 1\}$ is a test function that indicates whether or not to reject the null.

A *detection boundary* (Donoho and Jin; 2004) is a formula that separates the parameter space of alternatives into a detectable region (those which can be reliably distinguished from the null) and an undetectable region (those which cannot). To derive this formula, an asymptotic relationship is assumed between the dimensions of the data and the sparsity of signals. For the changepoint setting, the relationship is defined by a pair $(\beta, a) \in (0, 1) \times \mathbb{R}_+$, where for example, in Theorem 3.2.1,

$$\log \log \log n \sim a \log p \tag{3.1}$$

$$s \sim p^{1-\beta}. \tag{3.2}$$

This is the primary asymptotic regime that yields a non-trivial characterization of ρ^*, the critical signal size. As a function of a, β, the critical signal size $\rho^*(a, \beta)$ is referred to as the detection boundary throughout. The particular relationship between a, β, n, s is inspired by the form of the non-asymptotic critical radius derived by Liu et al. (2019), in which the term $p \log\log(8n)$ appears as an inflated problem dimension. To obtain a tradeoff between p and n, one can consider an asymptotic regime in which $\log\log n$ is polynomial in p, or equivalently, $\log\log\log n \asymp \log p$. The inflated problem dimension $p \log\log(8n)$ is described in more detail in Section 3.1.1. One of our main results, Theorem 3.2.2, is that the critical signal level $\rho^*(a, \beta)$ takes the form

$$\rho^*_{\text{2-side}}(a, \beta)^2 = \begin{cases} p^{\frac{a-(1-2\beta)}{2}} & a \leq 1-2\beta \\ (a-(1-2\beta))\log p & 1-2\beta < a \leq 1-4\beta/3 \\ 2(\sqrt{1-a}-\sqrt{1-a-\beta})^2 \log p & 1-4\beta/3 < a \leq 1-\beta \\ p^{a-(1-\beta)} & a > 1-\beta, \end{cases}$$

assuming n, p, s are related via (3.1) and (3.2). The above critical radius is the asymptotic analogue of the finite sample result of Liu et al. (2019). Notably, it provides the exact constants in the logarithmic regime $1-2\beta < a \leq 1-\beta$, where previously only the dependence on p was known (Liu et al.; 2019). We note that the appearance of $\rho^*_{\text{2-side}}$ is similar to a detection boundary studied by Chan et al. (2015) in a related changepoint problem. The formulation of the changepoint in their setting involves mean vectors with at least three piecewise constant segments, and the asymptotic calibration of p and n are linked to the lengths of these pieces. A detailed comparison with our work is given in Section 3.3.2. The key methodological ideas used in this chapter to prove the feasibility of the testing problem in the detectable region are inspired by those found in Chan et al. (2015) and Liu et al. (2019). Specifically, penalization of a Berk–Jones type statistic (Berk and Jones; 1979), together with a construction of geometrically growing candidate changepoint locations, motivates a test statistic that adaptively tests for a changepoint with asymptotically negligible error. Further description of this testing procedure is given in Section 3.4.1.

In the formula $\rho^*_{\text{2-side}}$, the Ingster–Donoho–Jin (IDJ) boundary is recovered by nullifying the changepoint component of the problem, i.e. the IDJ formula, shown in (3.5) and discussed below in Section 3.1.1, coincides with $\rho^*_{\text{2-side}}$ when $a \to 0$. This formula also coincides with the non-asymptotic result of Liu et al. (2019). A more detailed comparison is presented in Sections 3.3.1 and 3.3.3. The connection between asymptotic and finite sample results is elaborated upon in the following literature review.

3.1.1 Related literature in sparse signal detection

A canonical problem in sparse signal detection involves testing whether a p dimensional Gaussian vector $X \sim N(\theta, I_p)$ arises from a mean $\theta = 0$, or a mean $\theta \neq 0$ with signal size $\|\theta\|_2 \geq \rho$, supported on $s \leq p$ coordinates. One way to characterize the difficulty of this testing problem is to calculate the critical radius $\rho^*(p, s)$, a formula depending on the dimension p and sparsity s that describes the minimal signal strength required to distinguish between the two scenarios,

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \|\theta\|_2 \geq \rho^*(p, s).$$

Collier et al. (2017) showed that the critical radius takes the form,

$$\rho^*(p,s)^2 \asymp \begin{cases} \sqrt{p} & s \geq \sqrt{p} \\ s \log(ep/s^2) & s < \sqrt{p}. \end{cases} \tag{3.3}$$

This result is non-asymptotic in the following sense. In terms of the sparsity and problem dimension, the above radius gives the order of the signal size at which the testing problem becomes solvable. In other words, $\|\theta\|_2$ must be at least as large as $\rho^*(p, s)$ in order to guarantee the existence of a testing procedure which, with high probability, has small Type I and II errors at a fixed dimension p, and sparsity level s. In order to obtain a well defined, exact constant in the above formula, one specifies an asymptotic relationship between p and s. To capture the elbow effect in s which occurs at $s \asymp \sqrt{p}$ in expression (3.3), a natural asymptotic regime to consider is,

$$s \sim p^{1-\beta} \qquad \beta \in (0, 1).$$

The asymptotic constant when $\beta > 1/2$ was first derived by Ingster (1999) and again by Donoho and Jin (2004) in the following Bayesian version of the problem,

$$H_0 : X_j \overset{\text{iid}}{\sim} N(0, 1) \quad \text{versus} \quad H_1 : X_j \overset{\text{iid}}{\sim} (1 - \varepsilon)N(0, 1) + \varepsilon N(\mu, 1), \tag{3.4}$$

for $j = 1, \dots, p$, where $\varepsilon := p^{-\beta}$ is the expected fraction of non-null coordinates, and $\mu := \sqrt{2r \log p}$ is the signal size, where $r > 0$ is a constant. An analysis of the testing problem (3.4) gives the following form of the critical constant in terms of the sparsity parameter β,

$$r^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leq 3/4 \\ (1 - \sqrt{1 - \beta})^2 & \frac{3}{4} < \beta < 1. \end{cases} \tag{3.5}$$

The constant $r^*(\beta)$ is critical in the following sense: when $r < r^*(\beta)$, no test is able to separate the null and alternative scenarios in (3.4) with vanishing error, while when $r > r^*(\beta)$, a sequence of consistent tests exists. By the Neyman–Pearson Lemma, the likelihood ratio test achieves the smallest sum of Type I and II errors in the testing problem (3.4), and is therefore consistent throughout the detectable region $\{(r, \beta) : r > r^*(\beta)\}$. However, an obstacle in directly implementing the likelihood ratio test is that doing so requires knowledge of the true parameters β, r, which are unknown when one observes only the vector $X \in \mathbb{R}^p$. Donoho and Jin (2004) resolved this issue by introducing a testing procedure based on Higher Criticism (Tukey; 1989), and proved its adaptive consistency throughout the detectable region.

Recently, the work of Collier et al. (2017) was extended by Liu et al. (2019) to the high dimensional changepoint detection problem, in which each element in the collection of p observations is a sequence $X_{j:} \in \mathbb{R}^n$. In other words, the observed data is a matrix $X \in \mathbb{R}^{p \times n}$, where each n-dimensional row may be viewed as a unit of observation. Note that this view is in contrast with another view of the changepoint detection problem, where each column $X_{:i} \in \mathbb{R}^p$ constitutes a unit of observation, yielding a total of n observations. However, we prefer to use $p \to \infty$ as the limiting index in order to make direct comparisons to existing sparse signal detection results, and to lessen the notational burden in stating the lower bound constructions.

In the high dimensional changepoint detection problem studied by Liu et al. (2019), the signal size of the matrix $\theta \in \mathbb{R}^{p \times n}$ is defined as a normalized mean shift at the changepoint location. This formulation results in a critical radius that can be directly compared to (3.3),

$$\rho^*(p, n, s)^2 \asymp \begin{cases} \sqrt{p \log\log(8n)} & s > \sqrt{p \log\log(8n)} \\ s \log\left(\frac{ep \log\log(8n)}{s^2}\right) + \log\log(8n) & s \leq \sqrt{p \log\log(8n)}. \end{cases} \tag{3.6}$$

The two critical radii are nearly identical when the problem dimension p in (3.3) is replaced with the inflated problem dimension $p \log\log(8n)$ in (3.6). The additive contribution of the $\log\log(8n)$ term in the sparse regime $s \leq \sqrt{p \log\log(8n)}$ arises from the difficulty of the one dimensional changepoint detection problem $p = s = 1$, which was determined by Gao et al. (2020) to have critical radius,

$$\rho^*(1, n, 1)^2 \asymp \log\log(8n). \tag{3.7}$$

The sparse high dimensional changepoint problem is at least as difficult as both the one dimensional changepoint problem, and the sparse normal means problem with the inflated problem dimension $p \log\log(8n)$. Combining this observation with the formulae (3.3) and (3.7), the expression (3.6) for the critical radius appears reasonable.

In light of the relation between the finite sample analysis of Collier et al. (2017), and the asymptotic analysis of Ingster (1999) and Donoho and Jin (2004), it is natural to ask for the asymptotic analogue of the finite sample critical radius derived by Liu et al. (2019) in the high dimensional sparse changepoint problem. This analysis is the contribution of the present chapter. As in the sparse means problem, in order to calculate the sharp constant and obtain a formula for the detection boundary, an asymptotic regime is specified. The form $p \log\log n$ of the inflated problem dimension for the changepoint problem suggests that a non-trivial detection boundary may arise when n and p contribute comparably to the problem dimension, for instance with $\log\log n$ growing polynomially in p. This observation motivates the calibration,

$$\begin{aligned} \log\log\log n &\sim a \log p \qquad a > 0, \\ s &\sim p^{1-\beta} \qquad \beta \in (0, 1). \end{aligned} \tag{3.8}$$

Another regime considered in this chapter is,

$$\begin{aligned} \log\log n &\sim a \log p \qquad a > 0, \\ s &\sim p^{1-\beta} \qquad \beta \in (0, 1]. \end{aligned} \tag{3.9}$$

Once either of these asymptotic settings is assumed, the goal of this chapter is to establish the following two points.

1. Derive an expression for the critical radius of testing $\rho^*(a, \beta)$ for each pair (a, β) of calibration parameters. The parameter a controls the number of possible changepoint locations and β controls the sparsity of true signals. The resulting formula $\rho^*(a, \beta)$, called the *detection boundary*, characterizes the minimum amount of signal required to detect a single aligned changepoint affecting a sparse fraction of rows in X.

2. Construct an adaptive test that is able to separate the null and alternative hypotheses whenever the signal size ρ exceeds the critical radius $\rho^*(a, \beta)$. The test is adaptive in the sense that it does not require knowledge of the true parameters (ρ, a, β), and it is consistent in the sense that it has Type I and II errors tending to zero as the dimensions p and n tend to infinity according to either (3.8) or (3.9).

3.1.2 Related literature in changepoint detection

In this work, we address the problem of detecting a single common changepoint location across multiple data sequences. A closely related problem is that of localization, in which the objective is to determine the location of the change, having verified that a change has occurred. Dating back to the middle of the 20th century, Page (1954) studied changepoint detection for a single sequence of data. Since then, changepoint detection and localization in a single data source has been the subject of intensive research, as surveyed in, for example, Horváth and Rice (2014). For the single sequence problem with multiple changepoints, Killick et al. (2012) give an approach (PELT) based on penalized likelihood and minimum description length, while Frick et al. (2014) tackle both estimation and inference, providing efficient algorithms (SMUCE) based on dynamic programming.

The extension of the changepoint estimation problem to multiple data sources has been studied since at least Horváth et al. (1999), who tested for changes in the mean in multiple dependent sequences. In a similar setting, Bai (2010) proposed a least squares method for estimating a change in the mean, and a method based on quasi-maximum likelihood that can also detect a change in the variance. Fewer methods are available for changepoint estimation in the high dimensional setting subject to a sparsity constraint on the number of sequences exhibiting a change; examples based on CUSUM statistics include Cho and Fryzlewicz (2015) and Wang and Samworth (2018). Zhang et al. (2010) develop a generalized likelihood ratio test based on a chi–squared statistic from each data sequence, emphasizing detection of rare and weak changes in the signal. Kovács, Li, Bühlmann and Munk (2020), Liu et al. (2019), and Kovács, Li, Haubner, Munk and Bühlmann (2020) provide methods using an idea based on checking a geometrically growing grid of candidate changepoint locations, one which is closely related to the optimal testing procedures presented in the current chapter. An overview of recent advances in minimax optimal changepoint detection and localization in the high dimensional setting can be found in Verzelen et al. (2020) and Pilliat et al. (2020).

Chapter organization. The main results for the asymptotic regime (3.8) are stated in Section 3.2.2, whereas those for the regime (3.9) are given in Section 3.3.3. Various comparisons are made to existing results in the signal detection literature in Section 3.3 and 3.4.2 An overview of the lower bound construction and the intuition for the test statistic used to achieve the upper bound is given in Section 3.4. All rigorous arguments are contained in Section 3.5, and specific technical details are deferred to Section 3.6.

Notation. For $p \in \mathbb{N}$, we write $[p] = \{1, \dots, p\}$. Given $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two positive sequences a_n and b_n, we write $a_n \lesssim b_n$ and $a_n \leq O(b_n)$ to mean that there exists a constant $C > 0$ independent of n such that $a_n \leq C b_n$ for all n; moreover, $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We overload notation by using $a_n \sim b_n$ to denote $a_n / b_n \to 1$, and $X \sim F$ to mean the random variable X is distributed according to F; the intended usage of "$\sim$" will be made clear from context. We write $a_n = o(b_n)$ when $\limsup_{n\to\infty} a_n/b_n = 0$ and $a_n = \omega(b_n)$ when $\liminf_{n\to\infty} a_n/b_n = \infty$. For a set S, we use $\mathbf{1}_S$ and $|S|$ to denote its indicator function and cardinality respectively. For a vector $v = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, we define $\|v\|^2 = \sum_{\ell=1}^d v_\ell^2$. For a matrix $X \in \mathbb{R}^{p\times n}$ we let $X_{j:}$ denote the j^{th} row of X, and $X_{:i}$ denote the i^{th} column of X. The notation $X_{j,1:i} \in \mathbb{R}^i$ denotes the first i elements of the j^{th} row in X. We use ϕ and Φ to denote the standard normal density and CDF, respectively, and put $\bar{\Phi} := 1 - \Phi$. The notation $\mathbb{P}$ and $\mathbb{E}$ are generic probability and expectation operators whose distribution is determined from the context. In the context of testing H_0 vs H_1, the notation $\mathbb{P}_0, \mathbb{E}_0, \mathrm{Var}_0$ and $\mathbb{P}_1, \mathbb{E}_1, \mathrm{Var}_1$ refer to the probability, expectation, and variance operators in H_0 and H_1 respectively. Unless otherwise specified, log denotes the natural log. Z denotes a generic standard normal variable, whose dimensions will be clear from context.

3.2 Detection Boundary and Minimax Testing

3.2.1 Problem setting

An observation takes the form,

$$X = \theta + E \in \mathbb{R}^{p\times n},$$

where each $E_{ij} \overset{\text{iid}}{\sim} N(0,1)$, and the mean matrix θ may contain a shift in mean at some column index. The null scenario is that each row of θ is a constant vector,

$$\Theta_0(p,n) := \big\{\theta \in \mathbb{R}^{p\times n} : \text{ For all } j \leq p, \text{ there exists a } \mu_j \in \mathbb{R} \text{ s.t. } \theta_{ji} = \mu_j \text{ for all } 1 \leq i \leq n.\big\}$$

The alternative scenario allows for a fraction of rows of θ to share a common index $t^* \in [n-1]$, before which the entries in the row are constant, and after which the entries are equal to a different constant; t^* is called the changepoint location. We consider two alternative scenarios, called the one sided and two sided versions of the alternative, that lead to slightly different formulae for the critical testing radius.

One sided changepoint. In the one-sided version of the problem, the mean shift at the changepoint location is decreasing, i.e. the mean of the observations in a non-null row is larger before t^* than afterwards,

$$\begin{aligned}\Theta_1^{\text{1-side}}(p,n,\rho,s) := \Bigg\{\theta \in \mathbb{R}^{p\times n} : \sum_{j=1}^{p} \mathbf{1}_{\{\text{row } j \text{ has a changepoint}\}} \geq s, \text{ and } \exists t^* \in [n-1] \text{ s.t.} \\ \sqrt{\frac{t^*(n-t^*)}{n}}(\mu_{j1} - \mu_{j2}) \geq \rho \text{ for } j \in [p] \text{ non-null, where } \theta_{ji} = \mu_{j1} \text{ for } i \leq t^*, \\ \text{and } \theta_{ji} = \mu_{j2} \text{ for } i > t^*\Bigg\}, \qquad (3.10)\end{aligned}$$

where the signal size is $\rho > 0$. The factor $\frac{t^*(n-t^*)}{n}$ in the normalization of the signal in the definition of $\Theta_1^{\text{1-side}}$ is the precision of the natural test statistic for testing for a mean shift in row j when the changepoint location t^* is known,

$$\mathsf{Var}_0\left(\frac{1}{t^*}\sum_{i=1}^{t^*} X_{ji} - \frac{1}{n-t^*}\sum_{i=t^*+1}^{n} X_{ji}\right) = \frac{n}{t^*(n-t^*)}.$$

This precision factor can be thought of as an effective sample size given the changepoint; the above statistic is most powerful when t^* is close to $n/2$.

Two sided changepoint. In the two-sided version of the problem, the shift in mean at the changepoint location is no longer required to be decreasing. The parameter space is written,

$$\begin{aligned}\Theta_1^{\text{2-side}}(p,n,\rho,s) := \Bigg\{\theta \in \mathbb{R}^{p\times n} : \sum_{j=1}^{p} \mathbf{1}_{\{\text{row } j \text{ has a changepoint}\}} \geq s, \text{ and } \exists t^* \in [n-1] \text{ s.t.} \\ \sqrt{\frac{t^*(n-t^*)}{n}}|\mu_{j1} - \mu_{j2}| \geq \rho \text{ for } j \in [p] \text{ non-null, where } \theta_{ji} = \mu_{j1} \text{ for } i \leq t^*, \\ \theta_{ji} = \mu_{j2} \text{ for } i > t^*\Bigg\}. \qquad (3.11)\end{aligned}$$

Since $\Theta_1^{\text{1-side}}(p,n,\rho,s) \subset \Theta_1^{\text{2-side}}(p,n,\rho,s)$, the minimum signal size required for the existence of consistent tests is larger for the two sided version of the problem; it is at least as difficult as the one sided version of the problem.

3.2.2 Main Results

Consider the hypothesis testing problem,

$$H_0 : \theta \in \Theta_0(p, n) \quad \text{versus} \quad H_1 : \theta \in \Theta_1^{\text{1-side}}(p, n, \rho, s). \tag{3.12}$$

The minimax testing error is defined as

$$\mathcal{R}_{\text{1-side}}(p, n, \rho, s) := \inf_{\psi} \left[\sup_{\theta \in \Theta_0(p,n)} \mathbb{P}_\theta \psi + \sup_{\theta \in \Theta_1^{\text{1-side}}(p,n,\rho,s)} \mathbb{P}_\theta (1 - \psi) \right], \tag{3.13}$$

where the infimum is taken over all measurable test functions $\psi : \mathbb{R}^{p\times n} \to \{0, 1\}$ of X. The first calibration we consider is,

$$\begin{aligned} \log\log\log n &\sim a \log p \\ s &\sim p^{1-\beta}, \end{aligned}$$

where $\beta \in (0, 1)$ and $a > 0$. Our main result is that the following formula characterizes the limits of detection in the testing problem (3.12).

$$\rho^*_{\text{1-side}}(a, \beta)^2 := \begin{cases} p^{a-(1-2\beta)} & a \leq 1 - 2\beta \\ (a - (1 - 2\beta)) \log p & 1 - 2\beta < a \leq 1 - 4\beta/3 \\ 2(\sqrt{1-a} - \sqrt{1-a-\beta})^2 \log p & 1 - 4\beta/3 < a \leq 1 - \beta \\ p^{a-(1-\beta)} & a > 1 - \beta. \end{cases}$$

The characterization is in the following sense.

Theorem 3.2.1 (One sided)**.** *For the one sided testing problem, put $\rho := \rho^*_{\textit{1-side}}(a, \beta_1)$ for $a > 0$ and $\beta_1 \in (0, 1)$.*

1. *(Lower Bound) If $1 > \beta > \beta_1$, then $\mathcal{R}_{\textit{1-side}}(p, n, \rho, s) \not\to 0$, as $p \to \infty$.*
2. *(Upper Bound) If $0 < \beta < \beta_1$, then $\mathcal{R}_{\textit{1-side}}(p, n, \rho, s) \to 0$, as $p \to \infty$.*

When $\beta > \beta_1$, testing $\Theta_0(p, n)$ versus $\Theta_1^{\text{1-side}}(p, n, \rho, s = p^{1-\beta})$ is harder than testing $\Theta_0(p, n)$ versus $\Theta_1^{\text{1-side}}(p, n, \rho, s = p^{1-\beta_1})$, since the alternative in the former testing problem has a smaller fraction of non-null rows. Consequently, the first part of Theorem 3.2.1 states that $\rho^*_{\text{1-side}}(a, \beta_1)$ is a lower bound for the minimum required signal for a consistent testing procedure to exist. When $\beta < \beta_1$, a consistent testing procedure exists for testing $\Theta_0(p, n)$ versus $\Theta_1^{\text{1-side}}(p, n, \rho, s = p^{1-\beta})$, so that the second part of Theorem 3.2.1 implies that $\rho^*_{\text{1-side}}(a, \beta_1)$ is an upper bound for the minimum required signal. The result has been stated in terms of the sparsity parameter β in order to accommodate all regimes of the signal.

Indeed, the sharp constant is the coefficient of $\log p$ in the signal when $\rho^*_{\text{1-side}}$ is logarithmic in p, but when $\rho^*_{\text{1-side}}$ is polynomial in p, the corresponding constant is in the exponent of the signal size. Before making a comparison between Theorem 3.2.1, existing results in sparse signal detection, and the non-asymptotic rate of Liu et al. (2019), we state the analogous detection boundary for the two sided version of the problem,

$$H_0 : \theta \in \Theta_0(p, n) \quad \text{versus} \quad H_1 : \theta \in \Theta_1^{\text{2-side}}(p, n, \rho, s). \tag{3.14}$$

The minimum signal size required to separate H_0 from H_1 is characterized by the following detection boundary,

$$\rho^*_{\text{2-side}}(a, \beta)^2 := \begin{cases} p^{\frac{a-(1-2\beta)}{2}} & a \leq 1 - 2\beta \\ (a - (1 - 2\beta)) \log p & 1 - 2\beta < a \leq 1 - 4\beta/3 \\ 2(\sqrt{1-a} - \sqrt{1-a-\beta})^2 \log p & 1 - 4\beta/3 < a \leq 1 - \beta \\ p^{a-(1-\beta)} & a > 1 - \beta. \end{cases}$$

$\rho^*_{\text{2-side}}(a, \beta)$ is larger than $\rho^*_{\text{1-side}}(a, \beta)$ in the first case, $a < 1 - 2\beta$, corresponding to the dense regime. The two boundaries are identical in all other cases. Put

$$\mathcal{R}_{\text{2-side}}(p, \rho, a, \beta) := \inf_{\psi} \left[\sup_{\theta \in \Theta_0(p,n)} \mathbb{P}_\theta \psi + \sup_{\theta \in \Theta_1^{\text{2-side}}(p,n,\rho,s)} \mathbb{P}_\theta (1 - \psi) \right].$$

The following theorem is the two sided analogue of Theorem 3.2.1.

Theorem 3.2.2 (Two sided)**.** *For the two sided testing problem, put $\rho := \rho^*_{\textit{2-side}}(a, \beta_1)$ for $a > 0$ and $\beta_1 \in (0, 1)$.*

1. *(Lower Bound) If $1 > \beta > \beta_1$, then $\mathcal{R}_{\textit{2-side}}(p, n, \rho, s) \not\to 0$, as $p \to \infty$.*
2. *(Upper Bound) If $0 < \beta < \beta_1$, then $\mathcal{R}_{\textit{2-side}}(p, n, \rho, s) \to 0$, as $p \to \infty$.*

Next, we outline several connections between Theorems 3.2.1 and 3.2.2 and various results in the signal detection literature. The detection boundary result regarding the asymptotic regime (3.9) is motivated by the relation between Theorem 3.2.2 and the non-asymptotic rate of Liu et al. (2019), and is stated in Section 3.3.3. The proofs of Theorem 3.2.1 and 3.2.2 are contained in Sections 3.5.1, 3.5.2, 3.5.3, and 3.5.4.

3.3 Connections to related works

3.3.1 Connection to Ingster–Donoho–Jin boundary

We note that the minimum signal size $2r^*(\beta) \log p$ for detection in the sparse normal means problem studied by Ingster (1999) and Donoho and Jin (2004) is recovered by setting $a = 0$

in the one-sided changepoint problem,

$$\rho^*_{\text{1-side}}(0,\beta)^2 = \begin{cases} p^{2\beta-1} & 0 < \beta \leq \frac{1}{2} \\ (2\beta-1)\log p & \frac{1}{2} < \beta \leq \frac{3}{4} \\ 2(1-\sqrt{1-\beta})^2 \log p & \frac{3}{4} < \beta < 1. \end{cases}$$

The case $a = 0$ corresponds to the case when n does not grow with p. Then each row can effectively be treated as a scalar (the signal size) since the number of possible changepoint locations does not grow with p. In the $p \to \infty$ limit, the critical radius in the case $a = 0$ coincides with that of the sparse normal means problem, and its extension to the case $\beta \in (0, \frac{1}{2})$ by Cai et al. (2011). The two sided counterpart to the sparse normal means problem (3.4) is

$$H_0 : X_j \overset{\text{iid}}{\sim} N(0,1) \quad \text{versus} \quad H_1 : X_j \overset{\text{iid}}{\sim} (1-\varepsilon)N(0,1) + \varepsilon\left(\frac{1}{2}N(\mu,1) + \frac{1}{2}N(-\mu,1)\right). \tag{3.15}$$

Setting $a = 0$ in the two sided detection boundary $\rho^*_{\text{2-side}}(a,\beta)$ gives,

$$\rho^*_{\text{2-side}}(0,\beta)^2 = \begin{cases} p^{\beta-\frac{1}{2}} & 0 < \beta \leq \frac{1}{2} \\ (2\beta-1)\log p & \frac{1}{2} < \beta \leq \frac{3}{4} \\ 2(1-\sqrt{1-\beta})^2 \log p & \frac{3}{4} < \beta < 1, \end{cases}$$

recovering the formula given in Section 6A of Cai and Wu (2014) for the testing problem (3.15). Note that the two detection boundaries are identical except for the case $0 < \beta < \frac{1}{2}$. The factor of 2 difference in the exponent in this case is discussed in Section 6A of Cai and Wu (2014), where it is attributed to the symmetrization of the problem (3.15). Inspecting the formulae for $\rho^*_{\text{1-side}}(a,\beta)$ and $\rho^*_{\text{2-side}}(a,\beta)$ for general a, one sees that the same "symmetrization" phenomenon occurs in the dense regime ($a \leq 1 - 2\beta$) of the high dimensional changepoint problem.

3.3.2 Connection to the detection boundary in Chan et al. (2015)

Chan et al. (2015) studied a version of the high dimensional changepoint problem with multiple changepoints. In this model, the observation takes the form,

$$X = \theta + E \in \mathbb{R}^{p\times n},$$

where $E_{ij} \overset{\text{iid}}{\sim} N(0,1)$. The null hypothesis is simple, $\theta = 0$. The alternative is composite; under the alternative, there are $q > 0$ many disjoint intervals,

$$S_n^{(k)} := (j_n^{(k)}, j_n^{(k)} + \ell_n^{(k)}] \subset [n], \quad k = 1, \ldots, q,$$

over which the mean may be elevated. More precisely, each θ_{jt} is distributed,

$$\theta_{jt} = \begin{cases} \frac{\mu^{(k)} I_j^{(k)}}{\sqrt{\ell_n^{(k)}}} & t \in S_n^{(k)} \\ 0 & t \notin \cup_{k=1}^q S_n^{(k)} \end{cases}, \qquad I_j^{(k)} \overset{\text{iid}}{\sim} \text{Bern}(\varepsilon^{(k)}), \qquad (j,k) \in [p] \times [q],$$

where $\varepsilon^{(k)} := p^{-\beta^{(k)}}$ for $\beta^{(k)} \in (0,1)$, and $\mu^{(k)} > 0$ is the signal size. They consider the asymptotic setting

$$\log\log\left(\frac{n}{\ell_n^{(k)}}\right) \sim a^{(k)} \log p, \qquad a^{(k)} > 0. \qquad\qquad (k = 1, \ldots, q)$$

The detection boundary $\rho^*_{\text{1-side}}(a, \beta)$ for the one-sided changepoint problem appears as a critical radius for this problem in the following sense.

Theorem 3.3.1 (Chan and Walther)**.** *Put* $\mu^{(k)} := \rho^*_{\textit{1-side}}(a^{(k)}, \beta^{(k)})$.

1. *(Upper bound.) Put* $\varepsilon^{(k)} := p^{-\beta^{(k)} + \delta^{(k)}}$ *for some* $0 < \beta^{(k)} < 1$*, and* $0 < \delta^{(k)} \leq \beta^{(k)}$*. Then there exists a sequence of tests with Type I and II errors tending to zero as* $p \to \infty$*.*

2. *(Lower bound.) Put* $\varepsilon^{(k)} := p^{-\beta^{(k)} - \delta^{(k)}}$ *for some* $0 < \beta^{(k)} < 1$*, and* $\delta^{(k)} > 0$*. Then there does not exist a sequence of tests with Type I and II errors tending to zero as* $p \to \infty$*.*

Comparison of problem formulations and proof technique

To lessen the notational burden, and to make a direct comparison to the problem analyzed in this chapter, consider the case $q = 1$. We note that in Chan et al. (2015), the "Lower bound" of Theorem 3.3.1 is proved using a construction with $q = 1$. Since the hardest instance of this testing problem can be formulated in the $q = 1$ setting, by restricting attention to this case, no essential problem difficulties related to the form of the detection boundary are lost.

If the j^{th} row of θ is non-null, there is one contiguous region $S_n \subset [n]$ over which the mean vector $\theta_{j:}$ is elevated, i.e. $\theta_{jt} > 0$ for $t \in S_n$, and $\theta_{jt} = 0$ for $t \notin S_n$. Assuming this simplified setup, the following comparisons to the problem studied in the current chapter can be made.

1. According to the above asymptotic regime, the length $\ell_n = |S_n|$ of this elevated region is determined by the relation,

$$\log\log\left(\frac{n}{\ell_n}\right) \sim a \log p.$$

The calibration considered in this chapter is,

$$\log\log(\log n) \sim a \log p,$$

and the length of the elevated region is an integer $t^* \in [n-1]$, as defined in the parameter spaces $\Theta_1^{\text{1-side}}$ and $\Theta_1^{\text{2-side}}$. Essentially, the parenthesized terms n/ℓ_n and $\log n$ represent the number of candidate changepoint locations in the lower bound constructions, discussed in the next point.

2. In the lower bound construction for Theorem 3.3.1, the changepoint location is selected uniformly at random from n/ℓ_n many candidate positions in $[n - \ell_n]$. The likelihood ratio for this construction involves a sum of independent random variables, upon which the authors applied the Lyapunov Central Limit Theorem. In the lower bound construction for Theorem 3.2.1, a changepoint location is selected uniformly from a set of candidate positions $\mathcal{T} \subset [n]$. Since the elevated segment starts at index 1 and ends at the changepoint $t^* \in \mathcal{T}$, the resulting elevated segments (corresponding to the candidate changepoint locations $t^* \in \mathcal{T}$) overlap. As a result, the likelihood ratio for this construction involves a sum of dependent random variables. Instead of using the CLT, we use a more direct argument similar to the proof of Theorem 2(a) in Hu et al. (2021) to handle this case, and also some standard arguments based on bounding the second moment of the likelihood ratio for the simple versus simple hypothesis testing problem in the lower bound construction.

3. For the upper bound proof for Theorem 3.3.1, several procedures are provided in Chan et al. (2015). The one most similar to the procedure used to prove the upper bounds in Theorems 3.2.1, 3.2.2, and 3.3.2, takes a maximum over a collection of Berk–Jones statistics, each computed with respect to a candidate elevated segment in the non-null rows of θ. This maximum is then penalized by subtracting a quantity depending on the cardinality of the grid of candidate changepoint locations. The construction of an optimal test in the current chapter uses a geometrically growing changepoint locations similar to those used in the optimal test for achieving the lower bound in the finite sample detection boundary in Liu et al. (2019). The ideas from the procedures in Chan et al. (2015) and Liu et al. (2019), and how they are combined to yield optimal procedures for the testing problems considered in the current chapter, are detailed in Section 3.4.1.

3.3.3 Connection to non-asymptotic rate in Liu et al. (2019)

In Liu et al. (2019), the critical radius (3.6) is defined with respect to the 2-norm of the p dimensional vector of mean differences at the changepoint location $t^* \in [n]$,

$$\begin{aligned}\Theta^{(t^*)}(p,n,\rho,s) \coloneqq \Big\{&\theta = (\theta_1,\dots,\theta_n) \in \mathbb{R}^{p\times n} : \theta_t = \mu_1 \text{ for some } \mu_1 \in \mathbb{R}^p \text{ for all } 1 \le t \le t^*,\\ &\theta_t = \mu_2 \text{ for some } \mu_2 \in \mathbb{R}^p \text{ for all } t^*+1 \le t \le n,\\ &\|\mu_1-\mu_2\|_0 \le s, \sqrt{\frac{t^*(n-t^*)}{n}}\|\mu_1-\mu_2\|^2 \ge \rho^2\Big\}.\end{aligned}$$

Taking the union over all possible changepoint locations $t^* \in [n-1]$ gives the alternative hypothesis parameter space,

$$\Theta_1(p,n,\rho,s) \coloneqq \bigcup_{t^*=1}^{n-1} \Theta^{(t^*)}(p,n,\rho,s).$$

The null parameter space is the same as the one considered defined in the current chapter, $\Theta_0(p,n)$. The problem is to test,

$$H_0 : \theta \in \Theta_0(p,n) \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1(p,n,\rho,s),$$

for which it is shown that (3.6) is the detection boundary. Note that the signal size ρ in $\Theta_1(p,n,\rho,s)$ is defined with respect to the Euclidean norm of the entire vector (in $\mathbb{R}^p$) of changes, as opposed to the signal size required on each individual non-null row. To obtain a quantity comparable to the critical radius $\rho^*_{\text{2-side}}$, divide expression (3.6) by s for the minimum required signal in each non-null row,

$$\frac{\rho^*(p,n,s)^2}{s} \asymp \begin{cases} \frac{\sqrt{p\log\log(8n)}}{s} & s > \sqrt{p\log\log(8n)} \\ \log\left(\frac{ep\log\log(8n)}{s^2}\right) + \frac{\log\log(8n)}{s} & s \le \sqrt{p\log\log(8n)}. \end{cases} \tag{3.16}$$

According to the calibration,

$$\begin{aligned}\log\log\log n &\sim a\log p,\\ s &\sim p^{1-\beta},\end{aligned}$$

the "sparse" case, $s \le \sqrt{p\log\log n} \iff 1 + a - 2(1-\beta) \ge 0$, in (3.16) can be further split into the following cases,

$$\log\left(\frac{ep\log\log n}{s^2}\right) + \frac{\log\log n}{s} \asymp \begin{cases} \log\left(ep^{1+a-2(1-\beta)}\right) & a-(1-\beta) \le 0 \\ p^{a-(1-\beta)} & a-(1-\beta) > 0. \end{cases}$$

Then, rewriting (3.16) to include this further division, and ignoring constants, we have

$$\frac{\rho^*(p,n,s)^2}{s} \asymp \begin{cases} p^{\frac{a-(1-2\beta)}{2}} & a \leq 1-2\beta \\ \log p & 1-2\beta < a \leq 1-\beta \\ p^{a-(1-\beta)} & a > 1-\beta, \end{cases} \tag{3.17}$$

which coincides with the formula $\rho^*_{\text{2-side}}(a,\beta)^2$ up to the coefficient of the $\log p$ rate in the case $1-2\beta < a \leq 1-\beta$; the one-sided requirement that the change must be decreasing is not enforced in the non-asymptotic analysis of Liu et al. (2019). The form of the critical testing radius (3.16) suggests that a nontrivial interaction may arise between a and β within the detection boundary when $\log\log n$ grows polynomially in p, as discussed in Section 3.1.1. However, in the $s=1$ case, two competing terms in the non-asymptotic rate are $\log p$ and $\log\log n$, suggesting an asymptotic regime in which $\log\log n$ does not need to grow as fast as polynomial in p, in order for a nontrivial interaction to arise.

Another asymptotic regime. To balance the two terms in the sum $\log\left(\frac{ep\log\log n}{s^2}\right) + \frac{\log\log n}{s}$, a natural calibration to consider is,

$$\begin{aligned} \log\log n &\sim a\log p \qquad a > 0 \\ s &\sim p^{1-\beta} \qquad \frac{1}{2} < \beta \leq 1. \end{aligned}$$

Indeed, when s is a constant, the two terms are both of order $\log p$. The corresponding detection boundary is $\rho^*_2(a,\beta) := \sqrt{2r^*_2(a,\beta)\log p}$, where

$$r^*_2(a,\beta) := \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1-\sqrt{1-\beta})^2 & \frac{3}{4} < \beta < 1 \\ 1+a & \beta = 1. \end{cases}$$

We only state the result for the one sided version of the problem, since there is no difference in the one vs two sided detection boundaries for the sparse regime $\beta > \frac{1}{2}$. The squared critical radius for the two sided problem in the case $\beta \leq \frac{1}{2}$ is equal to $p^{\beta-\frac{1}{2}}$, which is directly implied by plugging $s = p^{1-\beta}$ and $\log\log n \asymp \log p$ into the non-asymptotic rate (3.16). Our analysis of the exact constant gives more information than the non-asymptotic rate (3.16) only in the case $\beta > \frac{1}{2}$. For this reason, we only state the portion of the detection boundary within the sparse regime $\beta > \frac{1}{2}$, below in Theorem 3.3.2. In the formula $r^*_2(a,\beta)$, the only case in which the boundary departs from the Ingster–Donoho–Jin boundary is when $\beta = 1$, corresponding to the case in which there are a constant number of non-null rows in θ. Note that taking $a \to 0$ in the expression for $r^*_2(a,\beta)$ recovers the Ingster–Donoho–Jin boundary, as this limit corresponds to removing the difficulty of an unknown changepoint location from the testing problem.

Theorem 3.3.2. *Put $\rho \coloneqq \sqrt{2r \log p}$ for $r > 0$ and $\beta \in (1/2, 1]$, and let $a > 0$.*

1. *(Lower Bound) If $r < r_2^*(a, \beta)$, then $\mathcal{R}_{\textit{1-side}}(p, n, \rho, s) \not\to 0$, as $p \to \infty$.*
2. *(Upper Bound) If $r > r_2^*(a, \beta)$, then $\mathcal{R}_{\textit{1-side}}(p, n, \rho, s) \to 0$, as $p \to \infty$.*

Note that the detection boundary $\rho_2^*(a, \beta)$ does not depend continuously on a as $\beta \to 1$. This non-intuitive result can be explained by the non-asymptotic rate (3.16). Indeed, when $\beta < 1$, we have $\log \log n = o(s)$, so there is no contribution from the additive $\frac{\log \log n}{s} \sim \frac{a \log p}{p^{1-\beta}}$ term in the limit as $p \to \infty$, and thus a has no contribution to the formula for the detection boundary. In contrast, when $\beta = 1$, this term becomes $\frac{\log \log n}{s} \sim a \log p$, which is non-negligible compared to the competing term $\log\left(\frac{ep \log \log n}{s^2}\right) \sim \log p$. We note here that taking $\beta = s = 1$ gives

$$\rho_2^*(a, 1)^2 = 2(1 + a) \log p \sim 2(\log p + \log \log n),$$

giving the minimax testing radius for the $s = 1$ high dimensional changepoint detection problem, with sharp constant 2. Taking $a \to \infty$ when $\beta = s = 1$ gives

$$\rho_2^*(a, 1)^2 = 2(1 + a) \log p \sim 2(\log p + \log \log n) = 2(1 + o(1)) \log \log n,$$

for fixed n and p. Hence the sharp constant in the $p = s = 1$ changepoint detection problem, as obtained in Verzelen et al. (2020), is consistent with Theorem 3.3.2. The proof of Theorem 3.3.2 is contained in Sections 3.5.5 and 3.5.6.

3.4 Overview of Calculations

To simplify the presentation while maintaining the core ideas, we only outline the upper and lower bound proofs for the one sided changepoint problem in the asymptotic calibration (3.8). All rigorous proofs are deferred to Section 3.5.

3.4.1 Upper bound

To derive the upper bound result of Theorem 3.2.1, it suffices to provide a testing procedure with Type I and II error tending to zero as $p \to \infty$. If the location $t^* \in [n-1]$ of the changepoint is known, one could compute a contrast for each row $j \in [p]$,

$$Y_{jt^*} \coloneqq \sqrt{\frac{t^*(n - t^*)}{n}} \left(\frac{1}{t^*} \sum_{t=1}^{t^*} X_{jt} - \frac{1}{n - t^*} \sum_{t=t^*+1}^{n} X_{jt} \right). \tag{3.18}$$

Since $Y_{jt} \sim N(0, 1)$ under H_0 for any $t \in [n-1]$, one could then count the number of contrasts exceeding a null quantile $\bar{\Phi}^{-1}(q)$ for some $q \in (0, 1)$, and compare the fraction of large signals to the expected fraction q under the null.

$$\bar{S}_{p,t^*}(q) \coloneqq \frac{1}{p}\sum_{j=1}^{p} \mathbf{1}_{\{Y_{jt^*} > \bar{\Phi}^{-1}(q)\}}$$
$$\mathbb{E}_0 \bar{S}_{p,t}(q) = q \text{ for any } t \in [n-1].$$

One way to compare these quantities is to compute $K(\bar{S}_{p,t^*}(q), q)$, where the function $K : (0,1)^2 \to \mathbb{R}$ is defined,

$$K(x,t) \coloneqq x \log \frac{x}{t} + (1-x)\log\frac{1-x}{1-t},$$

which is the KL divergence between two Bernoulli distributions with success parameters x and t, i.e. $K(x,t) = \mathrm{KL}(\mathrm{Bern}(x), \mathrm{Bern}(t))$. Hence, the statistic $K(\bar{S}_{p,t^*}(q), q)$ is large when the observed proportion differs greatly the expected proportion under the null. Different choices of q may result in larger values of $K(\bar{S}_{p,t^*}(q), q)$ for various settings of the parameters (a, β), whose exact setting is assumed unknown when testing between H_0 and H_1 having only observed the matrix $X \in \mathbb{R}^{p\times n}$. For this reason, consider taking the supremum of the quantity $K(\bar{S}_{p,t^*}(q), q)$ over all $q \in (0,1)$. This supremum is achieved at some maximizing p value, reducing this supremum to a finite maximum (Berk and Jones; 1979),

$$\sup_{q\in(0,1)} K(\bar{S}_{p,t^*}(q), q) = \max_{j\le p} K(j/p, p_{(j)}), \tag{3.19}$$

where $p_{(1)} < p_{(2)} < \cdots < p_{(p)}$ are the ordered p values $p_j \coloneqq \bar{\Phi}(Y_{jt^*})$. Since the true changepoint location t^* is typically not known a priori, a first thought is to take the maximum of (3.19) over all possible locations $t^* \in [n-1]$. However, the elevated regions corresponding to the changepoint locations t^* and t^*+1 are almost completely overlapping. For each j, the $(Y_{jt^*})_{t^*\in[n-1]}$ are highly dependent over $t^* \in [n-1]$, leading to frequent type I errors. A similar issue arises in the non-asymptotic analysis of the problem, where Liu et al. (2019) overcome this obstacle by taking a restricted maximum over a grid of geometrically growing candidate changepoint locations,

$$\{1, 2, 4, \ldots, 2^{\lfloor \log_2(n/2)\rfloor}\} \subset [n-1].$$

We adopt a similar approach in the current setting. The above grid is sufficient to obtain the optimal rate in the critical radius, but it does not give the sharp constant. In order to obtain the sharp constant, we use a dense and symmetric version of the above grid in order to ensure that for any true changepoint location $t^* \in [n-1]$, there exists an element of the grid close to t^*. The grid is defined as

$$\mathcal{T} \coloneqq \left\{ \lfloor (1+\delta)^0 \rfloor, \lfloor (1+\delta)^1 \rfloor, \ldots, \lfloor (1+\delta)^{\log_{1+\delta}\frac{n}{2}} \rfloor, \lfloor n - (1+\delta)^{\log_{1+\delta}\frac{n}{2}} \rfloor, \ldots, \lfloor n - (1+\delta)^0 \rfloor \right\},$$

where $\delta = \frac{1}{\log\log n} \to 0$. The final test statistic is a penalized maximum over this grid,

$$\mathsf{PBJ}_p := \left[\max_{t \in \mathcal{T}} \sup_{q \in (0,1)} pK(\bar{S}_{p,t}(q), q)\right] - 2\log|\mathcal{T}|. \tag{3.20}$$

The idea of subtracting a penalization term $2\log|\mathcal{T}|$ to offset the magnitude of the maximum scanning statistic comes from the procedure developed in Chan et al. (2015), where it serves a similar purpose. A union bound is used to show the Type I error control,

$$\begin{aligned}\mathbb{P}(\mathsf{PBJ}_p > 2(2+\gamma)\log p) &= \mathbb{P}\left(\max_{t \in \mathcal{T}, q \in (0,1)} pK(\bar{S}_{p,t}(q), q) > 2(2+\gamma)\log p + 2\log|\mathcal{T}|\right) \\ &\leq \sum_{t \in \mathcal{T}} \mathbb{P}\left(\max_{q \in (0,1)} pK(\bar{S}_{p,t}(q), q) - 2(2+\gamma)\log p > 2\log|\mathcal{T}|\right).\end{aligned}$$

The concentration result in Lemma 3.6.5, along with (3.19), allows one to show that each summand on the right hand side above is $o(|\mathcal{T}|^{-1})$, causing the bound to go to zero. Besides the Type I error, the penalty $2\log|\mathcal{T}|$ does not contribute crucially to the proof of Type II error control. We note that $\log|\mathcal{T}| = p^{a(1+o(1))}$ in the regime (3.8) whereas $\log|\mathcal{T}| = a(1+o(1))\log p$ in the regime (3.9), and that the test statistic (3.20) is used in the proofs of feasibility throughout the detectable region in both of these asymptotic settings.

3.4.2 Lower bound

A geometrically growing grid (similar to $\mathcal{T}$ from the upper bound) plays a key role in the lower bound construction. The true changepoint is first selected uniformly at random from this grid, and conditional on this selection, the rows of X are distributed iid from a mixture distribution. Note that the cardinality of the grid $\mathcal{T}$ from the proof of the upper bound behaves as,

$$|\mathcal{T}| \asymp \log_{1+\delta} n = \frac{\log n}{\log(1+\delta)} \asymp \frac{\log n}{(\log\log n)^{-1}},$$

since $\log(1+x) \asymp x$ as $x \to 0$. Thus, the dominant behavior is determined by the factor $\log n \asymp e^{p^{a(1+o(1))}}$. This exponentially large quantity turns out to be the correct order of the number of candidate changepoint locations in the grid for the lower bound construction. However, computing a contrast as in (3.18) for $t^* \in \mathcal{T}$ results in a collection of dependent variables (Y_{jt^*}). To establish a limiting distribution for a sum of dependent terms involving the variables (Y_{jt^*}), it suffices to dampen the dependence between these variables by further spacing out the candidate changepoint locations. In order to explain the failure of all testing procedures in the undetectable region, this dampening needs to be done while maintaining the overall exponential order of $e^{p^{a(1+o(1))}}$ many candidate changepoint locations. This is accomplished by using the grid,

$$\mathcal{T}' \coloneqq \left\{ \lfloor (\log n)^0 \rfloor, \lfloor (\log n)^1 \rfloor, \ldots, \lfloor (\log n)^{\log_{\log n}(n-1)} \rfloor \right\}. \tag{3.21}$$

Ignoring the symmetry of the grid $\mathcal{T}$, both of the grids $\mathcal{T}$ and $\mathcal{T}'$ are (roughly) of the form $\{b^0, b^1, \ldots, b^{\log_b n}\}$ with bases $b = 1+\delta$ and $b = \log n$ corresponding to $\mathcal{T}$ and $\mathcal{T}'$ respectively. Note that the cardinality of $\mathcal{T}'$ behaves as

$$|\mathcal{T}'| \asymp \log_{\log n} n = \frac{\log n}{\log\log n},$$

meaning that the dominant behavior of $|\mathcal{T}'|$ is determined by the factor $\log n \asymp e^{p^{a(1+o(1))}}$. The worst case testing error in (3.13) is lower bounded by the Bayes testing error of the two point problem,

$$H_0 : (X_{j:})_{j\in[p]} \sim \prod_{j=1}^{p} N(0, I_n) \tag{3.22}$$

$$H_1 : (X_{j:})_{j\in[p]} \sim \frac{1}{|\mathcal{T}'|} \sum_{k=1}^{|\mathcal{T}'|} \prod_{j=1}^{p} \left[(1-\varepsilon)N(0, I_n) + \varepsilon N(\rho\theta^{(k)}, I_n)\right], \tag{3.23}$$

where $\theta^{(k)} \in \mathbb{R}^n$ is the unique vector associated to the contrast (3.18). That is, $\langle X_{j:}, \theta^{(k)}\rangle = Y_{jt_k}$, where $t_k \coloneqq \lfloor (\log n)^k \rfloor \in \mathcal{T}'$ is the k^{th} grid element, and $\varepsilon = p^{-\beta}$ is the average fraction of non-null rows[1].

Remark 3.4.1. *A general framework for deriving a detection boundary in an iid testing problem of the form,*

$$H_0 : Y_i \overset{\text{iid}}{\sim} Q_n \quad \textit{versus} \quad H_1 : Y_i \overset{\text{iid}}{\sim} (1-\varepsilon_n)Q_n + \varepsilon_n G_n, \qquad (i = 1, \ldots, n)$$

where Q_n and G_n are distributions on $\mathbb{R}$, can be found in the chapter by Cai and Wu (2014). We note that the high dimensional changepoint problem considered here falls outside of this setting. Indeed, in the Bayesian two-point formulation (3.22) versus (3.23), each row $X_{j:}$ in the observed matrix $X \in \mathbb{R}^{p\times n}$ is not a scalar random variable, and the rows $(X_{j:})_{j\le p}$ are not independent. In this fomulation, the true changepoint location $k^* \in \mathcal{T}'$ is drawn uniformly from a set of candidate locations. Marginalizing over the uniformly drawn changepoint location yields a distribution on X for which the rows $(X_{j:})_{j\le p}$ are dependent.

An optimal procedure for testing (3.22) versus (3.23) is one that thresholds the likelihood

[1] In order for the realization of θ drawn in H_1 to be supported on $\Theta_1^{\text{1-side}}(p, n, \rho, s)$ with high probability, ε should technically be replaced with $\bar{\varepsilon} = p^{-\bar{\beta}}$ for some slightly smaller $\bar{\beta} < \beta$, where ‘slightly smaller’ is described more precisely in Section 3.5.1.

ratio,

$$\psi_{\mathsf{LRT}}(X) = \mathbf{1}\left\{\frac{f_1}{f_0}(X) > a_p\right\},$$

using a sequence (a_p) of rejection thresholds. Thus, the detection boundary can be derived by studying the asymptotic behavior of $\frac{P_1}{P_0}(X)$, where f_0 and f_1 are the densities associated with the distributions in H_0 and H_1 respectively. For most cases in the argument, this involves computing the second moment of the likelihood ratio, also known as the chi square divergence between f_0 and f_1, and choosing the signal size ρ so small that this moment is no larger than the typical size of $\frac{f_1}{f_0}(X)$ under the H_0. In this setting, the test ψ_{LRT} would be powerless in detecting a deviation from the null, implying the failure of an optimal test and thus the hardness of the testing problem.

In the case $1 - 4\beta/3 < a \leq 1 - \beta$, a separate argument involving truncation is used. The sum of type 1 and 2 errors for the likelihood ratio test is lower bounded, up to constants, by the probability of an event that does not depend on the original rejection threshold a_p, summarized in Lemma 3.6.2. Due to the mixture form of H_1, the likelihood ratio is an average of products, one for each candidate changepoint location,

$$\frac{f_1}{f_0}(X) = \frac{1}{|\mathcal{T}'|}\sum_{k=1}^{|\mathcal{T}'|} L_k(X),$$

where each $L_k(X)$ is defined

$$L_k(X) = \frac{\prod_{j=1}^{p}\left[(1-\varepsilon)\phi_n(X_{j:}) + \varepsilon\phi_n(X_{j:} - \rho\theta^{(k)})\right]}{\prod_{j=1}^{p}\phi_n(X_{j:})} = \prod_{j=1}^{p}\left[1 + \varepsilon\left(e^{\rho Y_{jt_k} - \rho^2/2} - 1\right)\right],$$

and ϕ_n is the density of a standard normal vector in $\mathbb{R}^n$. Under H_1, the dominant term in the average is L_{k^*}, whose log is a sum of terms,

$$\log L_{k^*}(X) = \sum_{j=1}^{p}\log\left[1 + \varepsilon\left(e^{\rho Y_{jt_{k^*}} - \rho^2/2} - 1\right)\right].$$

The terms in the sum are analyzed separately, depending on the size of each $Y_{jt_{k^*}}$, and the details of this calculation can be found in Case 3 of Section 3.5.1.

Connection to submatrix detection in Butucea and Ingster (2013)

The proof of the lower bound in Theorems 3.2.1 and 3.2.2 relies on the hardness of distinguishing between the hypotheses (3.22) and (3.23) for small enough ρ. As noted in the previous section, the likelihood ratio for this problem is of the form

$$\frac{f_1}{f_0}(X) = \frac{1}{|\mathcal{T}'|}\sum_{k=1}^{|\mathcal{T}'|}\prod_{j=1}^{p}\left[1+\varepsilon\left(e^{\rho Y_{jt_k}-\rho^2/2}-1\right)\right],$$

meaning that it only relies on the matrix X via the contrasts

$$Y_{jt_k} = \langle X_{j:}, \theta^{(k)}\rangle = \sqrt{\frac{\lfloor b^k\rfloor(n-\lfloor b^k\rfloor)}{n}}\left(\frac{1}{\lfloor b^k\rfloor}\sum_{t=1}^{\lfloor b^k\rfloor} X_{jt} - \frac{1}{n-\lfloor b^k\rfloor}\sum_{t=n-\lfloor b^k\rfloor}^{n} X_{jt}\right)$$

where $j = 1,\dots,p$ and $k = 1,\dots,|\mathcal{T}'|$, and the second equality holds by definition of $\theta^{(k)}$ (see part 1 of Lemma 3.7.1). In other words, it is sufficient to observe these contrasts in order to perform the optimal test, ψ_{LRT}, for testing (3.22) against (3.23), assuming oracle knowledge of ε and ρ. Since the true changepoint location is drawn uniformly at random from $\mathcal{T}'$, the contrasts can be written,

$$Y_{jt_k} = \langle X_{j:}, \theta^{(k)}\rangle = \langle \rho\theta^{(k^*)} + Z_{j:}, \theta^{(k)}\rangle = \rho\langle\theta^{(k^*)}, \theta^{(k)}\rangle + \langle Z_{j:}, \theta^{(k)}\rangle.$$

Letting $k^* \sim \mathrm{Unif}\{1,\dots,|\mathcal{T}'|\}$ denote the index of the true changepoint location, parts 2 and 3 of Lemma 3.7.1 together with the previous display imply

$$Y_{jt_k} \overset{(d)}{=} \begin{cases} \rho + Z & \text{if } k^* = k \\ o(1) + Z & \text{if } k^* \neq k, \end{cases}$$

since the grid $\mathcal{T}'$ has base $b = \log n = e^{p^{a(1+o(1))}}$ in the current regime (3.8). Further, the covariance between two contrasts Y_{jt_k} and Y_{jt_ℓ} corresponding to row j is

$$\mathrm{Cov}(Y_{jt_k}, Y_{jt_\ell}) = \langle\theta^{(k)}, \theta^{(\ell)}\rangle = \begin{cases} 1 & \text{if } k = \ell \\ O(b^{-\frac{|k-\ell|}{2}}) & \text{if } k \neq \ell, \end{cases}$$

by Part 3 of Lemma 3.7.1. In other words, the (Y_{jt_k}) are nearly independent Gaussians, and under H_1, there exists some $k^* \leq |\mathcal{T}'|$ for which $(Y_{jt_{k^*}})_{j\in[p]}$ have elevated mean ρ, and all other $(Y_{jt_k})_{j\in[p],k\neq k^*}$ are Gaussian variables that have unit variance and nearly zero mean.

Consider the matrix $Y \in \mathbb{R}^{p\times|\mathcal{T}'|}$ whose (j,k) entry is Y_{jt_k}. By the previous observation, the (Y_{jt_k}) are approximately independent, and the testing problem (3.22) versus (3.23) is essentially testing between,

$$H_0 : (Y_{j:})_{j\in[p]} \sim \prod_{j=1}^{p} N(0, I_{|\mathcal{T}'|})$$
$$H_1 : (Y_{j:})_{j\in[p]} \sim \frac{1}{|\mathcal{T}'|}\sum_{k=1}^{|\mathcal{T}'|}\prod_{j=1}^{p}\left[(1-\varepsilon)N(0, I_{|\mathcal{T}'|}) + \varepsilon N(\rho e_k, I_{|\mathcal{T}'|})\right],$$

where $e_k \in \mathbb{R}^{|\mathcal{T}'|}$ is the k^{th} standard basis vector. Let $I_j \in \{0,1\}$ indicates whether or not row j in Y has non-zero mean ρe_k, i.e. $I_j = 1 \iff$ row j is non-null. Under H_1, there is some rectangle

$$\{j \le p : I_j = 1\} \times \{k\} \subset [p] \times [|\mathcal{T}'|], \tag{3.24}$$

over which the entries Y_{jt_k} have elevated mean $\rho > 0$. This problem is a special case of the submatrix detection problem studied by Butucea and Ingster (2013), in which the entire observation is a matrix $Y \in \mathbb{R}^{N\times M}$ of unit variance Gaussian entries. Under the null, these entries have zero mean. Under the alternative, a submatrix $(Y_{ij})_{i\in A, j\in B}$ has elevated mean $s_{ij} \ge \rho$ for $(i,j) \in A\times B$ and zero mean $s_{ij} = 0$ for $(i,j) \notin A \times B \subset [N]\times[M]$. A and B are constrained to be of sizes $n \le N$ and $m \le M$ respectively. In this notation, the testing problem can be written,

$$H_0 : s_{ij} = 0 \text{ for all } i = 1,\dots,N, \text{ and } j = 1,\dots,M$$
$$H_1 : \exists A\times B \in \mathcal{C}_{nm} \text{ such that } s_{ij} = 0 \text{ if } (i,j) \notin A\times B, \text{ and } s_{ij} \ge \rho \text{ if } (i,j) \in A\times B,$$

where $\mathcal{C}_{nm}$ is the collection of rectangles $A \times B \subset [N]\times[M]$ satisfying $|A| = n$ and $|B| = m$. In their chapter, Butucea and Ingster have established both the rate and constant for the minimax separation ρ, in a specific asymptotic regime which requires,

$$n\log(N/n) \asymp m\log(M/m). \tag{3.25}$$

Their main result is that, under some asymptotic requirements restricting the relative growth of m and n, which include the condition (3.25), the minimax separation ρ satisfies,

$$\rho^2 = \frac{2(n\log(N/n) + m\log(M/m))}{nm}(1+o(1)). \tag{3.26}$$

In our setting, the rectangles (3.24) and matrix of contrasts $(Y_{jk}) \coloneqq (Y_{jt_k})$ have dimensions $(n, m, N, M) \coloneqq (p^{1-\beta}, 1, p, e^{p^{a(1+o(1))}})$. Then condition (3.25) roughly corresponds to $1-\beta = a$, but in general does not cover the range of parameters (a, β) considered in our chapter.

The result (3.26) then reduces to

$$\rho^2 = \frac{2(p^{1-\beta}\beta \log p + p^a)}{p^{1-\beta}} = 2\beta(1 + o(1)) \log p,$$

which coincides with the detection boundary in the case $a = 1 - \beta$,

$$\rho^*_{\text{1-side}}(1 - \beta, \beta) = 2\beta \log p.$$

3.5 Proofs

Recall that $\Theta_1^{\text{1-side}}(p,n,\rho,s) \subset \Theta_1^{\text{2-side}}(p,n,\rho,s)$ implies the two sided changepoint detection problem is at least as difficult as the one sided version. Note that the detection boundaries only differ when $a \leq 1-2\beta$. Then for the lower bounds, it suffices to provide the lower bound proof in the one sided problem for all cases, and provide the proof for the case $a \leq 1-2\beta$ in the two sided problem. Similarly, for the upper bounds, it suffices to provide the upper bound proof in the two sided problem in all cases, and the proof for the case $a \leq 1-2\beta$ in the one sided problem. Throughout, we let Z denote a standard normal variable, whose dimensions will be clear from the context.

3.5.1 Lower Bound for Theorem 3.2.1

The formula for the one sided boundary is,

$$\rho^*_{\text{1-side}}(a,\beta)^2 := \begin{cases} p^{a-(1-2\beta)} & a \leq 1-2\beta \\ (a-(1-2\beta))\log p & 1-2\beta < a \leq 1-4\beta/3 \\ 2(\sqrt{1-a}-\sqrt{1-a-\beta})^2 \log p & 1-4\beta/3 < a \leq 1-\beta \\ p^{a-(1-\beta)} & a > 1-\beta. \end{cases}$$

Put $\rho := \rho^*_{\text{1-side}}(a,\beta_1)$ for $a > 0$ and $\beta_1 \in (0,1)$. We will show that if $\beta > \beta_1$, then $\mathcal{R}_{\text{1-side}}(p,\rho,a,\beta) \to 1$, as $p \to \infty$. Consider the testing problem,

$$\begin{aligned} &H_0 : X_{j:} \overset{\text{iid}}{\sim} N(0, I_n) \\ &H_1 : k^* \sim \text{unif}\{1,\dots,|\mathcal{T}|\}, X_{j:} \mid k^* = k \overset{\text{iid}}{\sim} (1-\varepsilon)N(0,I_n) + \varepsilon N(\rho\theta^{(k)}, I_n), \end{aligned}$$

where $\varepsilon := p^{-\beta}$, $\mathcal{T}$ is the base $b := \log n \asymp e^{p^{a(1+o(1))}}$ grid, defined,

$$\mathcal{T} := \{\lfloor b^1 \rfloor, \lfloor b^2 \rfloor, \dots, \lfloor b^{\log_{\log n} n} \rfloor\},$$

and $\theta^{(k)} \in \mathbb{R}^n$ is the unique vector for which

$$\langle Z, \theta^{(k)} \rangle = \sqrt{\frac{\lfloor b^k \rfloor (n - \lfloor b^k \rfloor)}{n}} (\bar{Z}_{1:\lfloor b^k \rfloor} - \bar{Z}_{\lfloor b^k \rfloor + 1:n}).$$

More explicitly,

$$\theta_t^{(k)} := \begin{cases} \left(\frac{n-\lfloor b^k \rfloor}{n\lfloor b^k \rfloor}\right)^{1/2} & \text{if } t \leq b^k \\ -\left(\frac{\lfloor b^k \rfloor}{n(n-\lfloor b^k \rfloor)}\right)^{1/2} & \text{if } t > b^k. \end{cases}$$

By Part 2 of Lemma 3.7.1, $\langle \theta^{(k)}, \rho\theta^{(k)} \rangle = \rho$. Consequently, by Part 1 of Lemma 3.7.1 and the definition of the signal in $\Theta_1^{\text{1-side}}(p, n, \rho, s)$, if the draw of θ described in H_1 has more than $s \sim p^{1-\beta}$ non-null rows, then $\theta \in \Theta_1^{\text{1-side}}(p, n, \rho, s)$. Note however that the number of non-null rows under H_1 is distributed Binomial(p, ε), which implies that the resulting (random) instance of $\theta \in \mathbb{R}^{p\times n}$ is not necessarily an element of $\Theta_1^{\text{1-side}}(p, n, \rho, s)$. To remedy this issue, note that since $\beta > \beta_1$, there exists $\bar{\beta} \in (\beta_1, \beta)$ for which $\bar{\varepsilon} := p^{-\bar{\beta}} > p^{-\beta}$. Consider the modified testing problem,

$$\begin{aligned}
&H_0 : X_{j:} \overset{\text{iid}}{\sim} N(0, I_n) \\
&\bar{H}_1 : k^* \sim \text{unif}\{1, \ldots, |\mathcal{T}|\}, X_{j:} \mid k^* = k \overset{\text{iid}}{\sim} (1-\bar{\varepsilon})N(0, I_n) + \bar{\varepsilon}N(\rho\theta^{(k)}, I_n), \qquad (3.27)
\end{aligned}$$

and define the event,

$$G := \left\{ \sum_{j=1}^{p} \mathbf{1}_{\{\text{row } j \text{ is non-null}\}} \geq \frac{p\bar{\varepsilon}}{\log p} \right\}.$$

Each row $X_{j:}$ is non-null with probability $\bar{\varepsilon} = p^{-\bar{\beta}} = \omega(p^{-\beta})$ in (3.27). Thus, on the set G, the realizations of θ belong to the parameter space $\Theta_1^{\text{1-side}}(p, n, \rho, s)$. By Chebyshev's inequality, it is seen that $\bar{\mathbb{P}}_1(G^c) \to 0$, where $\bar{\mathbb{P}}_1$ denotes the distribution under $\bar{H}_1$. Then the minimax testing error can be lower bounded,

$$\begin{aligned}
\mathcal{R}_{\text{1-side}}(p, n, \rho, s) &:= \inf_{\psi} \left[\sup_{\theta \in \Theta_0(p,n)} \mathbb{P}_\theta \psi + \sup_{\theta \in \Theta_1^{\text{1-side}}(p,n,\rho,s)} \mathbb{P}_\theta(1-\psi) \right] \\
&\geq \inf_{\psi} \left[\mathbb{P}_0 \psi + \bar{\mathbb{P}}_1 (1-\psi) \mathbf{1}_G \right] \\
&\geq \inf_{\psi} \left[\mathbb{P}_0 \psi + \bar{\mathbb{P}}_1 (1-\psi) \right] - \bar{\mathbb{P}}_1(G^c) \qquad (3.28) \\
&= 1 - \mathsf{TV}(\mathbb{P}_0, \bar{\mathbb{P}}_1) - \bar{\mathbb{P}}_1(G^c). \qquad \text{(Neyman–Pearson)}
\end{aligned}$$

Then since $\bar{\mathbb{P}}_1(G^c) \to 0$, it suffices to show that $\liminf_{p\to\infty}(1 - \mathsf{TV}(\mathbb{P}_0, \bar{\mathbb{P}}_1)) \geq c$, for any $c \in (0, 1)$. To this end, the testing problem can equivalently be written,

$$\begin{aligned}
H_0 &: (X_{j:})_{j\in[p]} \sim \prod_{j=1}^{p} N(0, I_n) \\
\bar{H}_1 &: (X_{j:})_{j\in[p]} \sim \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \prod_{j=1}^{p} \left[(1-\bar{\varepsilon})N(0, I_n) + \bar{\varepsilon}N(\rho\theta^{(k)}, I_n) \right].
\end{aligned}$$

Let f_0 and f_1 be the densities of X corresponding to H_0 and $\bar{H}_1$. Note that,

$$1 - \mathsf{TV}(\mathbb{P}_0, \bar{\mathbb{P}}_1) = \int f_0 \wedge f_1 \geq \int_{\{f_1 \geq c f_0\}} f_0 \wedge f_1 \geq c \mathbb{P}_0 \left(\frac{f_1}{f_0} \geq c \right). \qquad \text{(for any } c \in (0,1))$$

Then for the rhs to tend to c, it suffices to show that

$$\frac{f_1}{f_0}(X) \overset{\mathbb{P}_0}{\to} 1.$$

The above convergence is implied by the condition, $\limsup_{p\to\infty} \mathbb{E}_0(\frac{f_1}{f_0}(X))^2 \leq 1$, since by Markov's inequality,

$$\mathbb{P}_0 \left(\left| \frac{f_1}{f_0}(X) - 1 \right| > \delta \right) \leq \frac{\mathbb{E}_0(\frac{f_1}{f_0}(X))^2 - 2\mathbb{E}_0(\frac{f_1}{f_0}(X)) + 1}{\delta^2} = \frac{\mathbb{E}_0(\frac{f_1}{f_0}(X))^2 - 1}{\delta^2}.$$

Cases 1 and 2: $a \leq 1 - 4\beta_1/3$

Letting $\bar{\mathbb{E}}_1$ denote the expectation with respect to $\bar{H}_1$, the second moment of the likelihood ratio is

$$\begin{aligned}
\mathbb{E}_0 \left(\frac{f_1}{f_0}(X) \right)^2 &= \bar{\mathbb{E}}_1 \frac{f_1}{f_0}(X) \\
&= \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \bar{\mathbb{E}}_1 \prod_{j=1}^{p} \left[1 - \bar{\varepsilon} + \bar{\varepsilon} \exp\left(\rho \langle X_{j:}, \theta^{(k)} \rangle - \rho^2/2 \right) \right] \\
&= \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \mathbb{E}_{l \sim \text{Unif}(\{1,\ldots,|\mathcal{T}|\})} \prod_{j=1}^{p} \left[1 - \bar{\varepsilon} + \bar{\varepsilon} \cdot \bar{\mathbb{E}}_1 (\exp(\rho \langle X_{j:}, \theta^{(k)} \rangle - \rho^2/2 \mid k^* = l) \right] \\
&= \frac{1}{|\mathcal{T}|^2} \sum_{k,l \leq |\mathcal{T}|} \prod_{j=1}^{p} \left[1 - \bar{\varepsilon} + \bar{\varepsilon} \cdot \mathbb{E}(\exp(\rho \langle \rho \theta^{(l)} I_j + Z_{j:}, \theta^{(k)} \rangle - \rho^2/2)) \right],
\end{aligned}$$

where $I_j = 1$ indicates that row j has a changepoint, and $I_j = 0$ otherwise ($I_j \overset{\text{iid}}{\sim} \text{Bern}(\bar{\varepsilon})$). Now,

$$\begin{aligned}
\mathbb{E}(\exp(\rho \langle \rho \theta^{(l)} I_j + Z_{j:}, \theta^{(k)} \rangle - \rho^2/2) &= (1 - \bar{\varepsilon}) \mathbb{E} e^{\rho Z - \rho^2/2} + \bar{\varepsilon} \cdot \mathbb{E} e^{\rho(\rho \langle \theta^{(l)}, \theta^{(k)} \rangle + Z) - \rho^2/2} \\
&= 1 + \bar{\varepsilon}(\exp(\rho^2 \langle \theta^{(l)}, \theta^{(k)} \rangle) - 1) \\
&\leq 1 + \bar{\varepsilon}(\exp(\rho^2 e^{-\frac{|k-l|}{2} p^a}) - 1),
\end{aligned}$$

where the last inequality follows by part 3 of Lemma 3.7.1, $Z_{j:} \in \mathbb{R}^n$ denotes a standard Gaussian vector, and Z denotes a standard Gaussian variable. Plugging back into the second moment, we have

$$\mathbb{E}_0\left(\frac{f_1}{f_0}(X)\right)^2 \leq \frac{1}{|\mathcal{T}|^2}\sum_{k=l}\prod_{j=1}^{p}\left[1+\bar{\varepsilon}^2(e^{\rho^2}-1)\right] + \frac{1}{|\mathcal{T}|^2}\sum_{k\neq l}\prod_{j=1}^{p}\left[1+\bar{\varepsilon}^2(\exp(\rho^2 e^{-\frac{|k-l|}{2}p^a})-1)\right].$$

Since $k \neq l$ implies $\rho^2 e^{-\frac{|k-l|}{2}p^a} \leq \rho^2 e^{-\frac{1}{2}p^a} \to 0$ as $p \to \infty$, we have that when $k \neq l$,

$$\prod_{j=1}^{p}\left[1+\bar{\varepsilon}^2(\exp(\rho^2 e^{-\frac{|k-l|}{2}p^a})-1)\right] \leq \left[1+2\bar{\varepsilon}^2\rho^2 e^{-\frac{1}{2}p^a}\right]^p \leq \exp\left(2p\bar{\varepsilon}^2\rho^2 e^{-\frac{1}{2}p^a}\right),$$

where the first inequality follows from $e^x - 1 < 2x$ for $0 < x < 1$, and the second inequality follows from $\log(1+x) \leq x$ for $x \in \mathbb{R}$. Note that $2p\bar{\varepsilon}^2\rho^2 e^{-\frac{1}{2}p^a} \to 0$ as $p \to \infty$, so the right hand side of the above tends to 1. Plugging the above back into the estimate for the second moment, we obtain

$$\begin{aligned}\mathbb{E}_0\left(\frac{f_1}{f_0}\right)^2 &\leq \frac{1}{|\mathcal{T}|^2}\sum_{k=l}\prod_{j=1}^{p}\left[1+\bar{\varepsilon}^2(e^{\rho^2}-1)\right] + \frac{1}{|\mathcal{T}|^2}\sum_{k\neq l}(1+o(1))\\ &= \frac{1}{|\mathcal{T}|}\exp(p\bar{\varepsilon}^2(e^{\rho^2}-1)) + 1 + o(1).\end{aligned}$$

Recall $|\mathcal{T}| \asymp e^{p^{a(1+o(1))}}$. In order for the right hand side to tend to 1, it suffices for the first term to tend to zero, i.e. $\exp(p\bar{\varepsilon}^2(e^{\rho^2}-1) - p^{a(1+o(1))}) \to 0$. Since $\rho^2 = \rho^*_{\text{1-side}}(a,\beta_1)^2 \sim \log(1+p^{a-(1-2\beta_1)})$ when $a \neq 1-2\beta_1$, we have

$$\begin{aligned}\exp\left(p\bar{\varepsilon}^2(e^{\rho^2}-1) - p^{a(1+o(1))}\right) &= \exp\left(p\bar{\varepsilon}^2 p^{(a-(1-2\beta_1))(1+o(1))} - p^{a(1+o(1))}\right)\\ &= \exp\left(p^{(a+2(\beta_1-\bar{\beta}))(1+o(1))} - p^{a(1+o(1))}\right)\\ &\to 0. \qquad (\bar{\beta} > \beta_1)\end{aligned}$$

When $a = 1-2\beta_1$, we have $\rho^2 = 1$, so that

$$\begin{aligned}\exp\left(p\bar{\varepsilon}^2(e^{\rho^2}-1) - p^{a(1+o(1))}\right) &= \exp\left(p\bar{\varepsilon}^2(e-1) - p^{a(1+o(1))}\right)\\ &= \exp\left(p^{(1-2\bar{\beta})(1+o(1))} - p^{(1-2\beta_1)(1+o(1))}\right)\\ &\to 0. \qquad (\bar{\beta} > \beta_1)\end{aligned}$$

Case 3: $1 - 4\beta_1/3 < a \le 1 - \beta_1$

Put $x \coloneqq \sqrt{1-a}$ and $y \coloneqq \sqrt{1-a-\beta_1}$. Then the signal ρ is of the form, $\rho = (x-y)\sqrt{2\log p}$. The likelihood ratio test for (3.27) is defined by

$$\psi_{\mathsf{LRT}}(X) \coloneqq \mathbf{1}\left\{\frac{f_1}{f_0}(X) > a_p\right\}$$

for some rejection level $a_p \in \mathbb{R}$. Then by (3.28), the minimax testing risk for the original testing problem is lower bounded by the testing risk of ψ_{LRT},

$$\mathcal{R}_{\text{1-side}}(p,n,\rho,s) \ge \inf_{\psi}[\mathbb{P}_0\psi + \bar{\mathbb{P}}_1(1-\psi)] - o(1) = \mathbb{P}_0\psi_{\mathsf{LRT}} + \bar{\mathbb{P}}_1(1-\psi_{\mathsf{LRT}}) - o(1).$$

The likelihood ratio is of the form,

$$\frac{f_1}{f_0}(X) = \frac{1}{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|} L_k$$
$$L_k \coloneqq \prod_{j=1}^{p}\left[1 - \bar{\varepsilon} + \bar{\varepsilon}e^{\rho\langle X_{j:}, \theta^{(k)}\rangle - \rho^2/2}\right].$$

By Lemma 3.6.2, the sum of errors from the likelihood ratio test can be lower bounded for any a_p,

$$\mathbb{P}_0(\psi_{\mathsf{LRT}} = 1) + \bar{\mathbb{P}}_1(\psi_{\mathsf{LRT}} = 0) \ge \frac{1}{3}\bar{\mathbb{P}}_1(A_3),$$

where $A_3 \coloneqq \left\{\frac{f_1}{f_0}(X) \le 3\right\}$. Now note

$$\bar{\mathbb{P}}_1(A_3) \ge \bar{\mathbb{P}}_1\left(\frac{1}{|\mathcal{T}|}\sum_{k\neq k^*} L_k \le 2, L_{k^*} \le |\mathcal{T}|\right).$$

Now by Lemma 3.6.3,

$$\bar{\mathbb{E}}_1\left(\frac{1}{|\mathcal{T}|}\sum_{k\neq k^*} L_k\right) = \mathbb{E}_1\left(\frac{1}{|\mathcal{T}|}\sum_{k\neq k^*} \mathbb{E}_1(L_k \mid k^*)\right) = \frac{1}{|\mathcal{T}|}\cdot(|\mathcal{T}|-1)(1 + O(e^{-\frac{1}{2}p^a(1+o(1))})),$$

which is $\le 1 + o(1)$ for $a > 0$ as $p \to \infty$. It now follows from Markov's inequality that

$$\mathbb{P}_1\left(\frac{1}{|\mathcal{T}|}\sum_{k\neq k^*} L_k \le 2\right) \ge \frac{1}{2} - o(1).$$

Thus, to show that $\bar{\mathbb{P}}_1(A_3)$ is asymptotically no smaller than $1/2$, it suffices to show

$$\mathbb{P}_1(L_{k^*} > |\mathcal{T}|) \to 0. \tag{3.29}$$

To this end, split the indices $j \in [p]$ into four sets. Letting $Q_j = 1\{\text{row } j \text{ is non-null}\}$, put

$$\begin{aligned}
&\Gamma_0 := \{j : Q_j = 0\} \\
&\Gamma_1 := \{j : Q_j = 1, \langle X_{j:}, \theta^{(k^*)}\rangle \leq \sqrt{2(1-a)\log p}\} \\
&\Gamma_2 := \{j : Q_j = 1, \sqrt{2(1-a)\log p} < \langle X_{j:}, \theta^{(k^*)}\rangle \leq 2\sqrt{2\log p}\} \\
&\Gamma_3 := \{j : Q_j = 1, \langle X_{j:}, \theta^{(k^*)}\rangle > 2\sqrt{2\log p}\}.
\end{aligned}$$

Then $\log L_{k^*} = \sum_{i=0}^{3} R_i$, where

$$R_i := \sum_{j \in \Gamma_i} \log\left(1 + \bar{\varepsilon}\left(\exp\left(\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2\right) - 1\right)\right), \tag{3.30}$$

for $i = 0, 1, 2, 3$, and the probability in condition (3.29) is bounded,

$$\bar{\mathbb{P}}_1(L_{k^*} > |\mathcal{T}|) = \bar{\mathbb{P}}_1\left(\sum_{i=0}^{3} R_i > \log|\mathcal{T}|\right) \leq \sum_{i=0}^{3} \bar{\mathbb{P}}_1\left(R_i > \frac{1}{4}\log|\mathcal{T}|\right).$$

The result now follows upon showing that

$$\bar{\mathbb{P}}_1\left(R_i > \frac{1}{4}\log|\mathcal{T}|\right) \to 0, \quad \text{for } i = 0, 1, 2, 3. \tag{3.31}$$

For $i = 0$, note that any j with $Q_j = 0$ has $\langle X_{j:}, \theta^{(k^*)}\rangle \sim N(0, 1)$, so that

$$\mathbb{E}_1 e^{R_0} = \mathbb{E}_1(\mathbb{E}_1(e^{R_0} \mid \Gamma_0)) = \mathbb{E}_1 \prod_{j \in \Gamma_0} \mathbb{E}(1 + \bar{\varepsilon}(e^{\rho Z - \rho^2/2} - 1)) = 1.$$

Thus

$$\bar{\mathbb{P}}_1\left(R_0 > \frac{1}{4}\log|\mathcal{T}|\right) = \bar{\mathbb{P}}_1\left(e^{R_0} > |\mathcal{T}|^{1/4}\right) \leq \frac{1}{|\mathcal{T}|^{1/4}} \to 0. \tag{3.32}$$

For $i = 1$, note that each summand in (3.30) is no smaller than $\log(1 + \bar{\varepsilon}(0 - 1)) = \log(1 - \bar{\varepsilon})$. Since there are $|\Gamma_1|$ many summands in R_1, Markov's inequality can be applied to the difference $R_1 - |\Gamma_1|\log(1 - \bar{\varepsilon}) > 0$,

$$\mathbb{P}\left(R_1 > \frac{1}{4}\log|\mathcal{T}|\right) \leq \mathbb{P}\left(R_1 - |\Gamma_1|\log(1 - \bar{\varepsilon}) > \frac{1}{4}\log|\mathcal{T}|\right) \leq \frac{4\bar{\mathbb{E}}_1(R_1 - |\Gamma_1|\log(1 - \bar{\varepsilon}))}{\log|\mathcal{T}|},$$

since $\log(1-\bar{\varepsilon}) < 0$. Since $\log(1-\bar{\varepsilon}) \asymp -\bar{\varepsilon}$ as $\bar{\varepsilon} \to 0$, and $\log|\mathcal{T}| \asymp p^{a(1+o(1))}$, for the rhs of the above to go to zero, it suffices to show

$$p^{-a(1+o(1))} \cdot \bar{\mathbb{E}}_1 R_1 \to 0 \tag{3.33}$$

$$p^{-a(1+o(1))-\bar{\beta}} \cdot \bar{\mathbb{E}}_1 |\Gamma_1| \to 0. \tag{3.34}$$

The first expectation is

$$\begin{aligned}
\bar{\mathbb{E}}_1 R_1 &= \bar{\mathbb{E}}_1 \sum_{j \in \Gamma_1} \log\left(1 + \bar{\varepsilon}\left(e^{\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2} - 1\right)\right) \\
&= \sum_{j=1}^{p} \bar{\mathbb{E}}_1 \left[1_{\{Q_j=1, \langle X_{j:}, \theta^{(k^*)}\rangle \leq \sqrt{2(1-a)\log p}\}} \log\left(1 + \bar{\varepsilon}\left(e^{\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2} - 1\right)\right)\right] \\
&\leq \sum_{j=1}^{p} \bar{\mathbb{E}}_1 \left[1_{\{Q_j=1, \langle X_{j:}, \theta^{(k^*)}\rangle \leq \sqrt{2(1-a)\log p}\}} \bar{\varepsilon} \cdot e^{\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2}\right] \\
&\leq \sum_{j=1}^{p} \bar{\varepsilon}^2 \cdot \mathbb{E}\left[1_{\{\rho + Z \leq \sqrt{2(1-a)\log p}\}} \cdot e^{\rho(\rho+Z) - \rho^2/2}\right], \qquad (Z \sim N(0,1))
\end{aligned}$$

since conditional on $Q_j = 1$, we have that $\langle X_{j:}, \theta^{(k^*)}\rangle \sim N(\rho, 1)$. Directly integrating, the above is equal to

$$\begin{aligned}
&= p\bar{\varepsilon}^2 e^{\rho^2/2} \int_{-\infty}^{\sqrt{2(1-a)\log p} - \rho} e^{\rho z} \phi(z) dz \\
&= p\bar{\varepsilon}^2 e^{\rho^2} \int_{-\infty}^{\sqrt{2(1-a)\log p} - \rho} \phi(z - \rho) dz \\
&= p\bar{\varepsilon}^2 e^{\rho^2} \Phi(\sqrt{2(1-a)\log p} - 2\rho). \qquad (3.35)
\end{aligned}$$

Recall that $\rho = \sqrt{2\log p} \cdot (\sqrt{1-a} - \sqrt{1-a-\beta_1}) =: \sqrt{2\log p} \cdot (x - y)$. The condition $a > 1 - 4\beta_1/3$ implies

$$\sqrt{2(1-a)\log p} - 2\rho = \sqrt{2\log p} \cdot (2y - x) = \sqrt{2\log p} \cdot (2\sqrt{1-a-\beta_1} - \sqrt{1-a}) < 0,$$

and so expression (3.35) becomes (up to log factors in p)

$$= p\bar{\varepsilon}^2 e^{\rho^2} \Phi(\sqrt{2\log p} \cdot (x - 2(x-y))) = p^{1-2\bar{\beta}+2(x-y)^2-(2y-x)^2} = p^{1-2\bar{\beta}-2y^2+x^2},$$

from which (3.33) follows, since

$$p^{-a(1+o(1))} \cdot \bar{\mathbb{E}}_1 R_1 = p^{-a(1+o(1))+1-2\bar{\beta}-2(1-a-\beta_1)+1-a} = p^{-2(\bar{\beta}-\beta_1)} \to 0.$$

Next, note that

$$\begin{aligned}\bar{\mathbb{E}}_1|\Gamma_1| &= \sum_{j=1}^{p} \bar{\mathbb{P}}_1(Q_j = 1, \langle X_{j:}, \theta^{(k^*)}\rangle \le \sqrt{2(1-a)\log p}) \\ &= p^{1-\bar{\beta}}\mathbb{P}(\rho + Z \le \sqrt{2(1-a)\log p}) && (Z \sim N(0,1)) \\ &\le p^{1-\bar{\beta}}.\end{aligned}$$

(3.34) now follows since $\bar{\beta} > \beta_1$ implies

$$p^{-a(1+o(1))-\bar{\beta}} \cdot \bar{\mathbb{E}}_1|\Gamma_1| \le p^{-a(1+o(1))+1-2\bar{\beta}} \le p^{-a(1+o(1))+1-2\beta_1} \to 0,$$

because $a > 1 - 4\beta_1/3 > 1 - 2\beta_1$. We have now shown (3.33) and (3.34), which imply condition (3.31) holds with $i = 1$.

Next we check condition (3.31) for $i = 2$. To this end, first note that each summand in R_2 is positive, because $\langle X_{j:}, \theta^{(k^*)}\rangle > \sqrt{2(1-a)\log p}$ implies

$$\begin{aligned}\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2 &> \rho\sqrt{2(1-a)\log p} - \rho^2/2 \\ &= \rho\sqrt{2\log p} \cdot \left(\sqrt{1-a} - \frac{1}{2}(\sqrt{1-a} - \sqrt{1-a-\beta_1})\right) > 0,\end{aligned}$$

which implies $\log\left(1 + \bar{\varepsilon}\left(\exp\left(\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2\right) - 1\right)\right) > 0$. Thus Markov's inequality can be applied to obtain

$$\mathbb{P}\left(R_2 > \frac{1}{4}\log|\mathcal{T}|\right) \le \frac{4\bar{\mathbb{E}}_1 R_2}{\log|\mathcal{T}|} \tag{3.36}$$

The expectation is

$$\begin{aligned}\bar{\mathbb{E}}_1 R_2 &= \sum_{j=1}^{p} \bar{\mathbb{E}}_1\left[1_{\{j\in\Gamma_2\}}\log\left(1 + \bar{\varepsilon}\left(e^{\rho\langle X_{j:}, \theta^{(k^*)}\rangle - \rho^2/2} - 1\right)\right)\right] \\ &\le \sum_{j=1}^{p} \bar{\mathbb{E}}_1\left[1_{\{j\in\Gamma_2\}}\log\left(1 + \bar{\varepsilon}\left(e^{\rho\cdot 2\sqrt{2\log p} - \rho^2/2} - 1\right)\right)\right] \\ &= \bar{\mathbb{E}}_1|\Gamma_2| \cdot \log\left(1 + p^C\right)\end{aligned}$$

for some constant C that is fixed as $p \to \infty$. Thus the polynomial dependence is determined by $\bar{\mathbb{E}}_1|\Gamma_2|$, which is

$$\begin{aligned}
\bar{\mathbb{E}}_1|\Gamma_2| &= \bar{\mathbb{E}}_1 \sum_{j=1}^{p} 1\{Q_j = 1, \sqrt{2(1-a)\log p} < \langle X_{j:}, \theta^{(k^*)}\rangle \leq 2\sqrt{2\log p}\} \\
&= \sum_{j=1}^{p} p^{-\bar{\beta}} \mathbb{P}(\sqrt{2(1-a)\log p} < \rho + Z \leq 2\sqrt{2\log p}) && (Z \sim N(0,1)) \\
&\leq p^{1-\bar{\beta}} \bar{\Phi}(\sqrt{2(1-a)\log p} - \rho) \\
&\asymp p^{1-\bar{\beta}-(\sqrt{1-a}-(x-y))^2}. && \text{(up to log factors in } p)
\end{aligned}$$

Plugging this into (3.36), we have (up to log factors in p) that

$$\bar{\mathbb{P}}_1\left(R_2 > \frac{1}{4}\log|\mathcal{T}|\right) \lesssim \frac{p^{1-\bar{\beta}-(1-a-\beta_1)}}{p^{a(1+o(1))}} \to 0$$

since $\beta_1 < \bar{\beta}$ are constant in p.

Finally, we check condition (3.31) for $i = 3$. We have

$$\begin{aligned}
\bar{\mathbb{P}}_1\left(R_3 > \frac{1}{4}\log|\mathcal{T}|\right) &\leq \bar{\mathbb{P}}_1(R_3 > 0) \\
&= \bar{\mathbb{P}}_1(\Gamma_3 \neq \varnothing) \\
&= \bar{\mathbb{P}}_1\left(\bigcup_{j=1}^{p} \left\{Q_j = 1, \langle X_{j:}, \theta^{(k^*)}\rangle > 2\sqrt{2\log p}\right\}\right) \\
&\leq \sum_{j=1}^{p} \bar{\mathbb{P}}_1\left(Q_j = 1, \langle X_{j:}, \theta^{(k^*)}\rangle > 2\sqrt{2\log p}\right) \\
&= p\bar{\varepsilon} \cdot \mathbb{P}(\rho + Z > 2\sqrt{2\log p}) \\
&= p^{1-\bar{\beta}} \bar{\Phi}(\sqrt{2\log p} \cdot (2-(x-y))) \\
&= p^{1-\bar{\beta}-(2-(x-y))^2}, && \text{(up to log factors in } p)
\end{aligned}$$

which tends to zero polynomially in p if the exponent $1 - \bar{\beta} - (2-(x-y))^2 < 0$. This condition is equivalent to

$$\sqrt{1-\bar{\beta}} + \sqrt{1-a} - \sqrt{1-a-\beta_1} < 2,$$

since $x = \sqrt{1-a}$ and $y = \sqrt{1-a-\beta_1}$. This condition is satisfied because the left hand side is

$$\sqrt{1-\bar{\beta}} + \sqrt{1-a-\beta_1+\beta_1} - \sqrt{1-a-\beta_1} < \sqrt{1-\bar{\beta}} + \sqrt{\beta_1} < 2,$$

because $\bar{\beta}, \beta_1 \in (0,1)$.

We have now shown (3.31), completing the proof of this case.

Case 4: $a > 1 - \beta_1$

In this case, a slightly different testing problem is used, in which the $\log\log n = p^a$ signal is placed evenly across the first $s = p^{1-\beta_1}$ rows (recall that $\rho^2 = p^{a-(1-\beta_1)}$). The lower bound construction is

$$H_0 : (X_{j:})_{j \le p} \sim \prod_{j=1}^{p} N(0, I_n)$$

$$H_1 : k^* \sim \mathrm{Unif}\{1, \dots, |\mathcal{T}|\}, (X_{j:})_{j \le s} \mid k^* \sim \prod_{j=1}^{s} N(\rho\theta^{(k^*)}, I_n), \quad (X_{j:})_{j > s} \sim \prod_{j=s+1}^{p} N(0, I_n).$$

The second moment is

$$\begin{aligned}
\mathbb{E}_0 \left(\frac{f_1}{f_0}(X) \right)^2 &= \mathbb{E}_1 \frac{f_1}{f_0}(X) \\
&= \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \mathbb{E}_1 \prod_{j=1}^{s} \frac{\phi(X_{j:} - \rho\theta^{(k)})}{\phi(X_{j:})} \\
&= \frac{1}{|\mathcal{T}|^2} \sum_{k,l} \prod_{j=1}^{s} \mathbb{E}_1 (\exp(\rho \langle X_{j:}, \theta^{(k)} \rangle - \rho^2/2) \mid k^* = l) \\
&= \frac{1}{|\mathcal{T}|^2} \sum_{k,l} e^{s\rho^2 \langle \theta^{(k)}, \theta^{(l)} \rangle} \\
&\le \frac{1}{|\mathcal{T}|} e^{p^{1-\beta+a-(1-\beta_1)}} + 1 + o(1),
\end{aligned}$$

since $\langle \theta^{(k)}, \theta^{(l)} \rangle \le e^{-\frac{|k-l|}{2} p^a}$. Now since $|\mathcal{T}| = e^{p^{a(1+o(1))}}$, the first term is

$$\frac{1}{|\mathcal{T}|} e^{p^{1-\beta+a-(1-\beta_1)}} = \exp\left(p^{a+(\beta_1-\beta)} - p^{a(1+o(1))} \right) \to 0, \qquad (\beta_1 < \beta)$$

which gives $\limsup_{p\to\infty} \mathbb{E}_0 \left(\frac{f_1}{f_0}(X) \right)^2 \le 1$, as desired.

3.5.2 Lower Bound for Theorem 3.2.2

For the two sided problem, consider the testing problem,

$$H_0 : (X_{j:})_{j\in[p]} \sim \prod_{j=1}^{p} N(0, I_n)$$

$$\bar{H}_1 : (X_{j:})_{j\in[p]} \sim \frac{1}{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|}\prod_{j=1}^{p}\left[(1-\bar{\varepsilon})N(0, I_n) + \bar{\varepsilon}\cdot\frac{1}{2}\left(N(\rho\theta^{(k)}, I_n) + N(-\rho\theta^{(k)}, I_n)\right)\right],$$

where $\rho \coloneqq \rho^*_{\text{2-side}}(a, \beta_1)$ where $a \geq 0$ and $\beta_1 \in (0, 1)$, $\bar{\varepsilon} = p^{-\bar{\beta}}$, where $\bar{\beta} < \beta_1$, and $\theta^{(k)}$ is defined as in the proof for Theorem 3.2.1. Under $\bar{H}_1$, the mean vector for a non-null row is either $\rho\theta^{(k)}$ or $-\rho\theta^{(k)}$ for some $k \leq |\mathcal{T}|$. Observe that for a vector $v_\mu \in \mathbb{R}^n$, whose first $\lfloor b^k \rfloor$ components are equal to μ_1, and remaining $n - \lfloor b^k \rfloor$ components are equal to μ_2, that

$$|\langle\theta^{(k)}, v_\mu\rangle| = \sqrt{\frac{\lfloor b^k\rfloor(n - \lfloor b^k\rfloor)}{n}}|\mu_1 - \mu_2|,$$

by the defining property of $\theta^{(k)}$ (Part 1 of Lemma 3.7.1). Taking $v_\mu \in \{\rho\theta^{(k)}, -\rho\theta^{(k)}\}$ and noting that $|\langle\theta^{(k)}, v_\mu\rangle| = \rho$, shows that each non-null mean vector generated in $\bar{H}_1$ satisfies the two-sided signal requirement of $\Theta_1^{\text{2-side}}(p, n, \rho, s)$. Together with the discussion on the choice of $\bar{\beta}$ in Section 3.5.1, this shows that the random instance of θ generated under $\bar{H}_1$ is an element of $\Theta_1^{\text{2-side}}(p, n, \rho, s)$ with high probability.

Since $\bar{\beta} < \beta_1$, the calculation (3.28) discussed in Section 3.5.1 implies that it is enough to show

$$\limsup_{p\to\infty} \mathbb{E}_0\left(\frac{f_1}{f_0}(X)\right)^2 \leq 1,$$

where f_0 and f_1 are the densities for H_0 and $\bar{H}_1$ respectively. As discussed in the beginning of Section 3.5, since the detection boundaries for the one sided and two sided problems are the same except for case $a \leq 1 - 2\beta_1$, it suffices to show the above second moment condition in this case. To this end, we compute

$$\begin{aligned}
\mathbb{E}_0 \left(\frac{f_1}{f_0}(X)\right)^2 &= \bar{\mathbb{E}}_1 \frac{f_1}{f_0}(X) \\
&= \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \bar{\mathbb{E}}_1 \prod_{j=1}^{p} \left[1 - \bar{\varepsilon} + \bar{\varepsilon} \frac{e^{-\rho^2/2}}{2} \left(e^{\rho \langle X_{j:}, \theta^{(k)} \rangle} + e^{-\rho \langle X_{j:}, \theta^{(k)} \rangle} \right) \right] \\
&= \frac{1}{|\mathcal{T}|^2} \sum_{k,l \leq |\mathcal{T}|} \prod_{j=1}^{p} \left[1 - \bar{\varepsilon} + \bar{\varepsilon} \frac{e^{-\rho^2/2}}{2} \bar{\mathbb{E}}_1 \left(e^{\rho \langle X_{j:}, \theta^{(k)} \rangle} + e^{-\rho \langle X_{j:}, \theta^{(k)} \rangle} \mid k^* = l \right) \right].
\end{aligned} \tag{3.37}$$

By definition of $\bar{H}_1$, the rows $X_{j:}$ are independent conditional on the event $k^* = l$. Thus, the terms in the conditional expectation can be computed,

$$\begin{aligned}
\bar{\mathbb{E}}_1(e^{\rho \langle X_{j:}, \theta^{(k)} \rangle} \mid k^* = l) &= (1 - \bar{\varepsilon}) e^{\rho^2/2} + \bar{\varepsilon} \cdot \frac{1}{2} \left(\mathbb{E} e^{\rho \langle \rho \theta^{(l)} + Z_{j:}, \theta^{(k)} \rangle} + \mathbb{E} e^{\rho \langle -\rho \theta^{(l)} + Z_{j:}, \theta^{(k)} \rangle} \right) \\
&= (1 - \bar{\varepsilon}) e^{\rho^2/2} + \bar{\varepsilon} \cdot \frac{e^{\rho^2/2}}{2} \left(e^{\rho^2 \langle \theta^{(l)}, \theta^{(k)} \rangle} + e^{-\rho^2 \langle \theta^{(l)}, \theta^{(k)} \rangle} \right) \\
&\leq (1 - \bar{\varepsilon}) e^{\rho^2/2} + \bar{\varepsilon} \cdot e^{\rho^2/2} e^{\rho^4 \langle \theta^{(l)}, \theta^{(k)} \rangle^2/2},
\end{aligned}$$

where we have used the inequality $\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}$ in the last line. A nearly identical calculation gives

$$\bar{\mathbb{E}}_1(e^{-\rho \langle X_{j:}, \theta^{(k)} \rangle} \mid k^* = l) \leq (1 - \bar{\varepsilon}) e^{\rho^2/2} + \bar{\varepsilon} \cdot e^{\rho^2/2} e^{\rho^4 \langle \theta^{(l)}, \theta^{(k)} \rangle^2/2}.$$

Plugging these two estimates back into (3.37), we obtain

$$\begin{aligned}
\mathbb{E}_0 \left(\frac{f_1}{f_0}(X)\right)^2 &\leq \frac{1}{|\mathcal{T}|^2} \sum_{k,l \leq |\mathcal{T}|} \prod_{j=1}^{p} \left[1 - \bar{\varepsilon} + \bar{\varepsilon} \cdot e^{-\rho^2/2} \left((1 - \bar{\varepsilon}) e^{\rho^2/2} + \bar{\varepsilon} \cdot e^{\rho^2/2} e^{\rho^4 \langle \theta^{(l)}, \theta^{(k)} \rangle^2/2} \right) \right] \\
&= \frac{1}{|\mathcal{T}|^2} \sum_{k,l \leq |\mathcal{T}|} \prod_{j=1}^{p} \left[1 + \bar{\varepsilon}^2 \left(e^{\rho^4 \langle \theta^{(l)}, \theta^{(k)} \rangle^2/2} - 1 \right) \right] \\
&\leq \frac{1}{|\mathcal{T}|^2} \sum_{k,l \leq |\mathcal{T}|} \exp \left(p^{1-2\bar{\beta}} (e^{\rho^4 \langle \theta^{(l)}, \theta^{(k)} \rangle^2/2} - 1) \right) \qquad (\log(1+x) \leq x) \\
&\leq \frac{1}{|\mathcal{T}|^2} \sum_{k=l} \exp \left(p^{1-2\bar{\beta}} (e^{\rho^4/2} - 1) \right) + \frac{1}{|\mathcal{T}|^2} \sum_{k \neq l} \exp \left(p \bar{\varepsilon}^2 (e^{\rho^4 e^{-\frac{1}{2} p^{a(1+o(1))}}/2} - 1) \right),
\end{aligned}$$

where in the last inequality, we have used Part 3 of Lemma 3.7.1. Now since $\rho^4 e^{-\frac{1}{2} p^{a(1+o(1))}} \to$

0, we have by the inequality $e^x - 1 \leq 2x$ for $0 < x < 1$ that the above is bounded by

$$\begin{aligned}
&\leq \frac{1}{|\mathcal{T}|} \exp\left(p^{1-2\bar{\beta}}(e^{\rho^4/2} - 1)\right) + \frac{1}{|\mathcal{T}|^2} \sum_{k \neq l} \underbrace{\exp\left(p\bar{\varepsilon}^2 \rho^4 e^{-\frac{1}{2}p^{a(1+o(1))}}\right)}_{1+o(1)} \\
&= \exp\left(p^{1-2\bar{\beta}}(e^{\rho^4/2} - 1) - p^{a(1+o(1))}\right) + 1 + o(1).
\end{aligned}$$

In order for the right hand side to tend to 1, it suffices for the first term to tend to zero, i.e. $\exp(p^{1-2\bar{\beta}}(e^{\rho^4/2} - 1) - p^{a(1+o(1))}) \to 0$. If $a < 1 - 2\beta_1$, we have $\rho^2 \sim \sqrt{\log(1 + p^{a-(1-2\beta_1)})}$, and the first term becomes

$$\exp(p^{1-2\bar{\beta}}(e^{\rho^4/2} - 1) - p^{a(1+o(1))}) \leq \exp(p^{1-2\bar{\beta}+a-(1-2\beta_1)} - p^{a(1+o(1))}) \to 0,$$

since $\beta_1 < \bar{\beta}$. If $a = 1 - 2\beta_1$, then $\rho^2 = 1$, so that

$$\exp(p^{1-2\bar{\beta}}(e^{\rho^4/2} - 1) - p^{a(1+o(1))}) = \exp(p^{(1-2\bar{\beta})(1+o(1))} - p^{(1-2\beta_1)(1+o(1))}) \to 0,$$

since $\beta_1 < \bar{\beta}$.

3.5.3 Upper Bound for Theorem 3.2.2

Put $\beta < \beta_1$ and $\rho \coloneqq \rho^*_{\text{2-side}}(a, \beta_1)$ as in the statement of Theorem 3.2.2. To show the second part of Theorem 3.2.2, it suffices to show that for any $\theta \in \Theta_1^{\text{2-side}}(p, n, \rho, s)$, the Type I and II errors of the penalized Berk–Jones statistic tend to zero as $p \to \infty$. The test statistic in this setting is defined,

$$\mathsf{PBJ}_p \coloneqq \left[\max_{k \leq |\mathcal{T}|} \sup_{q \in (0,1)} pK(\bar{S}_{p,k}(q), q)\right] - 2\log|\mathcal{T}|,$$

where $\mathcal{T}$ and $\bar{S}_{p,k}(q)$ are defined

$$\begin{aligned}
\mathcal{T} &\coloneqq \left\{\lfloor(1+\delta)^0\rfloor, \lfloor(1+\delta)^1\rfloor, \ldots, \lfloor(1+\delta)^{\log_{1+\delta}\frac{n}{2}}\rfloor, \lfloor n - (1+\delta)^{\log_{1+\delta}\frac{n}{2}}\rfloor, \ldots, \lfloor n - (1+\delta)^0\rfloor\right\}, \\
\bar{S}_{p,k} &\coloneqq \frac{1}{p}\sum_{j=1}^{p} \mathbf{1}_{\{|Y_{jk}| > \bar{\Phi}^{-1}(q/2)\}}, \qquad (3.38)
\end{aligned}$$

and Y_{jk} is the contrast corresponding to the corresponding to the k^{th} element in the grid $t_k \in \mathcal{T}$,

$$Y_{jk} \coloneqq \sqrt{\frac{t_k(n-t_k)}{n}}(\bar{X}_{j,1:t_k} - \bar{X}_{j,t_k+1:n}), \qquad 1 \leq j \leq p, \quad 1 \leq k \leq |\mathcal{T}|.$$

Note that in the current asymptotic setting (3.8), the cardinality of the grid is $|\mathcal{T}| = e^{p^{a(1+o(1))}}$. The test is performed by checking if the penalized Berk–Jones statistic exceeds the level $2(2+\gamma)\log p$, where $\gamma > 0$ is a small constant,

$$\psi_{\mathsf{PBJ}}(X) \coloneqq \mathbf{1}_{\{\mathsf{PBJ}_p > 2(2+\gamma)\log p\}}.$$

Type I error

For any $\theta \in \Theta_0(p,n)$, we have $Y_{jk} \sim N(0,1)$. The maximum over $q > 0$ is equivalent to taking a maximum over a finite set (Berk and Jones; 1979),

$$\max_{q\in(0,1)} K(\bar{S}_{p,k}(q), q) = \max_{j\le p} K(j/p, p_{k,(j)}),$$

where $p_{k,j} \coloneqq 2\bar{\Phi}(|Y_{jk}|)$ is the two sided p value corresponding to the k^{th} element in $\mathcal{T}$, and $p_{k,(1)} < \cdots < p_{k,(p)}$ are the ordered p values. Under the null, they are distributed as the order statistics of p iid Uniform$(0,1)$ variables. Then by the union bound,

$$\begin{aligned}\mathbb{P}_\theta(\mathsf{PBJ}_p > 2(2+\gamma)\log p) &\le \mathbb{P}_\theta\left(\max_{k\le|\mathcal{T}|, j\le p} pK(j/p, p_{k,(j)}) > 2(2+\gamma)\log p + 2\log|\mathcal{T}|\right)\\ &\le |\mathcal{T}| \sum_{j=1}^{p} \mathbb{P}_\theta\left(pK(j/p, p_{1,(j)}) > 2(2+\gamma)\log p + 2\log|\mathcal{T}|\right).\end{aligned}$$

By Lemma 3.6.5, the above is bounded by

$$\begin{aligned}&\le |\mathcal{T}| \left[(1+9/e)e^{-2(2+\gamma)\log p - 2\log|\mathcal{T}|} + \sum_{j=2}^{p} e\sqrt{2}je^{-(1-1/j)(2(2+\gamma)\log p + 2\log|\mathcal{T}|)}\right]\\ &\le |\mathcal{T}| \left[(1+9/e)e^{-\gamma\log p - \log|\mathcal{T}|} + e\sqrt{2}p\sum_{j=2}^{p} e^{-\frac{1}{2}(2(2+\gamma)\log p + 2\log|\mathcal{T}|)}\right]\\ &= |\mathcal{T}| \left[(1+9/e)e^{-\gamma\log p - \log|\mathcal{T}|} + e\sqrt{2}p^2 e^{-(2+\gamma)\log p - \log|\mathcal{T}|}\right] \to 0.\end{aligned}$$

Type II error

For any $\theta \in \Theta_1^{\text{2-side}}(p,n,\rho,s)$, the Type II error is,

$$\mathbb{P}_\theta(\mathsf{PBJ}_p \le 2(2+\gamma)\log p) = \mathbb{P}_\theta\left(\max_{k\le|\mathcal{T}|, q>0} K(\bar{S}_{p,k}(q), q) \le \frac{2(2+\gamma)\log p + 2p^{a(1+o(1))}}{p}\right),$$

since $|\mathcal{T}| = e^{p^{a(1+o(1))}}$. Then it suffices to pick $k \leq |\mathcal{T}|$ and $q > 0$ such that $K(\bar{S}_{p,k}(q), q) = \omega(p^{a-1+\zeta})$ in probability for some $\zeta > 0$, i.e.

$$\frac{K(\bar{S}_{p,k}(q), q)}{p^{a-1}} \xrightarrow{\mathbb{P}_\theta} \infty \quad \text{polynomially fast in } p. \tag{3.39}$$

Denote by t^* the true changepoint location in $\theta \in \Theta_1^{\text{2-side}}(p, n, \rho, s)$. The Berk–Jones test is equivalent to a threshold test (see Section 1.1 of Arias-Castro and Ying (2019)) whose rejection region is of the form,

$$\bigcup_{k \leq \mathcal{T}, t \in \mathcal{S}} \{\bar{S}_{p,k}(t) \geq c_t\}, \tag{3.40}$$

for some subset $\mathcal{S} \subset \mathbb{R}$ and (c_t) a set of critical values.[2] It follows from the definition (3.38) that if the means of the Gaussian statistics Y_{jk} are increased, the power of the threshold test increases, i.e. the Type II error decreases. Thus we assume in the following calculations that the lower bounds on s and ρ in the definition (3.11) are achieved, that is,

$$\sqrt{\frac{t^*(n-t^*)}{n}} |\mu_{j1} - \mu_{j2}| = \rho$$

$$\sum_{j=1}^{p} \mathbf{1}_{\{\text{row } j \text{ has a changepoint}\}} = s.$$

Without loss of generality, suppose that $t^* \leq n/2$; by symmetry of $\mathcal{T}$, an analogous argument can be made for $t^* > n/2$. By Lemma 3.6.6, there exists some $\tilde{t} := \lfloor (1+\delta)^{\tilde{k}} \rfloor \in \mathcal{T}$ for which

$$|Y_{j\tilde{k}}| = |\mu + Z| \text{ where } \mu = (1 + o(1))\rho, \tag{3.41}$$

when j corresponds to a non-null row in θ, and $\rho := \rho^*_{\text{2-side}}(a, \beta_1)$. When j corresponds to a null row in θ,

$$Y_{jk} \sim N(0,1) \qquad \text{for all } k = 1, \ldots, |\mathcal{T}|.$$

Throughout the following calculations, μ refers to the mean satisfying (3.41).

[2] The Berk–Jones statistic considered here coincides with the definition in Berk and Jones (1979), which was introduced as an approximation to the cdf of the Beta distribution. The approximation was proposed due to past issues in computing the Beta cdf, which can be computed using modern mathematical packages that are now standard. The ‘exact’ Berk–Jones statistics are defined via the cdf of the Beta distribution and are equivalent to the formulation originally considered by Berk and Jones (see Section 3 of Moscovich et al. (2016)). The rejection region of the ‘exact’ Berk–Jones statistic can be written in the form of a threshold test (see Section 1.1 of Arias-Castro and Ying (2019)).

Case 1: $a \leq 1 - 2\beta_1$

In the case $a < 1 - 2\beta_1$, we have $\rho = p^{\frac{a-(1-2\beta_1)}{4}} \to 0$. Put $q := 2\bar{\Phi}(\mu) \to 1$, where $\mu = (1 + o(1))\rho = (1 + o(1))p^{\frac{a-(1-2\beta_1)}{4}} \to 0$. Then by (3.41), we have

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{P}_\theta \left(|Y_{j\tilde{k}}| > \bar{\Phi}^{-1}(q/2) \right) \\
&= (1-\varepsilon)\mathbb{P}(|Z| > \mu) + \varepsilon\mathbb{P}(|\mu + Z| > \mu) \\
&= (1-\varepsilon)2\bar{\Phi}(\mu) + \varepsilon(\bar{\Phi}(0) + \bar{\Phi}(2\mu)) \to 1, \\
\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q)) &\leq \frac{1}{p^2}\sum_{j=1}^{p} \mathbb{P}_\theta \left(|Y_{j\tilde{k}}| > \bar{\Phi}^{-1}(q/2) \right) \leq p^{-1},
\end{aligned}$$

where $\varepsilon :=$ the fraction of non-null rows in θ. By Chebyshev's inequality,

$$|\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)| \leq p^{-1/2} \log p,$$

with probability tending to 1. Note by the above calculation, and the Mean Value Theorem that,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q &= \varepsilon(\bar{\Phi}(0) + \bar{\Phi}(2\mu) - 2\bar{\Phi}(\mu)) \\
&= \varepsilon(\phi(0)\mu^2 + o(\mu^2)) \\
&= \phi(0)p^{-\beta + \frac{a-(1-2\beta_1)}{2}}(1 + o(1)) \\
&= \phi(0)p^{(\beta_1 - \beta) + \frac{a}{2} - \frac{1}{2}}(1 + o(1)) \\
&= \omega(p^{-1/2} \log p),
\end{aligned}$$

since $\beta_1 > \beta$. Hence, by Part 1 of Lemma 3.6.7, we have

$$\begin{aligned}
K(\bar{S}_{p,\tilde{k}}(q), q) &\geq 2(\bar{S}_{p,\tilde{k}}(q) - q)^2 \\
&= 2(\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) + \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2 \\
&= 2(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2(1 + o_{\mathbb{P}_\theta}(1)) \\
&= 2\phi'(0)^2 p^{2(\beta_1 - \beta) + a - 1}(1 + o_{\mathbb{P}_\theta}(1)).
\end{aligned}$$

Then (3.39) follows since $\beta_1 > \beta$.

In the case $a = 1 - 2\beta_1$, we have $\rho^2 = 1$. Put $q := 2\bar{\Phi}(\mu)$. Then by the same calculation in the previous case,

$$\begin{aligned}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= (1-\varepsilon)2\bar{\Phi}(\mu) + \varepsilon(\bar{\Phi}(0) + \bar{\Phi}(2\mu)) \\ \operatorname{Var}_\theta(\bar{S}_{p,\tilde{k}}(q)) &\leq p^{-1}.\end{aligned}$$

We then have

$$\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q = \varepsilon(\bar{\Phi}(0) + \bar{\Phi}(2\mu) - 2\bar{\Phi}(\mu)) \gtrsim p^{-\beta} = \omega(p^{-1/2}\log p),$$

since $\beta < \beta_1 = \frac{1-a}{2} < \frac{1}{2}$. By Chebyshev's inequality,

$$|\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)| \leq p^{-1/2}\log p,$$

with probability tending to 1. By Part 1 of Lemma 3.6.7,

$$\begin{aligned}K(\bar{S}_{p,\tilde{k}}(q), q) &\geq 2(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2(1 + o_{\mathbb{P}_\theta}(1)) \\ &\gtrsim p^{-2\beta}(1 + o_{\mathbb{P}_\theta}(1)).\end{aligned}$$

Now since $-2\beta - (a-1) > 1 - 2\beta_1 - a = 0$, the condition (3.39) is satisfied.

Case 2: $1 - 2\beta_1 < a \leq 1 - 4\beta_1/3$

In this case, $\rho = \sqrt{(a - (1-2\beta_1))\log p} \to \infty$. Put $q \coloneqq 2\bar{\Phi}(2\mu) \to 0$. We have

$$\begin{aligned}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{P}_\theta\left(|Y_{j\tilde{k}}| > \bar{\Phi}^{-1}(q/2)\right) \\ &= (1-\varepsilon)\mathbb{P}(|Z| > 2\mu) + \varepsilon\mathbb{P}(|\mu + Z| > 2\mu) \\ &= (1-\varepsilon)2\bar{\Phi}(2\mu) + \varepsilon(\bar{\Phi}(\mu) + \bar{\Phi}(3\mu)) \qquad (3.42) \\ \operatorname{Var}_\theta(\bar{S}_{p,\tilde{k}}(q)) &\leq \frac{1}{p^2}\sum_{j=1}^{p} \mathbb{P}_\theta\left(|Y_{j\tilde{k}}| > \bar{\Phi}^{-1}(q/2)\right) = p^{-1}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q). \qquad (3.43)\end{aligned}$$

It follows that

$$\begin{aligned}\frac{\operatorname{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} &\leq \frac{1}{p\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)} \\ &= \frac{1}{p\big((1-\varepsilon)2\bar{\Phi}(2\mu) + \varepsilon(\bar{\Phi}(\mu) + \bar{\Phi}(3\mu))\big)} \\ &\leq \frac{1}{p^{1-\beta-\frac{1}{2}(a-(1-2\beta_1))(1+o(1))}} \to 0,\end{aligned}$$

where the last inequality follows from Mill's ratio, $\bar{\Phi}(\mu) \geq p^{-\frac{1}{2}(a-(1-2\beta_1))(1+o(1))}$. The convergence to zero follows since the inequality $1-\beta-\frac{1}{2}(a-(1-2\beta_1)) > 0$ is implied by the inequalities $a \leq 1-4\beta_1/3$, $\beta < \beta_1$, and $\beta_1 \leq \frac{3}{4}$ (if $\beta_1 > \frac{3}{4}$, then $0 \leq a \leq 1-4\beta_1/3$ could not be satisfied).

There are two possibilities that determine the behavior of the term $\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)$,

$$-2(a-(1-2\beta_1)) \geq -\beta - \frac{1}{2}(a-(1-2\beta_1)), \tag{3.44}$$

$$-2(a-(1-2\beta_1)) < -\beta - \frac{1}{2}(a-(1-2\beta_1)). \tag{3.45}$$

First suppose (3.44) holds, so that $2\bar{\Phi}(2\mu)(1+o(1)) \leq \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \leq 3\bar{\Phi}(2\mu)(1+o(1))$. It then follows from Chebyshev's inequality (since $\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} \to 0$) that $\frac{\bar{S}_{p,\tilde{k}}(q)}{q} \leq \frac{3}{2} + o_{\mathbb{P}_\theta}(1) \leq 4$ on a set with probability tending to 1. Then on this high probability set, by Part 2 of Lemma 3.6.7,

$$K(\bar{S}_{p,\tilde{k}}(q), q) \geq \frac{(\bar{S}_{p,\tilde{k}}(q) - q)^2}{9q}.$$

Note by (3.42), (3.43), and (3.44) that

$$\sqrt{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))} \leq \sqrt{3p^{-1-2(a-(1-2\beta_1))}}(1+o(1)) = o(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q),$$

since by Mill's ratio, $\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q \geq \varepsilon \bar{\Phi}(\mu) = p^{-\beta-\frac{1}{2}(a-(1-2\beta_1))(1+o(1))}$, and since

$$-\frac{1}{2} - (a-(1-2\beta_1)) < -\beta - \frac{1}{2}(a-(1-2\beta_1))$$

is implied by $\beta_1 > \beta$. Then it follows from another application of Chebyshev's inequality that,

$$\begin{aligned} K(\bar{S}_{p,\tilde{k}}(q), q) &\geq \frac{(\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) + \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2}{9q} \\ &= \frac{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2}{9q}(1 + o_{\mathbb{P}_\theta}(1)). \end{aligned}$$

By Mill's ratio, $q = 2\bar{\Phi}(2\mu) \leq 2p^{-2(a-(1-2\beta_1))(1+o(1))}$, so the above becomes

$$K(\bar{S}_{p,\tilde{k}}(q), q) \geq \frac{1}{18} \cdot p^{2(a-(1-2\beta_1))(1+o(1)) - 2\beta - (a-(1-2\beta_1))(1+o(1))} = \omega(p^{a-1}),$$

since $\beta < \beta_1$. (3.39) follows in this case.

Now suppose that (3.45) holds, so that $\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) = \varepsilon\bar{\Phi}(\mu)(1+o(1))$. Since $\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} \to 0$, it follows that,

$$\begin{aligned}
\frac{\bar{S}_{p,\tilde{k}}(q)}{q} &= \frac{\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)(1+o_{\mathbb{P}_\theta}(1))}{2\bar{\Phi}(2\mu)} && \text{(Chebyshev)} \\
&= \frac{\varepsilon\bar{\Phi}(\mu)(1+o_{\mathbb{P}_\theta}(1))}{2\bar{\Phi}(2\mu)} && \text{(by (3.45))} \\
&\geq p^{-\beta-\frac{1}{2}(a-(1-2\beta_1))(1+o(1))+2(a-(1-2\beta_1))(1+o(1))}(1+o_{\mathbb{P}_\theta}(1)) \to \infty, && \text{(Mill's ratio)} \\
\bar{S}_{p,\tilde{k}}(q) &= (\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))(1+o_{\mathbb{P}_\theta}(1)) \to 0, && \text{(Chebyshev)}
\end{aligned}$$

Then on a set with probability tending to 1, $e^2 \leq \frac{\bar{S}_{p,\tilde{k}}(q)}{q}$ and $\frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \leq \bar{S}_{p,\tilde{k}}(q) \leq \frac{1}{2}$. Then on this set, by Part 3 of Lemma 3.6.7,

$$\begin{aligned}
K(\bar{S}_{p,\tilde{k}}(q), q) &\geq \frac{1}{2}\bar{S}_{p,\tilde{k}}(q)\log\frac{\bar{S}_{p,\tilde{k}}(q)}{q} \\
&\geq \frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \\
&= \frac{1}{2}\varepsilon\bar{\Phi}(\mu)(1+o(1)) \\
&\geq \frac{1}{2}p^{-\beta-\frac{1}{2}(a-(1-2\beta_1))(1+o(1))} = \omega(p^{a-1}),
\end{aligned}$$

since $-\beta - \frac{1}{2}(a-(1-2\beta_1)) > a-1$ is implied by $a \leq 1-4\beta_1/3 < 1$ and $\beta < \beta_1$. Thus, (3.39) also holds in this case.

Case 3: $1 - 4\beta_1/3 < a \leq 1 - \beta_1$

In this case, put $x \coloneqq \sqrt{1-a}$ and $y \coloneqq \sqrt{1-a-\beta_1}$. Then,

$$\begin{aligned}
&\rho = (\sqrt{1-a} - \sqrt{1-a-\beta_1})\sqrt{2\log p} \to \infty, \\
&q \coloneqq 2\bar{\Phi}(x\sqrt{2\log p}) \to 0.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{P}_\theta\left(|Y_{j\tilde{k}}| > \bar{\Phi}^{-1}(q/2)\right) \\
&= (1-\varepsilon)2\bar{\Phi}(x\sqrt{2\log p}) + \varepsilon\mathbb{P}(|\mu + Z| > x\sqrt{2\log p}) \\
&= (1-\varepsilon)2\bar{\Phi}(x\sqrt{2\log p}) + \varepsilon\left(\bar{\Phi}((y+o(1))\sqrt{2\log p}) + \Phi((y-2x)(1+o(1))\sqrt{2\log p})\right) \\
&= 2p^{-x^2(1+o(1))} + p^{-\beta-y^2+o(1)}(1+o(1)) \qquad\qquad (y < 2x - y) \\
&\asymp p^{-\beta-y^2+o(1)},
\end{aligned}$$

since $\beta < \beta_1$. Check that the variance condition is satisfied,

$$\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} \leq \frac{1}{p\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)} = \frac{1}{2p^{1-x^2(1+o(1))} + p^{1-\beta-y^2+o(1)}} \to 0,$$

since $y^2 = 1 - a - \beta_1$ and $\beta < \beta_1$. Then by Chebyshev's inequality and Mill's ratio,

$$\frac{\bar{S}_{p,\tilde{k}}(q)}{q} = \frac{\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)}{q}(1 + o_{\mathbb{P}_\theta}(1)) \asymp p^{-\beta-y^2+o(1)+(1-a)(1+o(1))}(1 + o_{\mathbb{P}_\theta}(1)) \to \infty,$$

since $y^2 = 1 - a - \beta_1$ and $\beta < \beta_1$. Then on a set with probability tending to 1, we have $e^2 \leq \frac{\bar{S}_{p,\tilde{k}}(q)}{q}$ and $\frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \leq \bar{S}_{p,\tilde{k}}(q) \leq \frac{1}{2}$. On this set, by Part 3 of Lemma 3.6.7,

$$\begin{aligned}
K(\bar{S}_{p,\tilde{k}}(q), q) &\geq \frac{1}{2}\bar{S}_{p,\tilde{k}}(q)\log\frac{\bar{S}_{p,\tilde{k}}(q)}{q} \\
&\geq \frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \\
&\geq \frac{1}{2}p^{-\beta-y^2+o(1)}(1+o(1)) \\
&= \frac{1}{2}p^{-\beta-(1-a-\beta_1)+o(1)}(1+o(1)) = \omega(p^{a-1}),
\end{aligned}$$

since $\beta < \beta_1$. (3.39) now follows.

Case 4: $a > 1 - \beta_1$

In this case, $\rho^2 = p^{a-(1-\beta_1)} \to \infty$. Put $q := 2\bar{\Phi}(\mu/2) \le e^{-\frac{1}{8}(1+o(1))p^{a-(1-\beta_1)}} \to 0$. Then,

$$\begin{aligned}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{P}_\theta(|Y_{j\tilde{k}}| > \mu/2)\\ &= (1-\varepsilon)2\bar{\Phi}(\mu/2) + \varepsilon\mathbb{P}(|\mu + Z| > \mu/2)\\ &= (1-\varepsilon)2\bar{\Phi}(\mu/2) + \varepsilon(\bar{\Phi}(-\mu/2) + \bar{\Phi}(3\mu/2))\\ &= \varepsilon(1+o(1)).\end{aligned}$$

Check the variance condition,

$$\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} \le \frac{1}{p\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)} = p^{-1+\beta}(1+o(1)) \to 0.$$

Then we have by Chebyshev's inequality and Mill's ratio that,

$$\begin{aligned}\frac{\bar{S}_{p,\tilde{k}}(q)}{q} &= \frac{\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)}{q}(1 + o_{\mathbb{P}_\theta}(1))\\ &\ge \frac{\varepsilon}{e^{-\frac{1}{8}(1+o(1))p^{a-(1-\beta_1)}}}(1 + o_{\mathbb{P}_\theta}(1))\\ &\to \infty,\end{aligned}$$

since $a > 1 - \beta_1$. Then on a set with probability tending to 1, we have $e^2 \le \frac{\bar{S}_{p,\tilde{k}}(q)}{q}$ and $\frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \le \bar{S}_{p,\tilde{k}}(q) \le \frac{1}{2}$. On this set, by Part 3 of Lemma 3.6.7,

$$\begin{aligned}K(\bar{S}_{p,\tilde{k}}(q), q) &\ge \frac{1}{2}\bar{S}_{p,\tilde{k}}(q)\log\frac{\bar{S}_{p,\tilde{k}}(q)}{q}\\ &\ge \frac{1}{4}(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))\log\frac{\frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)}{e^{-\frac{1}{8}(1+o(1))p^{a-(1-\beta_1)}}}\\ &= \frac{1}{32}(1+o(1))p^{-\beta+a-(1-\beta_1)} = \omega(p^{a-1}),\end{aligned}$$

since $\beta < \beta_1$. Now (3.39) follows.

3.5.4 Upper Bound for Theorem 3.2.1

As discussed in the beginning of Section 3.5, it suffices to show the upper bound in the case $a \le 1 - 2\beta_1$, as the upper bound for the other cases is implied by the calculation in the previous section. It suffices to show that for any $\theta \in \Theta_1^{\text{1-side}}(p, n, \rho, s)$, the Type I and II errors of the penalized Berk–Jones statistic tend to zero as $p \to \infty$. The test statistic is the same as for the two sided problem, except $\bar{S}_{p,k}(q)$ is defined in a one sided manner,

$$\bar{S}_{p,k} := \frac{1}{p}\sum_{j=1}^{p} \mathbf{1}_{\{Y_{jk} > \bar{\Phi}^{-1}(q)\}}.$$

Since the one sided p values $p_{k,j} := \bar{\Phi}(Y_{jk})$ are distributed as iid Uniform$(0,1)$ variables under the null, the proof of the Type I error is exactly the same as in the two sided setting. We proceed to show that the Type II error also goes to zero when $a \leq 1 - 2\beta_1$.

Type II error

Suppose $\theta \in \Theta_1^{\text{1-side}}(p, n, \rho, s)$, with $\beta < \beta_1$ as in the statement of the Theorem. It suffices to pick $k \leq |\mathcal{T}|$ and $q > 0$ for which (3.39) is satisfied. Denote by t^* the true changepoint location in θ. By the discussion in Section 3.5.3 about threshold tests (see (3.40)), assume that the lower bounds ρ and s on the signal size and sparsity in the definition of $\Theta_1^{\text{1-side}}(p, n, \rho, s)$ are achieved. Without loss of generality, assume that $t^* \leq n/2$. By Lemma 3.6.6, there exists some $\tilde{t} := \lfloor (1+\delta)^{\tilde{k}} \rfloor \in \mathcal{T}$ for which

$$Y_{j\tilde{k}} = \mu + Z \text{ where } \mu = (1 + o(1))\rho, \tag{3.46}$$

when j corresponds to a non-null row in θ, and $\rho := \rho^*_{\text{1-side}}(a, \beta_1)$. When j corresponds to a null row in θ,

$$Y_{jk} \sim N(0,1) \qquad \text{for all } k = 1, \ldots, |\mathcal{T}|.$$

Throughout the following calculation, μ refers to the mean satisfying (3.46).

Case 1: $a \leq 1 - 2\beta_1$

First suppose $a < 1 - 2\beta_1$. In this case, $\rho = p^{\frac{a-(1-2\beta_1)}{2}} \to 0$. Put $q = \bar{\Phi}(2\mu) \to \frac{1}{2}$. Then

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{P}_\theta(Y_{j\tilde{k}} > \bar{\Phi}^{-1}(q)) \\
&= (1-\varepsilon)\mathbb{P}(Z > 2\mu) + \varepsilon\mathbb{P}(Z > \mu) \to \frac{1}{2}, \\
\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}) &\leq \frac{1}{p^2}\sum_{j=1}^{p} \mathbb{P}_\theta(Y_{j\tilde{k}} > \bar{\Phi}^{-1}(q)) \leq p^{-1}.
\end{aligned}$$

Then since $\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} \to 0$, it follows by Chebyshev's inequality that

$$|\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)| \leq p^{-1/2} \log p,$$

on a set with probability tending to 1. By the Mean Value Theorem,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q &= \varepsilon(\bar{\Phi}(\mu) - \bar{\Phi}(2\mu)) \\
&= \varepsilon(\phi(0)\mu + o(\mu)) \\
&= \phi(0) p^{-\beta + \frac{a-(1-2\beta_1)}{2}} (1 + o(1)) \\
&= \omega(p^{-1/2} \log p),
\end{aligned}$$

since $\beta < \beta_1$. Then by Part 1 of Lemma 3.6.7,

$$\begin{aligned}
K(\bar{S}_{p,\tilde{k}}(q), q) &\geq 2(\bar{S}_{p,\tilde{k}}(q) - q)^2 \\
&= 2(\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) + \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2 \\
&= 2(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2 (1 + o_{\mathbb{P}_\theta}(1)) \\
&= 2\phi(0)^2 p^{-2\beta + a - (1-2\beta_1)} (1 + o_{\mathbb{P}_\theta}(1)) \\
&= \omega(p^{a-1}),
\end{aligned}$$

since $\beta < \beta_1$. Hence, (3.39) holds.

In the case $a = 1 - 2\beta_1$, we have $\rho = 1$. Put $q \coloneqq \bar{\Phi}(\mu)$. Then by the same calculation as in the previous case,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= (1-\varepsilon)\bar{\Phi}(\mu) + \varepsilon \bar{\Phi}(0) \\
\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q)) &\leq p^{-1}.
\end{aligned}$$

Then,

$$\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q = \varepsilon(\bar{\Phi}(0) - \bar{\Phi}(\mu)) \gtrsim p^{-\beta} = \omega(p^{-1/2} \log p),$$

since $\beta < \beta_1 = \frac{1-a}{2} < \frac{1}{2}$. By Chebyshev's inequality,

$$|\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)| \leq p^{-1/2} \log p,$$

with probability tending to 1. By Part 1 of Lemma 3.6.7,

$$\begin{aligned}
K(\bar{S}_{p,\tilde{k}}(q), q) &\geq 2(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2 (1 + o_{\mathbb{P}_\theta}(1)) \\
&\gtrsim p^{-2\beta} (1 + o_{\mathbb{P}_\theta}(1)).
\end{aligned}$$

Now since $-2\beta - (a-1) > 1 - 2\beta_1 - a = 0$, the condition (3.39) is satisfied.

3.5.5 Lower Bound for Theorem 3.3.2

Recall the formula $\rho_2^*(a,\beta) = \sqrt{2r_2^*(a,\beta)\log p}$, where

$$r_2^*(a,\beta) := \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1-\sqrt{1-\beta})^2 & \frac{3}{4} < \beta < 1 \\ 1 + a & \beta = 1, \end{cases}$$

and the current asymptotic setting is,

$$\begin{aligned} \log\log n &\sim a \log p \qquad a > 0 \\ s &\sim p^{1-\beta} \qquad \frac{1}{2} < \beta \leq 1. \end{aligned}$$

Put $\rho := \sqrt{2r\log p}$, where $0 < r < r_2^*(a,\beta)$ as in the statement of the theorem. We show below that if $\beta \in (\frac{1}{2}, 1]$, then $\mathcal{R}_{\text{1-side}}(p,n,\rho,s) \to 1$ as $p \to \infty$. When $\beta = 1$, a different prior on θ is used than when $\beta \in (\frac{1}{2}, 1)$.

Cases 1 and 2: $\beta \in (\frac{1}{2}, 1)$

First suppose $\beta \in (\frac{1}{2}, 1)$. Since $0 < r < r_2^*(a,\beta)$, which is increasing in β, there exists some $\beta_1 \in (\frac{1}{2}, \beta)$ for which $r_2^*(a,\beta_1) = r$. Consider the testing problem,

$$\begin{aligned} &H_0 : X_{j:} \overset{\text{iid}}{\sim} N(0, I_n) \\ &\bar{H}_1 : X_{j:} \overset{\text{iid}}{\sim} (1-\bar{\varepsilon})N(0,I_n) + \bar{\varepsilon}N(\rho\theta^{(1)}, I_n), \end{aligned}$$

where $\bar{\varepsilon} := p^{-\bar{\beta}}$ with $\bar{\beta} \in (\beta_1, \beta)$, and $\theta^{(1)}$ is defined as in Section 3.5.1, and the reason for the choice $\bar{\beta} \in (\beta_1, \beta)$ is discussed in the beginning of Section 3.5.1. Since $\theta^{(1)}$ is a unit vector (see Lemma 3.7.1), there exist unit vectors $u_2, \ldots, u_n \in \mathbb{R}^n$ such that $\{\theta^{(1)}, u_2, \ldots, u_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Then each row $X_{j:}$ can be rewritten in this basis,

$$X_{j:} \equiv (X_{j:}^\top \theta^{(1)}, X_{j:}^\top u_2, \ldots, X_{j:}^\top u_n). \tag{3.47}$$

Since $u_i^\top \theta^{(1)} = 0$, we have $X_{j:}^\top u_i \overset{\text{iid}}{\sim} N(0,1)$ under both H_0 and H_1 for each $i = 2, \ldots, n$, and

$$X_{j:}^\top \theta^{(1)} \overset{\text{iid}}{\sim} \begin{cases} N(0,1) & \text{under } H_0 \\ (1-\bar{\varepsilon})N(0,1) + \bar{\varepsilon}N(\rho,1) & \text{under } H_1. \end{cases}$$

Hence only the first coordinate $X_{j:}^\top \theta^{(1)}$ is informative for testing between H_0 and H_1, and the remaining coordinates in (3.47) can be ignored. Finally, note that $\bar\beta > \beta_1$, and $\rho = \sqrt{2r\log p} = \sqrt{2r_2^*(a,\beta_1)\log p}$ denotes the minimal signal strength for detecting a non-null fraction $p^{-\beta_1} > p^{-\bar\beta}$ of signals. The result now follows from the lower bound for the original Ingster–Donoho–Jin problem.

Case 3: $\beta = 1$

In this case, $\rho = \sqrt{2r\log p} < \sqrt{2(1+a)\log p}$. We show that $\mathcal{R}_{\text{1-side}}(p,n,\rho,s=1) \to 1$ as $p\to\infty$. Consider the testing problem,

$$\begin{aligned} H_0 : X_{j:} \overset{\text{iid}}{\sim} N(0,I_n) \quad \text{vs} \quad H_1 : k &\sim \text{Unif}\{1,\dots,|\mathcal{T}|\} \\ i &\sim \text{Unif}\{1,\dots,p\} \\ X_{j:} \mid k,i &\overset{\text{indep.}}{\sim} \begin{cases} N(\rho\theta^{(k)}, I_n) & j=i \\ N(0,I_n) & j\neq i, \end{cases} \end{aligned}$$

for $j=1,\dots,p$, where the grid $\mathcal{T}$ is defined,

$$\mathcal{T} \coloneqq \{\lfloor b^1\rfloor, \lfloor b^2\rfloor, \dots, \lfloor b^{\log_{\log n} n}\rfloor\},$$

with base $b = \log n$, and the $\theta^{(k)}$ are defined as in previous sections (see Lemma 3.7.1). Note that the cardinality of this grid is $|\mathcal{T}| \asymp \frac{\log n}{\log\log n} = p^{a(1+o(1))}$, according to the calibration (3.9). It suffices to show the following two conditions,

$$\mathbb{P}_1(A_p^c) \to 0 \tag{3.48}$$

$$\limsup_{p\to\infty} \mathbb{E}_0 \left(\frac{f_1}{f_0}(X)\right)^2 \mathbf{1}_{A_p} \le 1, \tag{3.49}$$

for a suitable truncation event A_p, where f_0 and f_1 are the densities of X corresponding to H_0 and $\bar H_1$ respectively. Indeed, if (3.48) and (3.49) hold, then by monotonicity and Markov's inequality,

$$\begin{aligned} \mathbb{P}_0\left(\left|\frac{f_1}{f_0}(X) - 1\right| > \eta\right) &\le \mathbb{P}_0\left(\left|\frac{f_1}{f_0}(X)\mathbf{1}_{A_p} - 1\right| > \eta\right) \qquad (\eta\in(0,1)) \\ &\le \frac{\mathbb{E}_0\left(\frac{f_1}{f_0}(X)\right)^2\mathbf{1}_{A_p} - 2\mathbb{E}_0\frac{f_1}{f_0}(X)\mathbf{1}_{A_p} + 1}{\eta^2} \to 0, \end{aligned}$$

since $\mathbb{P}_1(A_p^c) \to 0$ is equivalent to $\mathbb{E}_0 \frac{f_1}{f_0}(X)\mathbf{1}_{A_p} \to 1$. To this end, define the truncation event,

$$A_p := \left\{ \max_{k \le |\mathcal{T}|, j \le p} \langle X_{j:}, \theta^{(k)} \rangle \le \sqrt{2(1+a+\gamma)\log p} \right\},$$

for some small constant γ satisfying

$$0 < \gamma < 4r - (1+a). \tag{3.50}$$

Check that (3.48) holds,

$$\begin{aligned}
\mathbb{P}_1(A_p^c) &= \frac{1}{p|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \sum_{i=1}^{p} \mathbb{P}\left(\max_{m \le |\mathcal{T}|, j \le p} \langle \rho\theta^{(k)} \mathbf{1}_{\{i=j\}} + Z_{j:}, \theta^{(m)} \rangle > \sqrt{2(1+a+\gamma)\log p} \right) \\
&= \frac{1}{p|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \sum_{i=1}^{p} \sum_{m=1}^{|\mathcal{T}|} \sum_{j=1}^{p} \mathbb{P}\big(\langle \rho\theta^{(k)} \mathbf{1}_{\{i=j\}} + Z_{j:}, \theta^{(m)} \rangle > \sqrt{2(1+a+\gamma)\log p} \big),
\end{aligned}$$

where $Z_{j:} \overset{\text{iid}}{\sim} N(0, I_n)$. Since $\langle \theta^{(k)}, \theta^{(m)} \rangle \le b^{-\frac{|k-m|}{2}}$, which is $o(1/\rho)$ when $k \neq m$, the sum can be split into three pieces,

$$\begin{aligned}
&\le \frac{1}{p|\mathcal{T}|} \sum_{(k,m,i,j): j=i, m=k} \bar{\Phi}\big(\sqrt{2(1+a+\gamma)\log p} - \rho \big) \\
&+ \frac{1}{p|\mathcal{T}|} \sum_{(k,m,i,j): j=i, m \neq k} \bar{\Phi}\big(\sqrt{2(1+o(1))(1+a+\gamma)\log p} \big) \\
&+ \frac{1}{p|\mathcal{T}|} \sum_{(k,m,i,j): j \neq i} \bar{\Phi}\big(\sqrt{2(1+a+\gamma)\log p} \big),
\end{aligned}$$

where we have used that $\langle Z_{j:}, \theta^{(m)} \rangle \sim N(0,1)$ for every $m \le |\mathcal{T}|$ and $j \le p$. Apply Mill's ratio to the above tail probabilities, and note that $|\mathcal{T}| \asymp \frac{\log n}{\log\log n} = p^{a(1+o(1))}$ according to the calibration (3.9), so that the above becomes bounded by

$$\begin{aligned}
&\le \frac{1}{p^{1+a(1+o(1))}} p^{1+a(1+o(1))} p^{-(\sqrt{1+a+\gamma} - \sqrt{r})^2} \\
&+ \frac{1}{p^{1+a(1+o(1))}} p^{1+2a(1+o(1))} p^{-(1+a+\gamma)(1+o(1))} \\
&+ \frac{1}{p^{1+a(1+o(1))}} p^{2(1+a(1+o(1)))} p^{-(1+a+\gamma)} \to 0,
\end{aligned}$$

since $1 + a + \gamma > r$ and $\gamma > 0$. To check (3.49), compute the truncated second moment,

$$\begin{aligned}\mathbb{E}_0 \left(\frac{f_1}{f_0}(X)\right)^2 \mathbf{1}_{A_p} &= \mathbb{E}_1 \frac{f_1}{f_0}(X)\mathbf{1}_{A_p} \\ &= \frac{1}{p|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \sum_{i=1}^{p} \underbrace{\mathbb{E}_1 \left(e^{\rho\langle X_{i:}, \theta^{(k)}\rangle - \rho^2/2} \mathbf{1}_{A_p}\right)}_{(*)}.\end{aligned}$$

The term $(*)$ is,

$$\begin{aligned}(*) &= \frac{1}{p|\mathcal{T}|} \sum_{l=1}^{|\mathcal{T}|} \sum_{j=1}^{p} \mathbb{E} e^{\rho\langle \rho\theta^{(l)} \mathbf{1}_{\{i=j\}} + Z_{i,:}, \theta^{(k)}\rangle - \rho^2/2} \mathbf{1}_{A_p} \\ &\leq \frac{1}{p|\mathcal{T}|} \sum_{l=1}^{|\mathcal{T}|} \left[(p-1) + e^{\rho^2\langle\theta^{(l)}, \theta^{(k)}\rangle} \mathbb{E} e^{\rho Z - \rho^2/2} \mathbf{1}_{\{\rho\langle\theta^{(l)}, \theta^{(k)}\rangle + Z \leq \sqrt{2(1+a+\gamma)\log p}\}}\right] \\ &\leq 1 + \frac{1}{p|\mathcal{T}|} \left[(|\mathcal{T}| - 1)(1 + o(1)) + e^{\rho^2} \Phi(\sqrt{2(1+a+\gamma)\log p} - 2\rho)\right] \\ &\leq 1 + o(1) + \frac{1}{p^{1+a(1+o(1))}} \cdot p^{2r} \cdot p^{-(\sqrt{1+a+\gamma} - 2\sqrt{r})^2}\end{aligned}$$

where in the third line, we have used Part 3 of Lemma 3.7.1 to claim $\rho^2\langle\theta^{(l)}, \theta^{(k)}\rangle \to 0$ when $l \neq k$, and in the last line, we have used that $\sqrt{1+a+\gamma} - 2\sqrt{r} < 0$ is implied by the condition (3.50). The right hand side becomes

$$\begin{aligned}&= 1 + o(1) + p^{-(1+a)(1+o(1)) + 2(1+a)\frac{r}{1+a} - (1+a)\left(\sqrt{1+\frac{\gamma}{1+a}} - 2\sqrt{\frac{r}{1+a}}\right)^2} \\ &= 1 + o(1) + p^{-(1+a)(1+o(1))\left(1 - 2\frac{r}{1+a} + \left(\sqrt{1+\frac{\gamma}{1+a}} - 2\sqrt{\frac{r}{1+a}}\right)^2\right)}.\end{aligned}$$

As $\gamma \to 0$, the above becomes

$$(*) = 1 + o(1) + p^{-2(1+a)(1+o(1))\left(1 - \sqrt{\frac{r}{1+a}}\right)^2 + O(\gamma)},$$

It follows from the assumption $r < 1 + a$ that γ can be chosen small enough to make the exponent negative, yielding $(*) = 1 + o(1)$. Plugging back into the truncated second moment gives

$$\begin{aligned}\mathbb{E}_0\left(\frac{f_1}{f_0}(X)\right)^2\mathbf{1}_{A_p} &= \mathbb{E}_1\frac{f_1}{f_0}(X)\mathbf{1}_{A_p}\\ &\le \frac{1}{p|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|}\sum_{i=1}^{p}(1+o(1))\\ &= 1+o(1),\end{aligned}$$

as desired.

3.5.6 Upper Bound for Theorem 3.3.2

We consider the cases $\beta \in (1/2, 1)$ and $\beta = 1$ separately; the penalized Berk–Jones test gives the upper bound for the former case, while a simple maximum statistic gives the upper bound for the latter case. The tests can be combined via $\psi = \psi_{\mathsf{PBJ}} \vee \psi_{\max}$ to give an overall test achieving the upper bound in the theorem.

First let $\beta \in (1/2, 1)$. Suppose without loss of generality that $r \in (0, 1)$; if $r \ge 1$, then

$$\Theta_1^{\text{1-side}}(p, n, \rho = \sqrt{2r\log p}, s) \subset \Theta_1^{\text{1-side}}(p, n, \rho = \sqrt{2\log p}, s),$$

and thus the worst case testing error becomes smaller. Since $r > r^*(a, \beta)$, there exists some $\beta_1 > \beta$ for which $r_2^*(a, \beta_1) = r$. Put $\rho \coloneqq \sqrt{2r\log p}$ as in the statement of Theorem 3.3.2. To show the upper bound, it suffices to show that for any $\theta \in \Theta_1^{\text{1-side}}(p, n, \rho, s)$, the Type I and II errors of the penalized Berk–Jones statistic tend to zero as $p \to \infty$. Here the test statistic is,

$$\mathsf{PBJ}_p \coloneqq \left[\max_{k\le|\mathcal{T}|}\sup_{q\in(0,1)} pK(\bar{S}_{p,k}(q), q)\right] - 2\log|\mathcal{T}|$$

where we recall from the beginning of Section 3.5.3 that $\mathcal{T}$ and $\bar{S}_{p,k}(q)$ are defined

$$\mathcal{T} \coloneqq \left\{\lfloor(1+\delta)^0\rfloor, \lfloor(1+\delta)^1\rfloor, \ldots, \lfloor(1+\delta)^{\log_{1+\delta}\frac{n}{2}}\rfloor, \lfloor n-(1+\delta)^{\log_{1+\delta}\frac{n}{2}}\rfloor, \ldots, \lfloor n-(1+\delta)^0\rfloor\right\},$$

$$\bar{S}_{p,k} \coloneqq \frac{1}{p}\sum_{j=1}^{p}\mathbf{1}_{\{Y_{jk}>\bar{\Phi}^{-1}(q)\}},$$

and Y_{jk} is the contrast corresponding to the corresponding to the k^{th} element in the grid $t_k \in \mathcal{T}$,

$$Y_{jk} \coloneqq \sqrt{\frac{t_k(n-t_k)}{n}}(\bar{X}_{j,1:t_k} - \bar{X}_{j,t_k+1:n}), \qquad 1\le j\le p, \quad 1\le k\le|\mathcal{T}|. \tag{3.51}$$

Note that $|\mathcal{T}| = p^{a(1+o(1))}$ so that $2\log|\mathcal{T}| = 2a(1+o(1))\log p$ in the current asymptotic setting (3.9). The test is performed by checking if the penalized Berk–Jones statistic exceeds the level $2(2+\gamma)\log p$, where $\gamma > 0$ is a small constant,

$$\psi_{\mathsf{PBJ}}(X) := \mathbf{1}_{\{\mathsf{PBJ}_p > 2(2+\gamma)\log p\}}.$$

Since the one sided p values $p_{k,j} := \bar{\Phi}(Y_{jk})$ are distributed as iid Uniform$(0,1)$ variables under the null, the proof of the Type I error is exactly the same as in Section 3.5.3. We proceed to show that the Type II error also goes to zero.

For any $\theta \in \Theta_1^{\text{1-side}}(p,n,\rho,s)$, the Type II error is,

$$\mathbb{P}_\theta(\mathsf{PBJ}_p \le 2(2+\gamma)\log p) = \mathbb{P}_\theta\left(\max_{k\le|\mathcal{T}|,q>0} pK(\bar{S}_{p,k}(q),q) \le \big(2(2+\gamma)+2a(1+o(1))\big)\log p\right),$$

since $|\mathcal{T}| = p^{a(1+o(1))}$. Then it suffices to pick $k \le |\mathcal{T}|$ and $q > 0$ such that $K(\bar{S}_{p,k}(q),q) = \omega(p^{-1+\zeta})$ in probability for some $\zeta > 0$, i.e.

$$pK(\bar{S}_{p,k}(q),q) \xrightarrow{\mathbb{P}_\theta} \infty \quad \text{polynomially fast in } p. \tag{3.52}$$

Denote by t^* the true changepoint location in θ. By the discussion in Section 3.5.3 about threshold tests (see (3.40)), assume that the lower bounds ρ and s on the signal size and sparsity in the definition of $\Theta_1^{\text{1-side}}(p,n,\rho,s)$ are achieved. Without loss of generality, suppose $t^* \le n/2$; by symmetry of $\mathcal{T}$, an analogous argument can be made for $t^* > n/2$. By Lemma 3.6.6, there exists some $\tilde{t} := \lfloor(1+\delta)^{\tilde{k}}\rfloor \in \mathcal{T}$ for which

$$Y_{j\tilde{k}} \sim N(\mu,1) \text{ where } \mu = (1+o(1))\rho, \tag{3.53}$$

when j corresponds to a non-null row in θ. When j corresponds to a null row in θ,

$$Y_{jk} \sim N(0,1) \qquad \text{for all } k = 1,\ldots,|\mathcal{T}|.$$

Throughout the following calculations, μ refers to the mean satisfying (3.53). Since $r \in (0,1)$, there are two cases: $r \in (0,\frac{1}{4}]$ and $r \in (\frac{1}{4},1)$. Case 3 is when $\beta = 1$, where it is shown that as long as $r > 1+a$, we have $\mathcal{R}_{\text{1-side}}(p,n,\rho=\sqrt{2r\log p},1) \to 0$.

Case 1: $0 < r \le \frac{1}{4}$

In this case, $\rho = \sqrt{2r \log p} = \sqrt{(2\beta_1 - 1)\log p}$, and $\beta_1 \in (\frac{1}{2}, \frac{3}{4}]$. Put $q \coloneqq \bar{\Phi}(2\mu) \to 0$. Then,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= \frac{1}{p}\sum_{j=1}^{p} \mathbb{P}_\theta \left(Y_{j\tilde{k}} > \bar{\Phi}^{-1}(q) \right) \\
&= (1-\varepsilon)\mathbb{P}(Z > 2\mu) + \varepsilon \mathbb{P}(\mu + Z > 2\mu) \\
&= (1-\varepsilon)\bar{\Phi}(2\mu) + \varepsilon\bar{\Phi}(\mu), && (3.54) \\
\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q)) &\le \frac{1}{p^2}\sum_{j=1}^{p} \mathbb{P}_\theta(Y_{j\tilde{k}} > \bar{\Phi}^{-1}(q)) = p^{-1}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q), && (3.55)
\end{aligned}$$

where we have assumed without loss of generality that $\varepsilon = p^{-\beta}$ is the number of non-null rows in θ. Apply Mill's ratio to bound the difference $\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q$,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q &= \varepsilon(\bar{\Phi}(\mu) - \bar{\Phi}(2\mu)) \\
&= p^{-\beta}\big(p^{-\frac{1}{2}(2\beta_1-1)(1+o(1))} - p^{-2(2\beta_1-1)(1+o(1))}\big). && (3.56)
\end{aligned}$$

It is straightforward to check that the assumptions $\beta < \beta_1$ and $\beta_1 \le \frac{3}{4}$ imply the following inequalities,

$$\begin{aligned}
\frac{1}{2}(-1 - 2(2\beta_1 - 1)) &< -\beta - \frac{1}{2}(2\beta_1 - 1) \\
\frac{1}{2}\big(-1 - \beta - \frac{1}{2}(2\beta_1 - 1)\big) &< -\beta - \frac{1}{2}(2\beta_1 - 1).
\end{aligned}$$

Together with (3.54), (3.55), and (3.56), the above two inequalities imply that,

$$\begin{aligned}
\sqrt{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))} &\le \sqrt{p^{-1}\big(p^{-2(2\beta_1-1)(1+o(1))} + p^{-\beta-\frac{1}{2}(2\beta_1-1)(1+o(1))}\big)} \\
&= o(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q). && (3.57)
\end{aligned}$$

Now check that

$$\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} \le \frac{1}{p\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)} = \frac{1}{p(p^{-2(2\beta_1-1)(1+o(1))} + p^{-\beta-\frac{1}{2}(2\beta_1-1)(1+o(1))})} \to 0,$$

since $1 - \beta - \frac{1}{2}(2\beta_1 - 1) > 0$ is implied by the assumptions $\beta < \beta_1$ and $\beta_1 \leq \frac{3}{4}$. It then follows by Chebyshev's inequality and (3.56) that

$$\begin{aligned}
\frac{\bar{S}_{p,\tilde{k}}(q)}{q} &= \frac{\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)}{q}(1 + o_{\mathbb{P}_\theta}(1)) \\
&= \frac{q + \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q}{q}(1 + o_{\mathbb{P}_\theta}(1)) \\
&\leq \left(1 + p^{-\beta - \frac{1}{2}(2\beta_1 - 1)(1+o(1))}\right)(1 + o_{\mathbb{P}_\theta}(1)) \leq 4,
\end{aligned}$$

on a set with probability tending to 1. Then by Part 1 of Lemma 3.6.7,

$$\begin{aligned}
K(\bar{S}_{p,\tilde{k}}(q), q) &\geq \frac{(\bar{S}_{p,\tilde{k}}(q) - q)^2}{9q} \\
&= \frac{(\bar{S}_{p,\tilde{k}}(q) - \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) + \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2}{9q} \\
&= \frac{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) - q)^2}{9q}(1 + o_{\mathbb{P}_\theta}(1)),
\end{aligned}$$

where the last equality follows from Chebyshev's inequality and (3.57). Then to show (3.52) holds for $\tilde{k}$ and q, by (3.56) it suffices to check that

$$p \cdot \frac{\left(p^{-\beta - \frac{1}{2}(2\beta_1 - 1)(1+o(1))} - p^{-\beta - 2(2\beta_1 - 1)(1+o(1))}\right)^2}{p^{-2(2\beta_1 - 1)(1+o(1))}} \to \infty.$$

The above divergence is implied by $1 - 2\beta - (2\beta_1 - 1) + 2(2\beta_1 - 1) > 0$, which in turn is implied by the assumption $\beta < \beta_1$.

Case 2: $\frac{1}{4} < r < 1$

In this case, $\rho = \sqrt{2r \log p} = (1 - \sqrt{1 - \beta_1})\sqrt{2 \log p}$, and $\beta_1 \in (\frac{3}{4}, 1)$. Put $q := \bar{\Phi}(\sqrt{2 \log p}) \to 0$. Then,

$$\begin{aligned}
\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) &= (1 - \varepsilon)\mathbb{P}(Z > \sqrt{2 \log p}) + \varepsilon \mathbb{P}(\mu + Z > \sqrt{2 \log p}) \\
&= (1 - \varepsilon)\bar{\Phi}(\sqrt{2 \log p}) + \varepsilon \bar{\Phi}\left((1 + o(1))\sqrt{1 - \beta_1}\sqrt{2 \log p}\right) \\
&= p^{-(1+o(1))} + p^{-\beta - (1-\beta_1)(1+o(1))}, \\
\frac{\mathrm{Var}_\theta(\bar{S}_{p,\tilde{k}}(q))}{(\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q))^2} &\leq \frac{1}{p\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)} = \frac{1}{p(p^{-(1+o(1))} + p^{-\beta - (1-\beta_1)(1+o(1))})} \to 0,
\end{aligned}$$

since $\beta < \beta_1$. It then follows by Chebyshev's inequality that

$$\begin{aligned}\frac{\bar{S}_{p,\tilde{k}}(q)}{q} &= \frac{\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)}{q}(1 + o_{\mathbb{P}_\theta}(1)) \\ &= \frac{p^{-(1+o(1))} + p^{-\beta-(1-\beta_1)(1+o(1))}}{p^{-(1+o(1))}}(1 + o_{\mathbb{P}_\theta}(1)) \\ &\to \infty,\end{aligned}$$

since $\beta < \beta_1$. The same application of Chebyshev's inequality gives

$$\bar{S}_{p,\tilde{k}}(q) = \mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q)(1 + o_{\mathbb{P}_\theta}(1)) \to 0 \text{ in } \mathbb{P}_\theta.$$

Then on a set with probability tending to 1, we have that $e^2 \leq \frac{\bar{S}_{p,\tilde{k}}(q)}{q}$ and $\frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \leq \bar{S}_{p,\tilde{k}}(q) \leq \frac{1}{2}$. On this set, by Part 3 of Lemma 3.6.7,

$$\begin{aligned}K(\bar{S}_{p,\tilde{k}}(q), q) &\geq \frac{1}{2}\bar{S}_{p,\tilde{k}}(q) \log \frac{\bar{S}_{p,\tilde{k}}(q)}{q} \\ &\geq \frac{1}{2}\mathbb{E}_\theta \bar{S}_{p,\tilde{k}}(q) \\ &\asymp p^{-\beta-(1-\beta_1)(1+o(1))}.\end{aligned}$$

(3.52) now follows since $\beta < \beta_1$.

Case 3: $\beta = 1$

In this case, we show that if $r > 1 + a$ and $\rho := \sqrt{2r \log p}$, then $\mathcal{R}_{\text{1-side}}(p, n, \rho, 1) \to 0$, as $p \to \infty$. Any $\theta \in \Theta_1^{\text{1-side}}(p, n, \rho, 1)$ has at least one non-null row. Consider the test,

$$\psi_{\max}(X) := \mathbf{1}\left\{\max_{k \leq |\mathcal{T}|, j \leq p} Y_{jk} > \sqrt{(2+\gamma)\log(p|\mathcal{T}|)}\right\},$$

where the Y_{jk} are defined in (3.51), and $\gamma > 0$ is a small constant satisfying,

$$\sqrt{(2+\gamma)(1+a)} < \sqrt{2r}. \tag{3.58}$$

Since each $Y_{jk} \sim N(0,1)$ under any null parameter $\theta \in \Theta_0(p,n)$, a union bound yields the Type I error control,

$$\begin{aligned}\mathbb{P}_\theta\left(\max_{k \le |\mathcal{T}|, j \le p} Y_{jk} > \sqrt{(2+\gamma)\log(p|\mathcal{T}|)}\right) &\le \sum_{k=1}^{|\mathcal{T}|}\sum_{j=1}^{p} \bar{\Phi}\big(\sqrt{(2+\gamma)\log(p|\mathcal{T}|)}\big) \\ &= p^{1+a(1+o(1))} p^{-\frac{2+\gamma}{2}(1+a(1+o(1)))} \\ &\to 0,\end{aligned}$$

since $|\mathcal{T}| = p^{a(1+o(1))}$. Now suppose $\theta \in \Theta_1^{\text{1-side}}(p,n,\rho,1)$. The Type II error is

$$\mathbb{P}_\theta(\psi_{\max} = 0) = \mathbb{P}_\theta\left(\max_{k \le |\mathcal{T}|, j \le p} Y_{jk} \le \sqrt{(2+\gamma)\log(p|\mathcal{T}|)}\right) \le \mathbb{P}_\theta\big(Y_{jk} \le \sqrt{(2+\gamma)\log(p|\mathcal{T}|)}\big),$$

for any $j \le p$ and $k \le |\mathcal{T}|$. Since the max test only becomes more powerful if the means of Y_{jk} are increased, we may assume that the lower bound ρ on the signal size in the definition of $\Theta_1^{\text{1-side}}(p,n,\rho,s=1)$ is achieved. By definition of $\theta \in \Theta_1^{\text{1-side}}(p,n,\rho,s=1)$, there is some non-null row in θ. Denote the index of this non-null row as $i \in [p]$. Without loss of generality, suppose that the true changepoint location t^* in θ satisfies $t^* \le n/2$; by symmetry of the grid, an analogous argument can be made for $t^* > n/2$. By the second part of Lemma 3.6.6, there exists some $\tilde{t} := \lfloor (1+\delta)^{\tilde{k}} \rfloor \in \mathcal{T}$ for which

$$Y_{i\tilde{k}} \sim N(\mu, 1) \text{ where } \mu = (1+o(1))\rho,$$

under $\mathbb{P}_\theta$. Then the Type II error is bounded,

$$\begin{aligned}\mathbb{P}_\theta(\psi_{\max} = 0) &\le \mathbb{P}_\theta\big(Y_{i\tilde{k}} \le \sqrt{(2+\gamma)\log(p|\mathcal{T}|)}\big) \\ &= \mathbb{P}\big(\mu + Z \le \sqrt{(2+\gamma)(1+a)\log p}\big) \\ &= \Phi\left(\sqrt{(2+\gamma)(1+a)\log p} - (1+o(1))\sqrt{2r\log p}\right) \\ &\to 0,\end{aligned}$$

by the assumption that γ satisfies (3.58).

3.6 Technical Lemmas

Lemma 3.6.1. *Let $\mathcal{T}$ be the grid of geometrically growing changepoint locations with base b,*

$$\mathcal{T} := \{\lfloor b^1 \rfloor, \lfloor b^2 \rfloor, \dots, \lfloor b^{\log_{\log n} n} \rfloor\}.$$

We abuse notation and treat b^k as $\lfloor b^k \rfloor$ in the statements and calculations that follow. For each $k = 1, \dots, |\mathcal{T}|$, recall the definition of the vector $\theta^{(k)} \in \mathbb{R}^n$,

$$\theta_t^{(k)} := \begin{cases} \left(\frac{n-b^k}{nb^k}\right)^{1/2} & \text{if } t \leq b^k \\ -\left(\frac{b^k}{n(n-b^k)}\right)^{1/2} & \text{if } t > b^k. \end{cases}$$

Then the following are true for each $k, l \leq |\mathcal{T}|$, and $Z \in \mathbb{R}^n$.

1. $\langle Z, \theta^{(k)} \rangle = \sqrt{\frac{b^k(n-b^k)}{n}} \left(\overline{Z}_{1:b^k} - \overline{Z}_{b^k+1:n}\right)$.
2. $\|\theta^{(k)}\| = 1$.
3. $\langle \theta^{(k)}, \theta^{(l)} \rangle \leq b^{-\frac{|k-l|}{2}}$.

Proof. For the first identity, apply the definition of $\theta^{(k)}$ to obtain,

$$\begin{aligned} \langle Z, \theta^{(k)} \rangle &= \left(\left(\frac{n-b^k}{nb^k}\right)^{1/2} \sum_{i \leq b^k} Z_i - \left(\frac{b^k}{n(n-b^k)}\right)^{1/2} \sum_{i > b^k} X_i\right) \\ &= \left(\frac{b^k(n-b^k)}{n}\right)^{1/2} \cdot \left(b^{-k} \sum_{i \leq b^k} Z_i - (n-b^k)^{-1} \sum_{i > b^k} Z_i\right) \\ &= \sqrt{\frac{b^k(n-b^k)}{n}} \left(\overline{Z}_{1:b^k} - \overline{Z}_{b^k+1:n}\right). \end{aligned}$$

For the second and third results, we compute the inner product for arbitrary $k, l \leq |\mathcal{T}|$.

Suppose $k \leq l$. Then the inner product $\langle \theta^{(k)}, \theta^{(l)} \rangle$ is equal to,

$$
\begin{aligned}
&= b^k \left(\frac{n - b^k}{nb^k} \cdot \frac{n - b^l}{nb^l} \right)^{1/2} - (b^l - b^k) \left(\frac{b^k}{n(n - b^k)} \cdot \frac{n - b^l}{nb^l} \right)^{1/2} \\
&+ (n - b^l) \left(\frac{b^k}{n(n - b^k)} \cdot \frac{b^l}{n(n - b^l)} \right)^{1/2} \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n - b^k)(n - b^l))^{1/2} \left[b^k \cdot \frac{1}{b^{k+l}} - (b^l - b^k) b^{\frac{k-l}{2} - \frac{k+l}{2}} \frac{1}{n - b^k} + \frac{n - b^l}{(n - b^k)(n - b^l)} \right] \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n - b^k)(n - b^l))^{1/2} \left[\frac{1}{b^l} - (b^l - b^k) b^{-l} \frac{1}{n - b^k} + \frac{1}{n - b^k} \right] \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n - b^k)(n - b^l))^{1/2} \left[\frac{n - b^k}{b^l (n - b^k)} - \frac{b^l - b^k}{b^l (n - b^k)} + \frac{b^l}{b^l (n - b^k)} \right] \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n - b^k)(n - b^l))^{1/2} \left[\frac{n}{b^l (n - b^k)} \right] \\
&= b^{\frac{k-l}{2}} \cdot \left(\frac{n - b^l}{n - b^k} \right)^{1/2}.
\end{aligned}
$$

An identical calculation gives a symmetric expression when $l \leq k$. Putting them together gives,

$$
\langle \theta^{(k)}, \theta^{(l)} \rangle = b^{-\frac{|k-l|}{2}} \left(\frac{n - b^{k \vee l}}{n - b^{k \wedge l}} \right)^{1/2} \leq b^{-\frac{|k-l|}{2}},
$$

which yields the third claim. Taking $k = l$ in the first equality gives $\|\theta^{(k)}\| = 1$.

□

Lemma 3.6.2. *Consider the simple vs simple testing problem,*

$$
H_0 : X \sim f_0 \quad \text{versus} \quad H_1 : X \sim f_1.
$$

The likelihood ratio test is defined $\psi(x) = 1_{\left\{\frac{f_1}{f_0}(x) > c\right\}}$ *for some constant* $c > 0$. *The sum of type 1 and 2 errors is bounded below,*

$$
\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \geq \frac{1}{k} \mathbb{P}_1(A_k),
$$

for any $k \geq 1$, *where* $A_k := \left\{ \frac{f_1}{f_0}(X) \leq k \right\}$, *and* $\mathbb{P}_0, \mathbb{P}_1$ *are the probability measures associated with* f_0, f_1.

Proof. The type 1 error can be lower bounded,

$$\mathbb{P}_0(\psi = 1) \geq \mathbb{P}_0(\{\psi = 1\} \cap A_k) \geq \mathbb{E}_0\left(\frac{1}{k} \cdot \frac{f_1}{f_0}(X) 1_{\{\psi=1\}\cap A_k}\right) = \frac{\mathbb{P}_1(\{\psi = 1\} \cap A_k)}{k}.$$

The type 2 error is trivially lower bounded,

$$\mathbb{P}_1(\psi = 0) \geq \frac{\mathbb{P}_1(\{\psi = 0\} \cap A_k)}{k}.$$

Summing these two inequalities gives the result. □

Lemma 3.6.3. *For $k \neq k^*$, we have*

$$\bar{\mathbb{E}}_1(L_k \mid k^*) = 1 + O(e^{-\frac{1}{2}p^a(1+o(1))}),$$

where k^* is defined in (3.27) and $L_k \coloneqq \prod_{j=1}^p \left[1 - \bar{\varepsilon} + \bar{\varepsilon} e^{\rho\langle X_{j:}, \theta^{(k)}\rangle - \rho^2/2}\right]$.

Proof. Recall that

$$L_k = \prod_{j=1}^p \left(1 - \varepsilon + \varepsilon e^{\rho\langle X_{j:}, \theta^{(k)}\rangle - \rho^2/2}\right), \quad X_{j:} \mid k^* = \begin{cases} \rho\theta^{(k^*)} + Z & \text{if } Q_j = 1 \\ Z & \text{if } Q_j = 0 \end{cases}$$

where $Z \sim N(0, I_n)$. Also by part 3 of Lemma 3.7.1, we have $\langle \theta^{(k)}, \theta^{(l)}\rangle \leq b^{-\frac{|k-l|}{2}}$, which is $\leq b^{-\frac{1}{2}}$ if $k \neq l$, where b is related to the grid size, and in this setting is $b \sim \log n \sim e^{p^a}$. Then for $k \neq k^*$, we have

$$\langle \theta^{(k)}, \theta^{(k^*)}\rangle \lesssim e^{-\frac{1}{2}p^a}. \tag{3.59}$$

Since conditional on k^*, the rows of $X \in \mathbb{R}^{p \times n}$ are independent, we have

$$\bar{\mathbb{E}}_1(L_k \mid k^*) = \prod_{j=1}^p \left(1 - \varepsilon + \varepsilon\left[(1-\varepsilon)\mathbb{E}e^{\rho Z_1 - \rho^2/2} + \varepsilon \mathbb{E}e^{\rho\langle \rho\theta^{(k^*)} + Z, \theta^{(k)}\rangle - \rho^2/2}\right]\right),$$

where the expectation is now taken with respect to $Z \sim N(0, I_n)$, and $Z_1 \sim N(0,1)$. The above is equal to

$$= \prod_{j=1}^p \left(1 - \varepsilon + \varepsilon\left[(1-\varepsilon)\cdot 1 + \varepsilon \mathbb{E}e^{\rho^2\langle\theta^{(k)}, \theta^{(k^*)}\rangle + \rho Z_1 - \rho^2/2}\right]\right),$$

where we have used that $\langle \theta^{(k)}, Z\rangle \sim N(0,1)$ by part 2 of Lemma 3.7.1. Using (3.59), the above becomes

$$= \prod_{j=1}^{p} \left(1 - \varepsilon + \varepsilon \left[(1-\varepsilon) \cdot 1 + \varepsilon \cdot e^{\rho^2 \langle \theta^{(k)}, \theta^{(k^*)} \rangle}\right]\right)$$

$$= \left(1 - \varepsilon + \varepsilon - \varepsilon^2 + \varepsilon^2 e^{\rho^2 \langle \theta^{(k)}, \theta^{(k^*)} \rangle}\right)^p \sim \exp\left(p\varepsilon^2 \rho^2 e^{-\frac{1}{2}p^a}\right) = 1 + O(e^{-\frac{1}{2}p^a(1+o(1))})$$

□

Lemma 3.6.4. *Let $p_{(1)}, \ldots, p_{(n)}$ be the order statistics of n uniform(0,1) random variables. Then*

$$p_{(k)} \sim \mathit{Beta}(k, n-k+1).$$

Proof. If $X_1, \ldots, X_n \sim F$ are iid from a continuous CDF F, then the density of the k^{th} order statistic is

$$f_{(k)}(x) = nf(x)\binom{n-1}{k-1} F(x)^{k-1}(1-F(x))^{n-k}.$$

Plug in $f(x) = \mathbf{1}\{x \in [0,1]\}$ to obtain the Beta$(k, n-k+1)$ density. □

Lemma 3.6.5. *Let $p_{(1)}, \ldots, p_{(n)}$ be the order statistics of $n > 2$ many uniform(0,1) random variables, and put $K(x,t) \coloneqq x \log \frac{x}{t} + (1-x)\log\frac{1-x}{1-t}$. Then for any $s > 0$, and $j \in \{2, \ldots, n\}$,*

$$\mathbb{P}(nK(j/n, p_{(j)}) > s) \le e\sqrt{2}je^{-(1-1/j)s}.$$

We also have

$$\mathbb{P}(nK(1/n, p_{(1)}) > s) \le (1 + 9/e)e^{-s}.$$

Proof. First suppose $j \ge 2$. For any $\lambda \in (0, 1-1/j]$, Markov's inequality implies

$$\begin{aligned} \mathbb{P}(nK(j/n, p_{(j)}) > s) &\le e^{-\lambda s}\mathbb{E}\exp(\lambda n K(j/n, p_{(j)})) \\ &= e^{-\lambda s}\mathbb{E}\exp\left(\lambda j \log\frac{j/n}{p_{(j)}} + \lambda(n-j)\log\frac{(n-j)/n}{1-p_{(j)}}\right) \\ &= e^{-\lambda s}(j/n)^{\lambda j}(1-j/n)^{\lambda(n-j)}\mathbb{E}p_{(j)}^{-\lambda j}(1-p_{(j)})^{-\lambda(n-j)}. \end{aligned}$$

By Lemma 3.6.4, $p_{(j)} \sim$ Beta$(j, n-j+1)$. The above becomes

$$= e^{-\lambda s}(j/n)^{\lambda j}(1-j/n)^{\lambda(n-j)}\frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} \cdot \frac{\Gamma(j(1-\lambda))\Gamma((n-j)(1-\lambda)+1)}{\Gamma(n(1-\lambda)+1)}.$$

Note that $j(1-\lambda) \ge 1$ is equivalent to $\lambda \le 1 - 1/j$. Recall Stirling's approximation,

$$\sqrt{2\pi}\left(\frac{m}{e}\right)^m \sqrt{m} \le \Gamma(m+1) \le e\left(\frac{m}{e}\right)^m \sqrt{m}.$$

Then the bound on the tail probability becomes

$$\mathbb{P}(nK(j/n, p_{(j)}) > s) \le e^{-\lambda s}(j/n)^{\lambda j}(1-j/n)^{\lambda(n-j)}\frac{(n/e)^n\sqrt{n}}{(\frac{j-1}{e})^{j-1}\sqrt{j-1}(\frac{n-j}{e})^{n-j}\sqrt{n-j}}\times$$
$$\frac{(\frac{j(1-\lambda)-1}{e})^{j(1-\lambda)-1}\sqrt{j(1-\lambda)}(\frac{(n-j)(1-\lambda)}{e})^{(n-j)(1-\lambda)}\sqrt{(n-j)(1-\lambda)}}{(\frac{n(1-\lambda)}{e})^{n(1-\lambda)}\sqrt{n(1-\lambda)}}.$$

Since $j \ge 2$, we have $\sqrt{\frac{j}{j-1}} \le \sqrt{2}$, so the above is bounded by,

$$\begin{aligned}
&\le \sqrt{2}\cdot e^{-\lambda s}(j/n)^{\lambda j}(1-j/n)^{\lambda(n-j)}\frac{n^n(j(1-\lambda)-1)^{j(1-\lambda)-1}((n-j)(1-\lambda))^{(n-j)(1-\lambda)}}{(j-1)^{j-1}(n-j)^{n-j}(n(1-\lambda))^{n(1-\lambda)}}\\
&= \sqrt{2}\cdot e^{-\lambda s}(j/n)^{\lambda j}(1-j/n)^{\lambda(n-j)}\frac{n^n}{(j-1)^{j-1}(n-j)^{n-j}}\times\\
&\frac{(j(1-\lambda)-1)^{j(1-\lambda)}((n-j)(1-\lambda))^{(n-j)(1-\lambda)}}{(n(1-\lambda))^{n(1-\lambda)}}\cdot\frac{1}{j(1-\lambda)-1}\\
&= \sqrt{2}\cdot e^{-\lambda s}\frac{j^{\lambda j}(n-j)^{\lambda(n-j)-(n-j)}}{n^{\lambda n - n}}\cdot\frac{1}{(j-1)^{j-1}}\times\\
&\frac{(j-1/(1-\lambda))^{j(1-\lambda)}(n-j)^{(n-j)(1-\lambda)}}{n^{n(1-\lambda)}}\cdot\frac{1}{j(1-\lambda)-1}\\
&= \sqrt{2}\cdot e^{-\lambda s}\frac{j^{\lambda j}}{(j-1)^{j-1}}\cdot(j-1/(1-\lambda))^{j(1-\lambda)}\cdot\frac{1}{j(1-\lambda)-1}\\
&= \sqrt{2}\cdot e^{-\lambda s}j^{-j(1-\lambda)+1}\cdot\left(\frac{j}{j-1}\right)^{j-1}(j(1-\lambda)-1)^{j(1-\lambda)-1}\frac{1}{(1-\lambda)^{j(1-\lambda)}}\\
&\le \sqrt{2}e\cdot e^{-\lambda s}j^{-j(1-\lambda)+1}j^{j(1-\lambda)-1}\frac{1}{(1-\lambda)^{j(1-\lambda)}} \qquad \left(\left(\tfrac{j}{j-1}\right)^{j-1} < e\right)\\
&= \sqrt{2}e\cdot e^{-\lambda s}\cdot\frac{1}{(1-\lambda)^{j(1-\lambda)}}.
\end{aligned}$$

Since the above holds for $\lambda \in (0, 1-j^{-1}]$, let $\lambda = 1 - j^{-1}$, so the above is,

$$\sqrt{2}e\cdot e^{-\lambda s}\cdot\frac{1}{(1-\lambda)^{j(1-\lambda)}} = \sqrt{2}e\cdot je^{-(1-j^{-1})s},$$

as desired. When $j = 1$, we bound the tail probability by directly integrating the Beta density,

$$\begin{aligned}\mathbb{P}(nK(1/n, p_{(1)}) > s) &= \mathbb{P}\left(\log\frac{1/n}{p_{(1)}} + (n-1)\log\frac{(n-1)/n}{1-p_{(1)}} > s\right) \\ &= \mathbb{P}\left(\frac{1/n}{p_{(1)}}\left(\frac{(n-1)/n}{1-p_{(1)}}\right)^{n-1} > e^s\right) \\ &= \mathbb{P}\left(\frac{1/n}{p_{(1)}}\left(\frac{(n-1)/n}{1-p_{(1)}}\right)^{n-1} > e^s, (1-p_{(1)})^{n-1} > \frac{1}{2}\right) \qquad (3.60) \\ &+ \mathbb{P}\left(\frac{1/n}{p_{(1)}}\left(\frac{(n-1)/n}{1-p_{(1)}}\right)^{n-1} > e^s, (1-p_{(1)})^{n-1} \leq \frac{1}{2}\right). \qquad (3.61)\end{aligned}$$

(3.60) can be bounded as follows. By monotonicity,

$$\mathbb{P}\left(\frac{1/n}{p_{(1)}}\left(\frac{(n-1)/n}{1-p_{(1)}}\right)^{n-1} > e^s, (1-p_{(1)})^{n-1} > \frac{1}{2}\right) \leq \mathbb{P}\left(1/n\cdot e^{-s}\frac{(1-1/n)^{n-1}}{1/2} > p_{(1)}\right).$$

Since $p_{(1)} \sim \text{Beta}(1, n)$, and since $(1-1/n)^{n-1} < 1/2$ for $n > 2$, this probability is bounded by,

$$\mathbb{P}\left(1/n\cdot e^{-s}\frac{(1-1/n)^{n-1}}{1/2} > p_{(1)}\right) \leq \int_0^{\frac{1}{n}e^{-s}} n(1-x)^{n-1}dx \leq e^{-s},$$

since $(1-x)^{n-1} < 1$. To bound (3.61), notice $(1-p_{(1)})^{n-1} \leq \frac{1}{2}$ is equivalent to $1-\left(\frac{1}{2}\right)^{\frac{1}{n-1}} \leq p_{(1)}$. Then (3.61) is equal to

$$\begin{aligned}&= \mathbb{P}\left(\frac{1/n}{p_{(1)}}\left(\frac{(n-1)/n}{1-p_{(1)}}\right)^{n-1} > e^s, 1-2^{-\frac{1}{n-1}} \leq p_{(1)}\right) \\ &\leq \mathbb{P}\left(\frac{1/n}{1-2^{-\frac{1}{n-1}}}\cdot e^{-s}((n-1)/n)^{n-1} > (1-p_{(1)})^{n-1}\right) \\ &= \mathbb{P}\left(p_{(1)} > 1-\left(\frac{1/n}{1-2^{-\frac{1}{n-1}}}\cdot e^{-s}\right)^{\frac{1}{n-1}}\cdot\frac{n-1}{n}\right) \\ &= \int_L^1 n(1-x)^{n-1}dx,\end{aligned}$$

where $L \coloneqq 1 - \left(\frac{1/n}{1-2^{-\frac{1}{n-1}}} \cdot e^{-s}\right)^{\frac{1}{n-1}} \cdot \frac{n-1}{n}$. Evaluating the integral gives

$$\begin{aligned}\int_L^1 n(1-x)^{n-1}dx &= (1-L)^n \\ &= \left(\frac{1/n}{1-2^{-\frac{1}{n-1}}}\right)^{\frac{n}{n-1}} \cdot \left(\frac{n-1}{n}\right)^n \cdot e^{-\frac{sn}{n-1}} \\ &< \left(\frac{1/n}{1-2^{-\frac{1}{n-1}}}\right)^{\frac{n}{n-1}} \cdot \left(\frac{n-1}{n}\right)^n \cdot e^{-s}.\end{aligned}$$

Straightforward calculation will show that $\left(\frac{1/n}{1-2^{-\frac{1}{n-1}}}\right)^{\frac{n}{n-1}} \cdot \left(\frac{n-1}{n}\right)^n < 9/e$ for $n > 2$. Combining the bounds on (3.60) and (3.61), we obtain the desired estimate.

□

Lemma 3.6.6. *The grid $\mathcal{T}$ is defined*

$$\mathcal{T} \coloneqq \left\{\lfloor(1+\delta)^0\rfloor, \lfloor(1+\delta)^1\rfloor, \ldots, \lfloor(1+\delta)^{\log_{1+\delta}\frac{n}{2}}\rfloor, \lfloor n-(1+\delta)^{\log_{1+\delta}\frac{n}{2}}\rfloor, \ldots, \lfloor n-(1+\delta)^0\rfloor\right\},$$

with $\delta = \frac{1}{\log\log n} \to 0$. For $\theta \in \Theta_1^{\textit{2-side}}(p,n,\rho,s)$ with true changepoint location $t^* \le n/2$, and changepoint size $\rho > 0$ defined,

$$\rho \coloneqq \sqrt{\frac{t^*(n-t^*)}{n}}|\mu_{j1} - \mu_{j2}|, \qquad \text{(see definition of } \Theta_1^{\text{2-side}})$$

there exists an element $\tilde{t} \in \mathcal{T}$ with $\frac{t^*}{1+\delta} < \tilde{t} \le t^*$, for which

$$\sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}}\left|\left(\bar{X}_{j,1:\tilde{t}} - \bar{X}_{j,\tilde{t}+1:n}\right)\right| \overset{(d)}{=} |\mu + Z|, \qquad (Z \sim N(0,1))$$

where $\mu = (1+o(1))\rho$, for every non-null row j. For $\theta \in \Theta_1^{\textit{1-side}}(p,n,\rho,s)$ with true changepoint location $t^* \le n/2$ and changepoint size $\rho > 0$, there exists $\tilde{t} \in \mathcal{T}$ with $\frac{t^*}{1+\delta} < \tilde{t} \le t^*$, for which

$$\sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}}\left(\bar{X}_{j,1:\tilde{t}} - \bar{X}_{j,\tilde{t}+1:n}\right) \sim N(\mu, 1),$$

where $\mu = (1+o(1))\rho$, for every non-null row j.

Proof. First let $\theta \in \Theta_1^{\text{2-side}}(p,n,\rho,s)$ with true changepoint location $t^* \le n/2$. The existence of $\tilde{t}$ satisfying $\frac{t^*}{1+\delta} < \tilde{t} \le t^*$ follows from the definition of the grid $\mathcal{T}$ and the assumption $t^* \le n/2$. Since the normalized contrast $\sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}}\left(\bar{X}_{j,1:\tilde{t}} - \bar{X}_{j,\tilde{t}+1:n}\right)$ is a linear combination of normal variables each with variance 1, it also has variance 1. Hence, it suffices to compute its mean under the alternative parameter θ. Let $j \in [p]$ denote the index of a non-null row in θ. Then since $\tilde{t} \leq t^*$, we have

$$\begin{aligned}\mathbb{E}_\theta \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \left(\bar{X}_{j,1:\tilde{t}} - \bar{X}_{j,\tilde{t}+1:n}\right) &= \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \left(\mu_{j1} - \frac{1}{n-\tilde{t}}(\mu_{j1}(t^*-\tilde{t}) + \mu_{j2}(n-t^*))\right) \\ &= \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \cdot \frac{n-t^*}{n-\tilde{t}} \cdot (\mu_{j1} - \mu_{j2}) \\ &= \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \cdot \frac{n-t^*}{n-\tilde{t}} \cdot \sqrt{\frac{n}{t^*(n-t^*)}} \cdot \rho \cdot \mathrm{sign}(\mu_{j1} - \mu_{j2}),\end{aligned}$$

from which it follows that

$$\sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \left|\bar{X}_{j,1:\tilde{t}} - \bar{X}_{j,\tilde{t}+1:n}\right| \overset{(d)}{=} \left|\sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \cdot \frac{n-t^*}{n-\tilde{t}} \cdot \sqrt{\frac{n}{t^*(n-t^*)}} \cdot \rho + Z\right|,$$

since the distribution of $Z \sim N(0,1)$ is symmetric. Then it suffices to show that

$$\lim_{n\to\infty} \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \cdot \frac{n-t^*}{n-\tilde{t}} \cdot \sqrt{\frac{n}{t^*(n-t^*)}} = 1. \tag{3.62}$$

Observe that $\tilde{t}$ and t^* satisfy $\frac{1}{1+\delta} < \frac{\tilde{t}}{t^*} \leq 1$, which implies the above limit, by the choice of $\delta = \frac{1}{\log\log n} \to 0$ and the assumption $t^* \leq n/2$.

For $\theta \in \Theta_1^{\text{1-side}}(p,n,\rho,s)$, the same calculation as above gives

$$\mathbb{E}_\theta \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \left(\bar{X}_{j,1:\tilde{t}} - \bar{X}_{j,\tilde{t}+1:n}\right) = \sqrt{\frac{\tilde{t}(n-\tilde{t})}{n}} \cdot \frac{n-t^*}{n-\tilde{t}} \cdot \sqrt{\frac{n}{t^*(n-t^*)}} \cdot \rho,$$

since $\rho = \sqrt{\frac{t^*(n-t^*)}{n}}(\mu_{j1} - \mu_{j2})$ for the non-null row with index j. Then the conclusion follows from (3.62). □

Lemma 3.6.7. *Let $x, t \in (0,1)$ and denote $K(x,t) := x\log\frac{x}{t} + (1-x)\log\frac{1-x}{1-t}$. The following inequalities hold,*

1. *$K(x,t) \geq 2(x-t)^2$ for all $x, t \in (0,1)$*
2. *$K(x,t) \geq \frac{(x-t)^2}{9t}$ when $\frac{x}{t} \leq 4$.*
3. *$K(x,t) \geq \frac{1}{2}x\log\frac{x}{t}$ when $e^2 \leq \frac{x}{t}$ and $x \leq \frac{1}{2}$.*

Proof. The first inequality follows from Pinsker's inequality,

$$\frac{1}{2}D(P\|Q) \geq \mathsf{TV}(P,Q)^2,$$

applied to $P = \text{Bern}(x)$ and $Q = \text{Bern}(t)$. The second inequality follows from the inequality,

$$D(P\|Q) \geq H^2(P, Q),$$

together with the identity $(\sqrt{x} - \sqrt{t})(\sqrt{x} + \sqrt{t}) = x - t$, since

$$\begin{aligned} H^2(P, Q) &:= \mathbb{E}_Q\big(\sqrt{P/Q} - 1\big)^2 \\ &= t(\sqrt{x/t} - 1)^2 + (1-t)(\sqrt{(1-x)/(1-t)} - 1)^2 \\ &\geq (\sqrt{x} - \sqrt{t})^2 \\ &= \frac{(x-t)^2}{(\sqrt{x} + \sqrt{t})^2} \\ &\geq \frac{(x-t)^2}{9t}. && (x \leq 4t) \end{aligned}$$

For the third inequality, using $\log(1-x) \geq -2x$ when $x < 1/2$, and the inequality $\log(1-t) \leq -t$ for all $t \in \mathbb{R}$, we have

$$\begin{aligned} (1-x)\log\frac{1-x}{1-t} &= (1-x)(\log(1-x) - \log(1-t)) \\ &\geq \frac{1}{2}(t - 2x) \\ &= -\frac{1}{2}x\left(2 - \frac{t}{x}\right). \end{aligned}$$

Using $e^2 \leq \frac{x}{t}$, we have $0 \leq 2 - \frac{t}{x} \leq 2 \leq \log\frac{x}{t}$, so that $-\frac{1}{2}x\left(2 - \frac{t}{x}\right) \geq -\frac{1}{2}x\log\frac{x}{t}$, which implies the result. □

3.7 Mixed Changepoint Detection Boundary

In this section, we consider another case of the high dimensional changepoint detection problem. The goal is still to detect whether a sparse fraction of rows in $X \in \mathbb{R}^{p\times n}$ have a changepoint location, but in the version studied below, the changepoint location is allowed to differ by row. By contrast, in the previous part of this chapter, the changepoint location is required to be the same among all non-null rows in X.

In Section 3.7.1, we state the detection boundary for the problem where the changepoint location can be different in each non-null row of X. Recall that the global null hypothesis is that each row of the mean matrix $\theta = \mathbb{E}(X)$ is constant, and the alternative hypothesis is that some sparse fraction of rows in θ have a changepoint.

More formally, the null hypothesis is again that all rows of the true mean matrix $\theta \in \mathbb{R}^{p\times n}$ are constant,

$$\Theta_0(p, n) := \big\{\theta \in \mathbb{R}^{p\times n} : \text{For all } j \leq p, \text{ there exists a } \mu_j \in \mathbb{R} \text{ s.t. } \theta_{ji} = \mu_j \text{ for all } i \leq n\big\}$$

Define the space of alternatives, $\Theta_1(p, a, r, \beta)$, as the following set

$$\Big\{\theta \in \mathbb{R}^{p\times n} : \frac{\#\text{non-null rows } j \in [p]}{p} = p^{-\beta(1+o(1))}, \text{ where } \frac{t_j(n-t_j)}{n}(\mu_{j1} - \mu_{j2})^2 = \rho(a,r)$$
$$\text{for non-null row } j \text{ with changepoint } t_j, \text{ where } \theta_{ji} = \mu_{j1} \text{ for } i \leq t_j, \theta_{ji} = \mu_{j2} \text{ for } i > t_j\Big\}.$$

In this problem, we consider the calibration

$$\begin{aligned}\log\log n &:= a \log p \\ \varepsilon &:= p^{-\beta} \\ \rho &:= \sqrt{2(1+a)r\log p} = \sqrt{2r(\log p + \log\log n)},\end{aligned}$$

where $\beta \in (0, 1), r \in (0, 1)$ and $a \geq 0$. Formally the hypothesis testing problem is

$$H_0 : \theta \in \Theta_0(p, n) \quad \text{versus} \quad H_1 : \theta \in \Theta_1(p, a, r, \beta), \tag{3.63}$$

and the minimax risk is defined

$$\mathcal{R}(p, a, r, \beta) := \inf_{\psi}\left[\sup_{\theta\in\Theta_0(p,n)} \mathbb{P}_0\psi + \sup_{\theta\in\Theta_1(p,a,r,\beta)} \mathbb{P}_\theta(1-\psi)\right].$$

3.7.1 Main result

The goal of this section is to establish the detection boundary,

$$\beta^*(a, r) = \begin{cases} 1 + (1+a)\left(r - \frac{1}{2}\right) & r \leq \frac{1}{4}, \\ 1 - (1+a)(1-\sqrt{r})_+^2 & r > \frac{1}{4}. \end{cases}$$

We may interpret the detection boundary as a phase transition in the set of triples (a, r, β), above which all tests are asymptotically powerless, and below which there exist tests with asymptotically negligible error. More precisely, we will show the following.

Theorem 3.7.1. *For the testing problem (3.63),*

1. *(Lower Bound) If $\beta > \beta^*(a, r)$, then $\mathcal{R}(p, a, r, \beta) \to 1$.*
2. *(Upper Bound) If $\beta < \beta^*(a, r)$, then $\mathcal{R}(p, a, r, \beta) \to 0$,*

as $p \to \infty$.

3.7.2 Proof of Theorem 3.7.1

Lower Bound

Throughout the subsequent calculations, we will ignore polylog factors (in p) for clarity, as they behave as constants compared to the polynomial factors they multiply. Let $b := \log n$

and define the grid of candidate changepoints,

$$\mathcal{T} := \left\{ \lfloor b^1 \rfloor, \lfloor b^2 \rfloor, \dots, \lfloor b^{\log_{\log n} n} \rfloor \right\}$$

Define the prior $\pi_0 := \delta_0$, and π_1 as

$$\theta_{j:} \overset{\text{iid}}{\sim} (1-\varepsilon)\delta_0 + \varepsilon \cdot \text{Unif}\left\{ \rho\theta^{(1)}, \dots, \rho\theta^{(|\mathcal{T}|)} \right\},$$

where $\theta^{(k)} \in \mathbb{R}^n$ is the unique vector for which $\langle Z, \theta^{(k)} \rangle$ is the normalized difference in sample means over the constant regions defined by changepoint k, i.e.

$$\langle Z, \theta^{(k)} \rangle = \sqrt{\frac{b^k(n-b^k)}{n}} \left(\overline{Z}_{1:b^k} - \overline{Z}_{b^k+1:n} \right),$$

and ρ is the signal defined above. More explicitly,

$$\theta_t^{(k)} := \begin{cases} \left(\frac{n-b^k}{nb^k} \right)^{1/2} & \text{if } t \leq b^k \\ -\left(\frac{b^k}{n(n-b^k)} \right)^{1/2} & \text{if } t > b^k. \end{cases}$$

Lemma (3.7.1) summarizes some useful properties of the collection $\left(\theta^{(k)}\right)_{k \leq |\mathcal{T}|}$. Lemma (3.7.2) shows that π_1 is supported on $\Theta_1(p, a, r, \beta)$. Since the average is always less than the supremum, the minimax testing risk is lower bounded,

$$\mathcal{R}(p, a, r, \beta) \geq \inf_{\psi} \left[\mathbb{P}_{\pi_0}\psi + \mathbb{P}_{\pi_1}(1-\psi) \right] \geq 1 - \mathsf{TV}(\mathbb{P}_{\pi_0}, \mathbb{P}_{\pi_1}),$$

by the Neyman Pearson lemma. Since the total variation distance and Hellinger distance are equivalent, it suffices to show $\beta > \beta^*(a, r)$ implies that the Hellinger distances goes to zero. By the tensorization property of the Hellinger distance, and the product form the null and alternative hypotheses, this is equivalent to

$$H^2(P_0, (1-\varepsilon)P_0 + \varepsilon P_1) = o(p^{-1}),$$

where $P_0 := N(0, I_n)$, and $P_1 := \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} N(\rho\theta^{(k)}, I_n)$. Define the splitting set

$$D := \left\{ \max_{k \leq |\mathcal{T}|} \langle X, \theta^{(k)} \rangle \leq \sqrt{2(1+a)\log p} \right\},$$

where we have abused notation by letting $X \in \mathbb{R}^n$ denote a generic row of the data matrix. Then

$$\begin{aligned}
H^2(P_0, (1-\varepsilon)P_0 + \varepsilon P_1) &= 1 - \int \sqrt{p_0((1-\varepsilon)p_0 + \varepsilon p_1)} \\
&= 1 - \mathbb{E}_{P_0} \sqrt{1 + \varepsilon \left(\frac{p_1}{p_0} - 1 \right)} \\
&\leq 1 - \mathbb{E}_{P_0} \sqrt{1 + \varepsilon \left(\frac{p_1}{p_0} \mathbf{1}_D - 1 \right)} \\
&\leq -\frac{1}{2} \varepsilon \mathbb{E}_{P_0} \left(\frac{p_1}{p_0} \mathbf{1}_D - 1 \right) + \varepsilon^2 \mathbb{E}_{P_0} \left(\frac{p_1}{p_0} \mathbf{1}_D - 1 \right)^2,
\end{aligned}$$

using the inequality $\sqrt{1+t} \geq 1 + t/2 - t^2$ for all $t \in \mathbb{R}$. The above becomes

$$\begin{aligned}
&\leq \frac{1}{2} \varepsilon P_1(D^c) + \varepsilon^2 \mathbb{E}_{P_0} \left(\frac{p_1}{p_0} \mathbf{1}_D - 1 \right)^2 \\
&= \frac{1}{2} \varepsilon P_1(D^c) + \varepsilon^2 \left(\mathbb{E}_{P_0} \left(\frac{p_1}{p_0} \right)^2 \mathbf{1}_D - 2P_1(D) + 1 \right) \\
&\asymp \varepsilon P_1(D^c) + \varepsilon^2 \left(\mathbb{E}_{P_0} \left(\frac{p_1}{p_0} \right)^2 \mathbf{1}_D - P_1(D) \right).
\end{aligned}$$

Based on the above calculation, the problem is now reduced to showing that $\beta > \beta^*(a, r)$ implies

$$p \varepsilon P_1(D^c) = o(1) \tag{3.64}$$

$$p \varepsilon^2 \left(\mathbb{E}_{P_0} \left(\frac{p_1}{p_0} \right)^2 \mathbf{1}_D - P_1(D) \right) = o(1). \tag{3.65}$$

We first compute the first order term. Letting Z denote a generic standard normal vector, we have

$$\begin{aligned}
p\varepsilon P_1(D^c) &= p\varepsilon \frac{1}{|\mathcal{T}|}\sum_{l=1}^{|\mathcal{T}|} \mathbb{P}_{N(\rho\theta^{(l)},I_n)}\left(\max_{k\le|\mathcal{T}|}\langle X,\theta^{(k)}\rangle > \sqrt{2(1+a)\log p}\right)\\
&\le p\varepsilon \frac{1}{|\mathcal{T}|}\sum_{l=1}^{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|} \mathbb{P}\left(\langle\rho\theta^{(l)}+Z,\theta^{(k)}\rangle > \sqrt{2(1+a)\log p}\right)\\
&\le p\varepsilon \frac{1}{|\mathcal{T}|}\sum_{l=1}^{|\mathcal{T}|}\left(|\mathcal{T}|p^{-(1+a)} + p^{-(1+a)(1-\sqrt{r})_+^2}\right) && \text{(Mills ratio)}\\
&= p\varepsilon\left(p^{-1} + p^{-(1+a)(1-\sqrt{r})_+^2}\right)\\
&= p^{-\beta} + p^{1-\beta-(1+a)(1-\sqrt{r})_+^2},
\end{aligned}$$

which is $o(1)$ so long as $\beta > 1-(1+a)(1-\sqrt{r})_+^2$. Since

$$1+(1+a)\left(r-\frac{1}{2}\right) \ge 1-(1+a)(1-\sqrt{r})_+^2$$

for all $r\in(0,1)$, for every $a\ge 0$, (see `https://www.desmos.com/calculator/y85orkn7hm`) (3.64) is implied by $\beta > \beta^*(a,r)$.

To show (3.65), we first compute the second moment of the likelihood ratio restricted to D. First note we can write it as

$$\mathbb{E}_{P_0}\left(\frac{p_1}{p_0}\right)^2 \mathbf{1}_D = \int \frac{p_1^2}{p_0}\mathbf{1}_D = \mathbb{E}_{P_1}\frac{p_1}{p_0}\mathbf{1}_D = \mathbb{E}_{P_1}\frac{1}{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|}\frac{\varphi(X-\rho\theta^{(k)})}{\varphi(X)}\mathbf{1}_D.$$

Elementary calculations show $\frac{\varphi(X-\rho\theta^{(k)})}{\varphi(X)} = \exp\left(\rho\langle X,\theta^{(k)}\rangle - \rho^2/2\right)$. Then the above is equal to

$$
\begin{aligned}
&= \frac{1}{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|} \mathbb{E}_{P_1} \exp\left(\rho\langle X, \theta^{(k)}\rangle - \rho^2/2\right) \mathbf{1}_D \\
&= \frac{1}{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|} \frac{1}{|\mathcal{T}|}\sum_{l=1}^{|\mathcal{T}|} \int (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|x-\rho\theta^{(l)}\|^2 + \rho\langle x, \theta^{(k)}\rangle - \rho^2/2\right) \mathbf{1}_D \\
&= \frac{1}{|\mathcal{T}|}\sum_{k=1}^{|\mathcal{T}|} \frac{1}{|\mathcal{T}|}\sum_{l=1}^{|\mathcal{T}|} \int (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\|x\|^2 - 2\rho\langle x, \theta^{(l)}\rangle + \rho^2) + \rho\langle x, \theta^{(k)}\rangle - \rho^2/2\right) \mathbf{1}_D \\
&= \frac{1}{|\mathcal{T}|^2}\sum_{k,l\le|\mathcal{T}|} \int (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|x\|^2 + \rho\langle x, \theta^{(l)} + \theta^{(k)}\rangle - \rho^2\right) \mathbf{1}_D \\
&= \frac{1}{|\mathcal{T}|^2}\sum_{k,l\le|\mathcal{T}|} \int (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|x-\rho(\theta^{(k)} + \theta^{(l)})\|^2 + \frac{1}{2}\rho^2\|\theta^{(k)}+\theta^{(l)}\|^2 - \rho^2\right) \mathbf{1}_D \\
&= \frac{1}{|\mathcal{T}|^2}\sum_{k,l\le|\mathcal{T}|} e^{\rho^2\langle\theta^{(k)},\theta^{(l)}\rangle} \mathbb{P}_{N(\rho(\theta^{(k)}+\theta^{(l)}),I_n)}(D).
\end{aligned}
$$

We first estimate the diagonal terms. If $k=l$, then the corresponding summand is

$$
\begin{aligned}
&= e^{\rho^2}\mathbb{P}_{N(2\rho\theta^{(k)},I_n)}\left(\max_{m\le|\mathcal{T}|}\langle X, \theta^{(m)}\rangle \le \sqrt{2(1+a)\log p}\right) \\
&= p^{2(1+a)r}\mathbb{P}\left(\max_{m\le|\mathcal{T}|}\langle 2\rho\theta^{(k)} + Z, \theta^{(m)}\rangle \le \sqrt{2(1+a)\log p}\right) \\
&\le p^{2(1+a)r}\mathbb{P}\left(2\rho + \langle Z, \theta^{(k)}\rangle \le \sqrt{2(1+a)\log p}\right) \\
&= p^{2(1+a)r}\mathbb{P}\left(\langle Z, \theta^{(k)}\rangle \le \sqrt{2(1+a)\log p}(1-2\sqrt{r})\right) \\
&\lesssim \begin{cases} p^{2(1+a)r} & r \le \frac{1}{4} \\ p^{2(1+a)r}p^{-(1+a)(1-2\sqrt{r})^2} & r > \frac{1}{4}. \end{cases}
\end{aligned}
$$

For a non-diagonal term $k \neq l$, the corresponding summand is

$$\begin{aligned}
&= e^{\rho^2\langle\theta^{(k)},\theta^{(l)}\rangle}\mathbb{P}\left(\max_{m\leq|\mathcal{T}|}\langle\rho(\theta^{(k)}+\theta^{(l)})+Z,\theta^{(m)}\rangle \leq \sqrt{2(1+a)\log p}\right) \\
&\leq e^{\rho^2\langle\theta^{(k)},\theta^{(l)}\rangle}\mathbb{P}\left(\rho+\langle\rho\theta^{(l)}+Z,\theta^{(k)}\rangle \leq \sqrt{2(1+a)\log p}\right) \\
&= e^{\rho^2\langle\theta^{(k)},\theta^{(l)}\rangle}\mathbb{P}\left(\langle Z,\theta^{(k)}\rangle \leq \sqrt{2(1+a)\log p}(1-\sqrt{r})\{1+o(1)\}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Lemma (3.7.1) Part 3)} \\
&\leq 1+2(\rho\langle\theta^{(k)},\theta^{(l)}\rangle), \qquad\qquad\qquad\qquad\qquad (e^x \leq 1+2x \text{ as } x\to 0)
\end{aligned}$$

since probabilities are always less than 1, and since $k \neq l$ implies $\langle\theta^{(k)},\theta^{(l)}\rangle \lesssim (\log n)^{-1/2} = p^{-a/2}$, whereas ρ is logarithmic in p. Putting the diagonal terms together with the off diagonal terms, we have

$$\mathbb{E}_{P_0}\left(\frac{p_1}{p_0}\right)^2\mathbf{1}_D \lesssim \begin{cases} \frac{1}{|\mathcal{T}|}p^{2(1+a)r}+\frac{1}{|\mathcal{T}|^2}\sum_{k\neq l}(1+2\rho\langle\theta^{(k)},\theta^{(l)}\rangle) & r\leq\frac{1}{4} \\ \frac{1}{|\mathcal{T}|}p^{2(1+a)r-(1+a)(1-2\sqrt{r})^2}+\frac{1}{|\mathcal{T}|^2}\sum_{k\neq l}(1+2\langle\theta^{(k)},\theta^{(l)}\rangle) & r>\frac{1}{4}. \end{cases}$$

Using Lemma (3.7.1) Part 3 again, we have

$$\frac{1}{|\mathcal{T}|^2}\sum_{k\neq l}(1+2\rho\langle\theta^{(k)},\theta^{(l)}\rangle) \leq 1+2\rho\cdot\frac{1}{|\mathcal{T}|^2}\underbrace{\sum_{k\neq l}\langle\theta^{(k)},\theta^{(l)}\rangle}_{\lesssim b\cdot b^{-1/2}},$$

since the sum can be written as a sum of b geometric series, each summing to less than a constant multiple of $b^{-1/2}$ by Part 3 of the Lemma. Now since $|\mathcal{T}| \asymp \log n = p^a$, the above says

$$\frac{1}{|\mathcal{T}|^2}\sum_{k\neq l}(1+2\rho\langle\theta^{(k)},\theta^{(l)}\rangle) \leq 1+2\rho\cdot p^{-2a+\frac{a}{2}}.$$

Now since $\frac{a}{2}-2a < -a+2(1+a)r-(1+a)(1-2\sqrt{r})^2$ for all $r>\frac{1}{4}$, the restricted squared likelihood ratio is

$$\mathbb{E}_{P_0}\left(\frac{p_1}{p_0}\right)^2\mathbf{1}_D \lesssim \begin{cases} p^{-a+2(1+a)r}+1 & r\leq\frac{1}{4} \\ p^{-a+2(1+a)r-(1+a)(1-2\sqrt{r})^2}+1 & r>\frac{1}{4}. \end{cases}$$

We have already shown that

$$P_1(D^c) \leq p^{-1}+p^{-(1+a)(1-\sqrt{r})_+^2},$$

so that

$$P_1(D) \geq 1 - p^{-1} - p^{-(1+a)(1-\sqrt{r})_+^2}.$$

Putting it all together, (3.65) becomes $p\varepsilon^2 \left(\mathbb{E}_{P_0} \left(\frac{p_1}{p_0}\right)^2 \mathbf{1}_D - P_1(D) \right)$, which can be bounded up to constants by

$$\begin{aligned}
&\lesssim \begin{cases} p\varepsilon^2 \left(p^{-a+2(1+a)r} + 1 - 1 + p^{-1} + p^{-(1+a)(1-\sqrt{r})_+^2} \right) & r \leq \frac{1}{4} \\ p\varepsilon^2 \left(p^{-a+2(1+a)r-(1+a)(1-2\sqrt{r})^2} + 1 - 1 + p^{-1} + p^{-(1+a)(1-\sqrt{r})_+^2} \right) & r > \frac{1}{4} \end{cases} \\
&= \begin{cases} p^{1-2\beta-a+2(1+a)r} + p^{-2\beta} + p^{1-2\beta-(1+a)(1-\sqrt{r})_+^2} & r \leq \frac{1}{4} \\ p^{1-2\beta-a+2(1+a)r-(1+a)(1-2\sqrt{r})^2} + p^{-2\beta} + p^{1-2\beta-(1+a)(1-\sqrt{r})_+^2} & r > \frac{1}{4} \end{cases} \\
&= \begin{cases} p^{1-2\beta-a+2(1+a)r} + o(1) & r \leq \frac{1}{4} \\ p^{1-2\beta-a+2(1+a)r-(1+a)(1-2\sqrt{r})^2} + o(1) & r > \frac{1}{4} \end{cases},
\end{aligned}$$

since $\beta > \beta^*(a, r)$ implies

$$\beta > 1 - (1+a)(1-\sqrt{r})_+^2 \geq \frac{1}{2}\left(1 - (1+a)(1-\sqrt{r})_+^2\right).$$

The exponent in the $r \leq \frac{1}{4}$ case is negative when

$$\beta > 1 + (1+a)\left(r - \frac{1}{2}\right),$$

which is the detection boundary in this case. In the $r > \frac{1}{4}$ case, the exponent is negative when

$$\begin{aligned}
2\beta &> 1 - a + 2(1+a)r - (1+a)(1-2\sqrt{r})^2 \\
&= 1 - a - (1+a)(1 - 4\sqrt{r} + 2r) \\
&= 2 - 2(1+a)(1-\sqrt{r})^2 \\
&\geq 2 - 2(1+a)(1-\sqrt{r})_+^2.
\end{aligned}$$

Dividing by 2 gives the detection boundary in the $r > \frac{1}{4}$ case. This completes the proof of the lower bound.

Upper Bound

Define the higher criticism statistic HC_p,

$$\mathsf{HC}_p = \max_{q>0} \frac{S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}},$$

where S_q is defined

$$\begin{aligned}
S_q &:= \sum_{j=1}^{p} \sum_{l=1}^{\log n} \mathbf{1}\left\{ \max_{t \in \mathcal{T}_l'} A_{jt} > 2(1+a)q \log p \right\}, \\
A_{jt} &:= \frac{t(n-t)}{n} (\overline{X}_{1:t}^{j} - \overline{X}_{(t+1):n}^{j})^2 - 1 \\
\mathcal{T} &:= \left\{ \lfloor (1+\delta)^0 \rfloor, \lfloor (1+\delta)^1 \rfloor, \ldots, \lfloor (1+\delta)^{\log_{1+\delta} \frac{n}{2}} \rfloor, \lfloor n - (1+\delta)^{\log_{1+\delta} \frac{n}{2} + 1} \rfloor, \ldots, \lfloor n - (1+\delta)^0 \rfloor \right\},
\end{aligned}$$

for some small grid size $\delta > 0$, and $\mathcal{T}_l'$ are disjoint blocks of grid points in $\mathcal{T}$; specifically, they are $\{\mathcal{T}_l'\}_{l=1,\ldots,2\log_{\log n} \frac{n}{2}}$,

$$\begin{aligned}
\mathcal{T}_1' &= \{ \lfloor (1+\delta)^0 \rfloor, \lfloor (1+\delta)^1 \rfloor, \ldots, \lfloor (1+\delta)^{\log_{1+\delta} \log n} \rfloor \} \\
\mathcal{T}_2' &= \{ \lfloor (1+\delta)^{(\log_{1+\delta} \log n)+1} \rfloor, \ldots, \lfloor (1+\delta)^{2\log_{1+\delta} \log n} \rfloor \} \\
\mathcal{T}_3' &= \{ \lfloor (1+\delta)^{2(\log_{1+\delta} \log n)+1} \rfloor, \ldots, \lfloor (1+\delta)^{3\log_{1+\delta} \log n} \rfloor \} \\
&\vdots
\end{aligned}$$

and so on, so that $\mathcal{T}$ is the union $\mathcal{T} = \bigcup_i \mathcal{T}_i'$, and there are $\asymp \log n$ many blocks. Note that each $|\mathcal{T}_i'| \asymp \log_{1+\delta} \log n \asymp (\log \log n)^2$ assuming $\delta = \frac{1}{\log \log n}$. Throughout the proof we will use the notation $y_q := 2(1+a)q \log p$.

Type I Error

In Lemma (3.7.4), we have shown that

$$\mathrm{Var}_0 \left(\sum_{l=1}^{\log n} \mathbf{1}\{ \max_{t \in \mathcal{T}_l'} A_{jt} > y_q \} \right) \asymp \sum_{l=1}^{\log n} \mathrm{Var}_0 \left(\mathbf{1}\{ \max_{t \in \mathcal{T}_l'} A_{jt} > y_q \} \right).$$

By a gaussian tail bound, the right hand side is of order $(\log n) p^{-(1+a)q} = p^{a-(1+a)q}$, so that $\mathrm{Var}_0(S_q) \asymp p^{(1+a)(1-q)}$. Then, by definition,

$$\max_{q>0} \frac{S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \asymp \max_{q>0} \frac{\sum_{l=1}^{\log n} \sum_{j=1}^{p} \left(\mathbf{1}\{ \max_{t \in \mathcal{T}_l'} A_{jt} > y_q \} - \mathbb{P}_0(\max_{t \in \mathcal{T}_l'} A_{jt} > y_q) \right)}{p^{\frac{1}{2}(1+a)(1-q)}}.$$

The idea is to split the sum over l into a constant number of pieces m (independent of p), chosen during the analysis, so that the summands in each piece are nearly independent. Then we can use high probability bounds on the small gaussian representation terms (similar to what we have done in the proof of Lemma (3.7.4)) so that each piece behaves as a normalized uniform empirical process, and then estimate the error of this approximation. In this step, we will choose m so that the error tends to zero. To this end, the above is

$$\begin{aligned}
&= \max_{q>0} \sum_{i=1}^{m} \frac{\sum_{l\in\{i,i+m,i+2m,\dots\}} \sum_{j=1}^{p} \left(\mathbf{1}\left\{\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right\} - \mathbb{P}_0\left(\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right)\right)}{p^{\frac{1}{2}(1+a)(1-q)}} \\
&\le \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \max_{q>0} \frac{\sum_{l\in\{i,i+m,i+2m,\dots\}} \sum_{j=1}^{p} \left(\mathbf{1}\left\{\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right\} - \mathbb{P}_0\left(\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right)\right)}{m^{-1/2}\cdot p^{\frac{1}{2}(1+a)(1-q)}}
\end{aligned} \tag{3.66}$$

Since the analysis for each of these pieces is identical, we only analyze the first summand (corresponding to $i = 1$),

$$\frac{\sum_{l\in\{1,1+m,1+2m,\dots\}} \sum_{j=1}^{p} \left(\mathbf{1}\{\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\} - \mathbb{P}_0(\max_{t\in\mathcal{T}'_l} A_{jt} > y_q)\right)}{m^{-1/2}\cdot p^{\frac{1}{2}(1+a)(1-q)}}. \tag{3.67}$$

Suppose $t \in \mathcal{T}'_k$ where $k \in \{1, 1+m, 1+2m, \dots\}$. The goal now is to approximate the indicators in the double sum by indicators which are independent. Under H_0, we can represent A_{jt} as

$$A_{jt} = \left(Z_{jt} + \sum_{l\in\{1,1+m,1+2m,\dots\}\setminus\{k\}} \sum_{s\in\mathcal{T}'_l} Z_{jts} \right)^2 - 1,$$

where

$$\mathrm{Var}(Z_{jts}) = \mathrm{Cov}(\langle X_{j:}, \theta^{(t)}\rangle, \langle X_{j:}, \theta^{(s)}\rangle) \le b^{-\frac{|k-l|}{2}} := (\log n)^{-\frac{|k-l|}{2}} = p^{-a\frac{|k-l|}{2}}$$

(Lemma (3.7.1) Part 3)

$$\mathrm{Var}(Z_{jt}) = 1 - \sum_{l\in\{1,1+m,1+2m,\dots\}\setminus\{k\}} \sum_{s\in\mathcal{T}'_l} \mathrm{Var}(Z_{jts}),$$

and for each t, the $(Z_{jts})_{s\in\mathcal{T}'_l, l\in\{1,1+m,\dots\}\setminus\{k\}}$ are mutually independent, and the groups

$$\left\{(Z_{jt})_{t\in\mathcal{T}'_l}\right\}_{l\in\{1,1+m,1+2m,\dots\}}$$

are independent across l. Then since $|\mathcal{T}_l'| \asymp \log\log n$, we have that

$$\begin{aligned}
\sum_{l \in \{1, 1+m, 1+2m, \dots\} \setminus \{k\}} \sum_{s \in \mathcal{T}_l'} \operatorname{Var}(Z_{jts}) &\leq \sum_{l \in \{1, 1+m, 1+2m, \dots\} \setminus \{k\}} (\log\log n) b^{-\frac{|k-l|}{2}} \\
&= \mathsf{PL} \cdot \sum_{l \in \{1, 1+m, 1+2m, \dots\} \setminus \{k\}} b^{-\frac{|k-l|}{2}} \\
&\lesssim \mathsf{PL} \cdot \sum_{i=1}^{\infty} (b^{-\frac{1}{2}})^{mi} \\
&= \mathsf{PL} \cdot b^{-\frac{m}{2}} \sum_{i=0}^{\infty} (b^{-\frac{1}{2}})^{mi} \\
&\lesssim \mathsf{PL} \cdot b^{-\frac{m}{2}},
\end{aligned}$$

since the sum is a geometric series. This calculation implies that with high probability,

$$\max_{k \in \{1, 1+m, 1+2m, \dots\}} \left| \sum_{l \in \{1, 1+m, 1+2m, \dots\} \setminus \{k\}} \sum_{s \in \mathcal{T}_l'} Z_{jts} \right| \lesssim \mathsf{PL} \cdot b^{-\frac{m}{4}} \sqrt{\log \frac{b}{m}} = \mathsf{PL} \cdot b^{-\frac{m}{4}},$$

since $|\{1, 1+m, 1+2m, \dots\}| \asymp \frac{b}{m}$. Thus on this high probability event, (3.67) is bounded by

$$\begin{aligned}
&\leq \frac{1}{m^{-1/2} \cdot p^{\frac{1}{2}(1+a)(1-q)}} \sum_{l \in \{1, 1+m, 1+2m, \dots\}} \sum_{j=1}^{p} \Big(\mathbf{1}\{\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}\} \\
&- \mathbb{P}_0(\max_{t \in \mathcal{T}_l'} A_{jt} > y_q) \Big) \\
&\leq \frac{1}{m^{-1/2} \cdot p^{\frac{1}{2}(1+a)(1-q)}} \sum_{l \in \{1, 1+m, 1+2m, \dots\}} \sum_{j=1}^{p} \Big(\mathbf{1}\{\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}\} \\
&- \mathbb{P}_0(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q) \Big),
\end{aligned}$$

where the second inequality follows since $\mathbb{P}_0(\max_{t \in \mathcal{T}_l'} A_{jt} > y_q) \geq \mathbb{P}_0(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q)$. Adding and subtracting the null expectation $\mathbb{P}_0 \left(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}} \right)$ in the differences gives

$$= \frac{1}{(m^{-1/2} \cdot p^{\frac{1}{2}(1+a)(1-q)})} \sum_{l \in \{1, 1+m, 1+2m, \dots\}} \sum_{j=1}^{p} \Big(\mathbf{1}\{\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}\} \tag{3.68}$$

$$- \mathbb{P}_0(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}) \Big)$$

$$+ \frac{1}{m^{-1/2} \cdot p^{\frac{1}{2}(1+a)(1-q)}} \sum_{l \in \{1, 1+m, 1+2m, \dots\}} \sum_{j=1}^{p} \Big(\mathbb{P}_0(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}) \tag{3.69}$$

$$- \mathbb{P}_0(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q) \Big).$$

The first term above is equal in distribution to a normalized uniform empirical process, and we will choose m so that the second term is negligible. The difference is

$$\mathbb{P}_0\left(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}\right) - \mathbb{P}_0\left(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q\right)$$
$$= \mathbb{P}\left(y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}} < \max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 \le y_q\right)$$

Monotonicity yields,

$$\mathbb{P}\left(y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}} < \max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 \le y_q\right) \le \mathbb{P}\left(\bigcup_{t \in \mathcal{T}_l'} \left\{y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}} < Z_{jt}^2 - 1 \le y_q\right\}\right)$$
$$\le \sum_{t \in \mathcal{T}_l'} \mathbb{P}\left(y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}} < Z_{jt}^2 - 1 \le y_q\right). \tag{3.70}$$

Let f_{jt} denote the density of $Z_{jt}^2 - 1$, which is bounded because $\mathrm{Var}(Z_{jt}) \to 1$. As a result, each summand can be written

$$\mathbb{P}(y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}} < Z_{jt}^2 - 1 \le y_q) = \int_{y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}}^{y_q} f_{jt}(z) dz \lesssim \mathsf{PL} \cdot b^{-\frac{m}{4}}.$$

The above estimate with (3.70) imply that

$$\mathbb{P}_0\left(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q - \mathsf{PL} \cdot b^{-\frac{m}{4}}\right) - \mathbb{P}_0\left(\max_{t \in \mathcal{T}_l'} Z_{jt}^2 - 1 > y_q\right) \lesssim \mathsf{PL} \cdot b^{-\frac{m}{4}},$$

since $|\mathcal{T}_l'| \asymp \log\log n \in \mathsf{PL}$. Plugging this estimate back into (3.69) gives

$$\begin{aligned}
&\asymp \text{normalized empirical process} + \frac{\sum_{l\in\{1,1+m,1+2m,\dots\}} \sum_{j=1}^{p} \mathsf{PL}\cdot b^{-\frac{m}{4}}}{m^{-1/2}\cdot p^{\frac{1}{2}(1+a)(1-q)}} \\
&\asymp \text{normalized empirical process} + \mathsf{PL}\cdot p^{1+a}\cdot p^{-a\frac{m}{4}}\cdot p^{-\frac{1}{2}(1+a)(1-q)}.
\end{aligned}$$

From here it is clear that the condition

$$1+a-\frac{am}{4}-\frac{1}{2}(1+a)(1-q)\le 0$$

is satisfied when

$$m \ge \frac{4(1+a)}{a} \ge \frac{4}{a}\left(1+a-\frac{1}{2}(1+a)(1-q)\right).$$

Hence we may choose $m=\frac{4(1+a)}{a}$ so that the above condition is satisfied. It now follows that (3.66) is the sum of m many normalized empirical processes, and so grows sub polynomially, so the test

$$\psi(X) := \mathbf{1}\{\mathsf{HC}_p > C\sqrt{\log\log p}\}$$

has consistent Type I error for some universal constant $C>0$.

Type II error

By Lemma (3.7.5), it suffices to show that for any $\theta\in\Theta_1$, there exists some $q>0$ for which the following two conditions hold,

$$\frac{\mathrm{Var}_\theta(S_q)}{(\mathbb{E}_\theta S_q)^2} \to 0 \tag{3.71}$$

$$\frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \to \infty \text{ polynomially.} \tag{3.72}$$

By independence across $j\in[p]$, we have

$$\mathrm{Var}_\theta(S_q) = \sum_{j=1}^{p} \mathrm{Var}_\theta\left(\sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\right).$$

By Lemma (3.7.6), we have

$$\begin{aligned}
\sum_{j=1}^{p} \mathrm{Var}_\theta\left(\sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right\}\right) &\lesssim \sum_{j=1}^{p}\sum_{l=1}^{\log n} \mathrm{Var}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right\}\right) + o(1) \\
&\leq \sum_{j=1}^{p}\sum_{l=1}^{\log n} \mathbb{P}_\theta\left(\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right) + o(1) \\
&= \mathbb{E}_\theta S_q + o(1).
\end{aligned}$$

Thus, to show that (3.71) holds, it suffices to choose q so that $\mathbb{E}_\theta S_q \to \infty$. For the j^{th} row of $\theta \in \mathbb{R}^{p\times n}$, if it is non-null, let $t_j^* \leq n$ denote the true changepoint. By definition of the grid $\mathcal{T}$, there exists some $\tilde{t}_j \in \mathcal{T}$ for which

$$\frac{t_j^*}{1+\delta} < \tilde{t}_j \leq t_j^*. \tag{3.73}$$

Assume without loss of generality that $t_j^* \leq \frac{n}{2}$; by symmetry of the grid $\mathcal{T}$, an analagous argument can be made for $t_j^* > \frac{n}{2}$. Let $\tilde{\mathcal{T}}'_j$ denote the block of size $\asymp \log\log n$ that contains $\tilde{t}_j$. Then by definition of S_q,

$$\begin{aligned}
\mathbb{E}_\theta S_q &= \mathbb{E}_\theta \sum_{j=1}^{p}\sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right\} \\
&= \sum_{\{\text{non-null } j\}}\sum_{l=1}^{\log n} \mathbb{P}_{\theta_{j:}}\left(\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right) + \sum_{\{\text{null } j\}}\sum_{l=1}^{\log n} \mathbb{P}_0\left(\max_{t\in\mathcal{T}'_l} A_{jt} > y_q\right). \quad (3.74)
\end{aligned}$$

Note that $A_{j\tilde{t}_j}$ is distributed like $(\mu + Z)^2 - 1$, where $Z \sim N(0,1)$, and

$$\begin{aligned}
\mu^2 &= \frac{\tilde{t}_j(n-\tilde{t}_j)}{n}\left(\mu_{j1} - \frac{1}{n-\tilde{t}_j}(\mu_{j1}(t_j^* - \tilde{t}_j) + \mu_{j2}(n - t_j^*))\right)^2 \\
&= \frac{\tilde{t}_j(n-\tilde{t}_j)}{n}\left(\frac{n-t_j^*}{n-\tilde{t}_j}(\mu_{j1} - \mu_{j2})\right)^2 \\
&= \frac{\tilde{t}_j(n-t_j^*)}{t_j^*(n-\tilde{t}_j)} \cdot \underbrace{\frac{t_j^*(n-t_j^*)}{n}(\mu_{j1} - \mu_{j2})^2}_{=\rho^2} \\
&\geq \frac{1}{1+\delta} \cdot \frac{n-t_j^*}{n-\tilde{t}_j} \cdot \rho^2 && \text{(by (3.73))} \\
&\geq \frac{1}{1+\delta} \cdot \frac{n-(1+\delta)\tilde{t}_j}{n-\tilde{t}_j} \cdot \rho^2 \\
&= \frac{1}{1+\delta} \cdot \left(1 - \frac{\delta\tilde{t}_j}{n-\tilde{t}_j}\right) \cdot \rho^2 \\
&\geq \frac{1-\delta}{1+\delta} \cdot \rho^2 && \left(\tilde{t}_j \leq \tfrac{n}{2} \Rightarrow \tfrac{\tilde{t}_j}{n-\tilde{t}_j} \leq 1\right)
\end{aligned}$$

Recalling $y_q := 2(1+a)q\log p$, we then have

$$\begin{aligned}
\mathbb{P}_\theta\left(\max_{t\in\tilde{\mathcal{T}}_j'} A_{jt} > y_q\right) &\geq \mathbb{P}_\theta\left(A_{j\tilde{t}_j} > y_q\right) \\
&\geq \mathbb{P}\left(\sqrt{\frac{1-\delta}{1+\delta}} \cdot \rho + Z > \sqrt{y_q}\right) \\
&= \mathbb{P}\left(N(0,1) > \sqrt{2(1+a)q\log p} - \sqrt{\frac{1-\delta}{1+\delta} \cdot 2(1+a)r\log p}\right) \\
&= \mathbb{P}\left(Z > \sqrt{2(1+a)\log p}\left(\sqrt{q} - \sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)\right).
\end{aligned}$$

For all other $\mathcal{T}_l'$, we have the trivial bound

$$\mathbb{P}_\theta\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right) \geq \mathbb{P}_0\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right).$$

Since there are $\asymp \log_{\log n} n$ (which, ignoring polylog factors in p, behaves like $\log n$) many blocks $\mathcal{T}_l'$, the left term in (3.74) becomes

$$\begin{aligned}
&\geq \sum_{\{\text{non-null } j\}} \left[\mathbb{P}\left(Z > \sqrt{2(1+a)\log p}\left(\sqrt{q} - \sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)\right)\right. \\
&\left. + \sum_{\mathcal{T}_l' : \mathcal{T}_l' \neq \tilde{\mathcal{T}}_j'} \mathbb{P}_0\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right)\right] \\
&\asymp \sum_{\{\text{non-null } j\}} \left[\mathbb{P}\left(Z > \sqrt{2(1+a)\log p}\left(\sqrt{q} - \sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)\right)\right. \\
&\left. + (\log n)\mathbb{P}_0\left(\max_{t\in\mathcal{T}_1'} A_{jt} > y_q\right)\right] \\
&\gtrsim \sum_{\{\text{non-null } j\}} \left[p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + (\log n)\mathbb{P}_0(A_{j1} > y_q)\right] \\
&\gtrsim \sum_{\{\text{non-null } j\}} \left[p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + p^{a-(1+a)q}\right].
\end{aligned}$$

Hence (3.74) becomes

$$\geq \sum_{\{\text{non-null } j\}} \left[p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + p^{a-(1+a)q}\right] + \sum_{\{\text{null } j\}} \sum_{l=1}^{\log n} \mathbb{P}_0(A_{jt} > y_q),$$

where t is any element of $\mathcal{T}$ (under the null, all A_{jt} have the same distribution). Applying the gaussian tail bound again, the above is lower bounded by

$$\begin{aligned}
&\geq \sum_{\{\text{non-null } j\}} \left[p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + p^{a-(1+a)q}\right] + \sum_{\{\text{null } j\}} \sum_{l=1}^{\log n} p^{-(1+a)q} \\
&= p\varepsilon \left[p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + p^{a-(1+a)q}\right] + p(1-\varepsilon)p^{a-(1+a)q} \\
&\asymp p^{1-\beta-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + p^{-\beta+(1+a)(1-q)} + p^{(1+a)(1-q)} - p^{-\beta+(1+a)(1-q)} \\
&= p^{1-\beta-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} + p^{(1+a)(1-q)}.
\end{aligned}$$

Letting $\delta \to 0$ as $\delta = \frac{1}{\log\log n}$, the above says

$$\mathbb{E}_\theta S_q \gtrsim p^{1-\beta-(1+a)(\sqrt{q}-\sqrt{r})_+^2(1+o(1))} + p^{(1+a)(1-q)}. \tag{3.75}$$

Hence condition (3.71) has been reduced to choosing a $q > 0$ for which the right hand side of the above tends to infinity. We now reduce condition (3.72) to a similar condition, since the choice of q needs to satisfy both. To this end, the difference in the numerator of (3.72) is

$$\begin{aligned}
\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q &= \Bigg[\mathbb{E}_\theta \sum_{\{\text{non-null } j\}} \sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\} \\
&\quad + \mathbb{E}_\theta \sum_{\{\text{null } j\}} \sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\Bigg] - \mathbb{E}_0 S_q \\
&= \sum_{\{\text{non-null } j\}} \sum_{l=1}^{\log n} \left[\mathbb{P}_{\theta_{j:}}\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right) - \mathbb{P}_0\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right)\right]
\end{aligned}$$

Under the alternative, $A_{jt} = (\mu_{jt} + Z_{jt})^2 - 1$ for some $\mu_{jt} \in \mathbb{R}$ and $Z_{jt} \sim N(0,1)$, whereas under the null, $A_{jt} = Z_{jt}^2 - 1$. Hence we have a trivial bound $\mathbb{P}_{\theta_{j:}}\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right) - \mathbb{P}_0\left(\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right) \geq 0$. For $\tilde{t}_j \in \tilde{\mathcal{T}}_j'$ (the closest grid point to the true changepoint t_j^*), monotonicity, a union bound, and the gaussian tail bound give

$$\begin{aligned}
\mathbb{P}_{\theta_{j:}}\left(\max_{t\in\tilde{\mathcal{T}}_j'} A_{jt} > y_q\right) - \mathbb{P}_0\left(\max_{t\in\tilde{\mathcal{T}}_j'} A_{jt} > y_q\right) &\geq \mathbb{P}_{\theta_{j:}}\left(A_{j\tilde{t}_j} > y_q\right) - (\log\log n)\mathbb{P}_0\left(A_{j\tilde{t}_j} > y_q\right) \\
&\gtrsim p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2} - \mathsf{PL}\cdot p^{-(1+a)q} \\
&\asymp p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2}
\end{aligned}$$

Combined with the trivial bound for all $\mathcal{T}_l'$ not equal to $\tilde{\mathcal{T}}_j'$, the above implies the following lower bound on the difference $\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q$,

$$\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q \gtrsim \sum_{\{\text{non-null } j\}} \left[p^{-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2}\right] \asymp p^{1-\beta-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2}$$

By Lemma (3.7.4), we have

$$\mathrm{Var}_0(S_q) \asymp \sum_{j=1}^{p} \sum_{l=1}^{\log n} \mathrm{Var}_0\left(\mathbf{1}\left\{\max_{t \in \mathcal{T}_l'} A_{jt} > y_q\right\}\right) \asymp p^{(1+a)(1-q)}.$$

Together with the lower bound on $\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q$, this gives

$$\frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \gtrsim p^{1-\beta-(1+a)\left(\sqrt{q}-\sqrt{\frac{1-\delta}{1+\delta}}\sqrt{r}\right)_+^2 - \frac{1}{2}(1+a)(1-q)}. \tag{3.76}$$

Letting $\delta \to 0$ as $\delta := \frac{1}{\log\log n}$, (3.75) and (3.76) translate the conditions (3.71) and (3.72) into

$$p^{1-\beta-(1+a)(\sqrt{q}-\sqrt{r})_+^2(1+o(1))} + p^{(1+a)(1-q)} \to \infty$$
$$p^{1-\beta-(1+a)(\sqrt{q}-\sqrt{r})_+^2(1+o(1))-\frac{1}{2}(1+a)(1-q)} \to \infty.$$

Taking $q = 1$ when $r > 1/4$ gives the requirement $\beta < 1 - (1+a)(1-\sqrt{r})_+^2$. Taking $q = 4r$ when $r \leq \frac{1}{4}$ gives the requirement $\beta < 1 + (1+a)\left(r - \frac{1}{2}\right)$. Hence we have shown that $\beta < \beta^*(a,r)$ implies a consistent Type II error for the Higher Criticism statistic.

3.7.3 Technical Lemmas

Lemma 3.7.1. *Put $b := \log n$. For each $k = 1, \ldots, |\mathcal{T}|$, recall the definition of the vector $\theta^{(k)} \in \mathbb{R}^n$,*

$$\theta_t^{(k)} := \begin{cases} \left(\frac{n-b^k}{nb^k}\right)^{1/2} & \text{if } t \leq b^k \\ -\left(\frac{b^k}{n(n-b^k)}\right)^{1/2} & \text{if } t > b^k. \end{cases}$$

Then the following are true for each $k, l \leq |\mathcal{T}|$, and $Z \in \mathbb{R}^n$.

1. $\langle Z, \theta^{(k)} \rangle = \sqrt{\frac{b^k(n-b^k)}{n}}\left(\overline{Z}_{1:b^k} - \overline{Z}_{b^k+1:n}\right)$.

2. $\|\theta^{(k)}\| = 1$.

3. $\langle \theta^{(k)}, \theta^{(l)} \rangle \leq b^{-\frac{|k-l|}{2}}$.

Proof. For the first identity, apply the definition to get,

$$\begin{aligned}
\langle Z, \theta^{(k)} \rangle &= \left(\left(\frac{n-b^k}{nb^k} \right)^{1/2} \sum_{i \le b^k} Z_i - \left(\frac{b^k}{n(n-b^k)} \right)^{1/2} \sum_{i > b^k} X_i \right) \\
&= \left(\frac{b^k(n-b^k)}{n} \right)^{1/2} \cdot \left(b^{-k} \sum_{i \le b^k} Z_i - (n-b^k)^{-1} \sum_{i > b^k} Z_i \right) \\
&= \sqrt{\frac{b^k(n-b^k)}{n}} \left(\overline{Z}_{1:b^k} - \overline{Z}_{b^k+1:n} \right).
\end{aligned}$$

For the second and third results, we compute the inner product for arbitrary $k, l \le |\mathcal{T}|$. Let $k \le l$ without loss of generality. Then

$$\begin{aligned}
\langle \theta^{(k)}, \theta^{(l)} \rangle &= b^k \left(\frac{n-b^k}{nb^k} \cdot \frac{n-b^l}{nb^l} \right)^{1/2} - (b^l - b^k) \left(\frac{b^k}{n(n-b^k)} \cdot \frac{n-b^l}{nb^l} \right)^{1/2} \\
&+ (n-b^l) \left(\frac{b^k}{n(n-b^k)} \cdot \frac{b^l}{n(n-b^l)} \right)^{1/2} \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n-b^k)(n-b^l))^{1/2} \left[\frac{1}{b^l} - (b^l - b^k) b^{-l} \frac{1}{n-b^k} + \frac{1}{n-b^k} \right] \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n-b^k)(n-b^l))^{1/2} \left[\frac{n-b^k}{b^l(n-b^k)} - \frac{b^l-b^k}{b^l(n-b^k)} + \frac{b^l}{b^l(n-b^k)} \right] \\
&= \frac{b^{\frac{k+l}{2}}}{n} ((n-b^k)(n-b^l))^{1/2} \left[\frac{n}{b^l(n-b^k)} \right] \\
&= b^{\frac{k-l}{2}} \cdot \left(\frac{n-b^l}{n-b^k} \right)^{1/2}.
\end{aligned}$$

In general, symmetry implies

$$\langle \theta^{(k)}, \theta^{(l)} \rangle = b^{-\frac{|k-l|}{2}} \left(\frac{n-b^{k \vee l}}{n-b^{k \wedge l}} \right)^{1/2} \le b^{-\frac{|k-l|}{2}},$$

which yields the third claim. Taking $k = l$ in the first equality gives $\|\theta^{(k)}\| = 1$.

□

Lemma 3.7.2. *The prior π_1 defined by*

$$\theta_{j:} \overset{\text{iid}}{\sim} (1-\varepsilon)\delta_0 + \varepsilon \cdot \mathit{Unif}\left\{ \rho\theta^{(1)}, \ldots, \rho\theta^{(|\mathcal{T}|)} \right\},$$

is supported on $\Theta_1(p, a, r, \beta)$.

Proof. Since b^k is the changepoint corresponding to the mean vector $\rho\theta^{(k)}$, the normalized signal is

$$\begin{aligned}
&= \rho^2 \cdot \frac{b^k(n-b^k)}{n}\left[\left(\frac{n-b^k}{nb^k}\right)^{1/2} + \left(\frac{b^k}{n(n-b^k)}\right)^{1/2}\right]^2 \\
&= \rho^2 \cdot \frac{b^k(n-b^k)}{n}\left[\left(\frac{n-b^k}{nb^k}\right) + 2\left(\frac{b^k}{n(n-b^k)}\right)^{1/2}\left(\frac{n-b^k}{nb^k}\right)^{1/2} + \left(\frac{b^k}{n(n-b^k)}\right)\right] \\
&= \rho^2 \cdot \frac{b^k(n-b^k)}{n}\left[\left(\frac{n-b^k}{nb^k}\right) + \frac{2}{n} + \left(\frac{b^k}{n(n-b^k)}\right)\right] \\
&= \rho^2 \cdot \left[\frac{(n-b^k)^2}{n^2} + \frac{2b^k(n-b^k)}{n^2} + \frac{(b^k)^2}{n^2}\right] \\
&= \rho^2.
\end{aligned}$$

To see that the fraction of non sparse rows is $p^{-\beta(1+o_{\mathbb{P}}(1))}$, it suffices to show that

$$\frac{\sum_{j=1}^p \mathbf{1}\{\theta_{j\cdot} \text{ has a changepoint}\}}{p} = p^{-\beta(1+o_{\mathbb{P}}(1))}.$$

By definition of the prior π_1, the numerator on the left hand side is distributed $\mathrm{Binom}(p, \varepsilon)$. Then it is sufficient for the following equality to hold,

$$\log \frac{\sum_{j=1}^p \mathbf{1}\{\theta_{j\cdot} \text{ has a changepoint}\}}{p} = -\beta(1 + o_{\mathbb{P}}(1)) \log p.$$

Subtracting $\log \varepsilon = \log p^{-\beta}$ from both sides, the above is implied by

$$\frac{\log \frac{\sum_{j=1}^p \mathbf{1}\{\theta_{j\cdot} \text{ has a changepoint}\}}{p\varepsilon}}{\log p} = o_{\mathbb{P}}(1). \tag{3.77}$$

By Markov's inequality, for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{\log \frac{\sum_{j=1}^{p} \mathbf{1}\{\theta_{j:} \text{ has a changepoint}\}}{p\varepsilon}}{\log p} > \delta\right) &= \mathbb{P}\left(\frac{\sum_{j=1}^{p} \mathbf{1}\{\theta_{j:} \text{ has a changepoint}\}}{p\varepsilon} > p^{\delta}\right) \\
&\leq p^{-\delta} \frac{1}{p\varepsilon} \sum_{j=1}^{p} \mathbb{P}(\theta_{j:} \text{ has a changepoint}) \\
&= p^{-\delta} \to 0.
\end{aligned}$$

This implies (3.77) and completes the proof. □

Lemma 3.7.3. *Let $X_1, \ldots, X_n$ be gaussian variables with* $\mathrm{Cov}(X_i, X_j) \coloneqq c_{ij} \geq 0$*, and* $\mathrm{Var}(X_i) = 1$ *for $i \leq n$. Then $X_1, \ldots, X_n$ are jointly equal in distribution to*

$$\begin{aligned}
W_1 &\coloneqq Z_1 + Z_{12} + Z_{13} + \cdots + Z_{1n} \\
W_2 &\coloneqq Z_2 + Z_{21} + Z_{23} + \cdots + Z_{2n} \\
&\vdots \\
W_n &\coloneqq Z_n + Z_{n1} + Z_{n2} + \cdots + Z_{n(n-1)},
\end{aligned}$$

where $Z_{ij} = Z_{ji}$ and Z_i are independent mean zero gaussians with variances

$$\begin{aligned}
\mathrm{Var}(Z_i) &= 1 - \sum_{1 \leq j \leq n : j \neq i} c_{ij} \\
\mathrm{Var}(Z_{ij}) &= c_{ij}.
\end{aligned}$$

Proof. It is straightforward to verify that the W_i have unit variance and the correct covariances. □

Lemma 3.7.4. *We have*

$$\mathrm{Var}_0(S_q) \asymp \sum_{j=1}^{p} \sum_{l=1}^{\log n} \mathrm{Var}_0(\boldsymbol{1}\{\max_{t \in \mathcal{T}_l'} A_{jt} > y_q\}).$$

Proof. We split the sum over pairs $(\mathcal{T}_k', \mathcal{T}_l')$ with $|k - l| \leq m$ and $|k - l| > m$, where m is to be chosen to make the cross term contribution negligible.

$$\begin{aligned}\mathrm{Var}_0\left(\sum_{l=1}^{\log n}\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\}\right) &= \sum_{|k-l|\le m}\mathrm{Cov}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'}A_{jt}>y_q\},\mathbf{1}\{\max_{s\in\mathcal{T}_l'}A_{js}>y_q\})\\ &\quad+\sum_{|k-l|>m}\mathrm{Cov}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'}A_{jt}>y_q\},\mathbf{1}\{\max_{s\in\mathcal{T}_l'}A_{js}>y_q\}).\end{aligned} \tag{3.78}$$

The first term on the rhs of (3.78) is of order

$$\begin{aligned}&=\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\})+\sum_{1\le|k-l|\le m}\mathrm{Cov}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'}A_{jt}>y_q\},\mathbf{1}\{\max_{s\in\mathcal{T}_l'}A_{js}>y_q\})\\ &\le\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\})\\ &+\sum_{1\le|k-l|\le m}\left(\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'}A_{jt}>y_q\})\mathrm{Var}_0(\mathbf{1}\{\max_{s\in\mathcal{T}_l'}A_{js}>y_q\})\right)^{1/2}\\ &\le\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\})\\ &+\sum_{1\le|k-l|\le m}\frac{1}{2}\left(\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'}A_{jt}>y_q\})+\mathrm{Var}_0(\mathbf{1}\{\max_{s\in\mathcal{T}_l'}A_{js}>y_q\})\right)\\ &\le\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\})+m^2\cdot\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\})\\ &\lesssim\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_l'}A_{jt}>y_q\}).\end{aligned}$$

It remains to show that the sum of non-adjacent cross terms with $|k-l|>m$ are of order no larger than the sum of diagonal terms, i.e. we want to show

$$\sum_{|k-l|>m}\mathrm{Cov}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'}A_{jt}>y_q\},\mathbf{1}\{\max_{s\in\mathcal{T}_l'}A_{js}>y_q\})=O\left(\sum_{l=1}^{\log n}\mathrm{Var}_0(\mathbf{1}\{\max_{t\in\mathcal{T}'}A_{jt}>y_q\})\right).$$

To do this, we use the gaussian representation lemma. We can write each covariance term

$\mathrm{Cov}_0(\mathbf{1}\{\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\}, \mathbf{1}\{\max_{s\in\mathcal{T}_l'} A_{js} > y_q\})$ as

$$= \mathbb{P}_0\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q, \max_{s\in\mathcal{T}_l'} A_{js} > y_q\right) - \mathbb{P}_0\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\mathbb{P}_0\left(\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right). \quad (3.79)$$

Then we may write

$$A_{js} = \left(Z_{js} + \sum_{t\in\mathcal{T}_k'} Z_{jts}\right)^2 - 1,$$

where by Part 3 of Lemma (3.7.1), we have

$$\begin{aligned} \mathrm{Var}(Z_{jts}) &= \mathrm{Cov}(\langle X_{j:}, \theta^{(t)}\rangle, \langle X_{j:}, \theta^{(s)}\rangle) = \langle \theta^{(t)}, \theta^{(s)}\rangle \le b^{-\frac{|k-l|}{2}} := (\log n)^{-\frac{|k-l|}{2}} = p^{-a\frac{|k-l|}{2}} \\ \mathrm{Var}(Z_{js}) &= 1 - \sum_{t\in\mathcal{T}_k'} \mathrm{Var}(Z_{jts}) \asymp 1 - (\log\log n) p^{-a\frac{|k-l|}{2}} = 1 + o(1), \end{aligned}$$

and these gaussians are all independent of each other. Since $|\mathcal{T}_k'| \asymp |\mathcal{T}_l'| \asymp \log\log n$, we have that

$$\mathrm{Var}\left(\sum_{t\in\mathcal{T}_k'} Z_{jts}\right) \lesssim (\log\log n) b^{-\frac{|k-l|}{2}} := \mathsf{PL}\cdot b^{-\frac{|k-l|}{2}}, \quad (3.80)$$

where PL denotes a polylogarithmic factor in p. Then it follows that

$$\max_{s\in\mathcal{T}_l'}\left|\sum_{t\in\mathcal{T}_k'} Z_{jts}\right| \lesssim \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}$$

with high probability. The joint probability in the covariance term (3.79) is

$$\lesssim \mathbb{P}_0\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q, \max_{s\in\mathcal{T}_l'} Z_{js}^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right) + \mathbb{P}\left(\max_{s\in\mathcal{T}_l'}\left|\sum_{t\in\mathcal{T}_k'} Z_{jts}\right| \ge \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right). \quad (3.81)$$

The second term can be made arbitrarily small, since (3.80) implies

$$\mathbb{P}\left(\max_{s\in\mathcal{T}_l'}\left|\sum_{t\in\mathcal{T}_k'} Z_{jts}\right| \geq \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right) \leq \mathbb{P}\left(\max_{s\in\mathcal{T}_l'}\frac{\left|\sum_{t\in\mathcal{T}_k'} Z_{jts}\right|}{\sqrt{\mathrm{Var}\left(\sum_{t\in\mathcal{T}_k'} Z_{jts}\right)}} \geq \mathsf{PL}\right)$$
$$\leq (\log\log n)\mathbb{P}\left(|N(0,1)| > \mathsf{PL}\right),$$

which can be made arbitrarily polynomially small in p by Mills ratio. Hence we can ignore this term, e.g. choose $\mathsf{PL} = \sqrt{2(a-(1+a)q)(1+\gamma)\log p}$ with $\gamma > 0$. Now by independence from the Gaussian representation lemma, (3.81) becomes

$$= \mathbb{P}_0\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\mathbb{P}_0\left(\max_{s\in\mathcal{T}_l'} Z_{js}^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right).$$

Thus the entire covariance term is

$$\begin{aligned}
&\asymp \mathbb{P}_0\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\left[\mathbb{P}_0\left(\max_{s\in\mathcal{T}_l'} Z_{js}^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right) - \mathbb{P}_0\left(\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right)\right]\\
&\leq \mathbb{P}_0\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\left[\mathbb{P}_0\left(\max_{s\in\mathcal{T}_l'} Z_{js}^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right) - \mathbb{P}_0\left(\max_{s\in\mathcal{T}_l'} Z_{js}^2 - 1 > y_q\right)\right],
\end{aligned} \tag{3.82}$$

since A_{js} has more noise inside the square than does $Z_{js}^2 - 1$. It suffices to estimate the bracketed difference. This difference is equal to

$$\begin{aligned}
&= \mathbb{P}_0\left(y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}} < \max_{s\in\mathcal{T}_l'} Z_{js}^2 - 1 \leq y_q\right)\\
&\leq \mathbb{P}_0\left(\bigcup_{s\in\mathcal{T}_l'}\left\{y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}} < Z_{js}^2 - 1 \leq y_q\right\}\right)\\
&= \sum_{s\in\mathcal{T}_l'}\mathbb{P}_0\left(y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}} < Z_{js}^2 - 1 \leq y_q\right)\\
&= \sum_{s\in\mathcal{T}_l'}\int_{y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}}^{y_q} f_{js}(z)dz,
\end{aligned}$$

where $f_{js}(z)$ is the density of $Z_{js}^2 - 1$. Now since Z_{js} is centered normal with variance going to 1, the density of $Z_{js}^2 - 1$ is bounded by a constant. Hence the above is bounded by

$$\lesssim \sum_{s \in \mathcal{T}_l'} \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}} = \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}},$$

since $|\mathcal{T}_l'| \asymp \log\log n$. Using the above estimate on the bracketed difference in (3.82), we have

$$\sum_{|k-l|>m} \mathrm{Cov}_0(\mathbf{1}\{\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\}, \mathbf{1}\{\max_{s \in \mathcal{T}_l'} A_{js} > y_q\}) \lesssim \sum_{|k-l|>m} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right) \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}}.$$

Bounding the probability in each summand by the max of these probabilities, the above is less than

$$\begin{aligned}
&\leq \left[\max_{k \leq \log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right)\right] \mathsf{PL} \cdot \sum_{k,l \leq \log n, |k-l|>m} (b^{-1/4})^{|k-l|} \\
&= \left[\max_{k \leq \log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right)\right] \mathsf{PL} \cdot \sum_{k,l \leq \log n, |k-l|>m} (b^{-1/4})^{|k-l|-(m+1)+(m+1)} \\
&\lesssim \left[\max_{k \leq \log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right)\right] \mathsf{PL} \cdot b^{-\frac{m+1}{4}} \sum_{k=1}^{\log n} \sum_{i=0}^{\infty} (b^{-1/4})^i \\
&\lesssim \left[\max_{k \leq \log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right)\right] \mathsf{PL} \cdot b^{-\frac{m+1}{4}+1}
\end{aligned}$$

since $\sum_0^\infty (b^{-1/4})^i \lesssim 1$ is a geometric series. Picking $m = 4$ will suffice, and the above then reads

$$\sum_{|k-l|>m} \mathrm{Cov}_0(\mathbf{1}\{\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\}, \mathbf{1}\{\max_{s \in \mathcal{T}_l'} A_{js} > y_q\}) \lesssim \max_{k \leq \log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right).$$

The result now follows because

$$\sum_{l=1}^{\log n} \mathrm{Var}_0\left(\mathbf{1}\{\max_{t \in \mathcal{T}_l'} A_{jt} > y_q\}\right) \asymp \sum_{l=1}^{\log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_l'} A_{jt} > y_q\right) \geq \max_{k \leq \log n} \mathbb{P}_0\left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q\right).$$

□

Lemma 3.7.5. *For consistent Type II error, it suffices to have some $q > 0$ for which*

$$\frac{\mathrm{Var}_\theta(S_q)}{(\mathbb{E}_\theta S_q)^2} \to 0 \tag{3.83}$$

$$\frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \to \infty \textit{ polynomially.} \tag{3.84}$$

Proof. By definition of the supremum, the Type II error is bounded as

$$\begin{aligned}\mathbb{P}_\theta\left(\mathsf{HC}_p \le C\sqrt{\log\log p}\right) &:= \mathbb{P}_\theta\left(\sup_{q>0}\frac{S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \le C\sqrt{\log\log p}\right)\\ &\le \mathbb{P}_\theta\left(\frac{S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \le C\sqrt{\log\log p}\right)\\ &\le \mathbb{P}_\theta\left(\frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} - \frac{S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} \ge \frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} - C\sqrt{\log\log p}\right).\end{aligned}$$

By the condition (3.84), the right hand side of the inequality inside the probability is non-negative. Thus by monotonicity and Chebyshev's inequality, the probability is bounded above,

$$\le \mathbb{P}_\theta\left(\left(\frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} - \frac{S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}}\right)^2 \ge \left(\frac{\mathbb{E}_\theta S_q - \mathbb{E}_0 S_q}{\sqrt{\mathrm{Var}_0(S_q)}} - C\sqrt{\log\log p}\right)^2\right) \lesssim \frac{\mathrm{Var}_\theta(S_q)}{(\mathbb{E}_\theta S_q)^2},$$

since condition (3.84) implies that $\mathbb{E}_\theta S_q \to \infty$ polynomially. By condition (3.83), the right hand side tends to zero, completing the proof.

□

Lemma 3.7.6. *For any $\theta \in \Theta_1(p, a, r, \beta)$, we have*

$$\mathrm{Var}_\theta\left(\sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\right) \lesssim \sum_{l=1}^{\log n}\mathrm{Var}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\right) + o(p^{-1}).$$

Proof. The variance of the sum can be written as

$$\begin{aligned}\mathrm{Var}_\theta\left(\sum_{l=1}^{\log n} \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\right) &= \sum_{|k-l|\le m}\mathrm{Cov}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right\}, \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\right)\\ &\quad + \sum_{|k-l|>m}\mathrm{Cov}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right\}, \mathbf{1}\left\{\max_{t\in\mathcal{T}_l'} A_{jt} > y_q\right\}\right),\end{aligned} \tag{3.85}$$

where m is to be chosen during the analysis. By the same reasoning from the proof of Lemma (3.7.4), the first term on the right hand side of (3.85) is of order

$$\asymp \sum_{l=1}^{\log n} \mathrm{Var}_\theta \left(\mathbf{1} \left\{ \max_{t \in \mathcal{T}_l'} A_{jt} > y_q \right\} \right).$$

It remains to show the contribution from the cross terms with $|k-l| > m$ are of smaller order than p^{-1}, i.e. that

$$\sum_{|k-l|>m} \mathrm{Cov}_\theta \left(\mathbf{1} \left\{ \max_{t \in \mathcal{T}_k'} A_{jt} > y_q \right\}, \mathbf{1} \left\{ \max_{s \in \mathcal{T}_l'} A_{js} > y_q \right\} \right) = o(p^{-1}). \tag{3.86}$$

To this end, we use the gaussian representation lemma (3.7.3). We can write the $(k,l)^{th}$ covariance term as

$$= \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q, \max_{s \in \mathcal{T}_l'} A_{js} > y_q \right) - \mathbb{P}_\theta \left(\max_{t \in \mathcal{T}_k'} A_{jt} > y_q \right) \mathbb{P}_\theta \left(\max_{s \in \mathcal{T}_l'} A_{js} > y_q \right). \tag{3.87}$$

By the Lemma, we may write

$$A_{js} = \left(\mu_{js} + Z_{js} + \sum_{t \in \mathcal{T}_k'} Z_{jts} \right)^2 - 1$$

where μ_{js} is a mean parameter determined by θ, and

$$\begin{aligned} \mathrm{Var}(Z_{jts}) &= \mathrm{Cov}\left(\langle X_{j:}, \theta^{(t)} \rangle, \langle X_{j:}, \theta^{(s)} \rangle \right) \leq b^{-\frac{|k-l|}{2}} \\ \mathrm{Var}(Z_{js}) &= 1 - \sum_{t \in \mathcal{T}_k'} \mathrm{Var}(Z_{jts}) \asymp 1 - (\log\log n) b^{-\frac{|k-l|}{2}} = 1 + o(1), \qquad (|\mathcal{T}_k'| \asymp \log\log n) \end{aligned}$$

and these gaussians are independent of each other. Now since

$$\mathrm{Var}\left(\sum_{t \in \mathcal{T}_k'} Z_{jts} \right) \lesssim (\log\log n) b^{-\frac{|k-l|}{2}} = \mathsf{PL} \cdot b^{-\frac{|k-l|}{2}}.$$

Splitting the joint probability on the event

$$\left\{ \max_{s \in \mathcal{T}_l'} \left| \sum_{t \in \mathcal{T}_k'} Z_{jts} \right| \leq \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}} \right\},$$

the joint probability in the covariance term (3.87) is bounded by

$$\leq \mathbb{P}_\theta\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q, \max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right)$$
$$+\,\mathbb{P}\left(\max_{s\in\mathcal{T}_l'}\left|\sum_{t\in\mathcal{T}_k'} Z_{jts}\right| > \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right).$$

By the same reasoning in the proof of Lemma (3.7.4), the second term above can be made negligibly (polynomially) small. Thus we focus on the first term. By independence, it is equal to

$$= \mathbb{P}_\theta\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right).$$

Plugging this into the covariance term (3.87), the $(k,l)^{th}$ covariance term is

$$\begin{aligned}
&= \mathbb{P}_\theta\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\left[\mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right)\right.\\
&\left. - \mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right)\right]\\
&\leq \mathbb{P}_\theta\left(\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right)\left[\mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right)\right.\\
&\left. - \mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q\right)\right]\\
&\leq \mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q - \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}\right) - \mathbb{P}_\theta\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q\right),
\end{aligned}$$

since A_{js} has more noise inside the square than does $(\mu_{js}+Z_{js})^2 - 1$, which implies the inequality $\mathbb{P}\left(\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right) \geq \mathbb{P}\left(\max_{s\in\mathcal{T}_l'}(\mu_{js}+Z_{js})^2 - 1 > y_q\right)$. It suffices to estimate this difference. By monotonicity and a union bound, it is

$$\begin{aligned}
&= \mathbb{P}_\theta \left(y_q - \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}} < \max_{s \in \mathcal{T}_l'} (\mu_{js} + Z_{js})^2 - 1 \leq y_q \right) \\
&\leq \mathbb{P}_\theta \left(\bigcup_{s \in \mathcal{T}_l'} \left\{ y_q - \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}} + 1 < (\mu_{js} + Z_{js})^2 \leq y_q + 1 \right\} \right) \\
&\leq \sum_{s \in \mathcal{T}_l'} \int_{y_q - \mathsf{PL} \cdot b^{-\frac{|k-l|}{4}} + 1}^{y_q + 1} f_{js}(z) dz, \qquad (3.88)
\end{aligned}$$

where f_{js} denotes the density of $(\mu_{js} + Z_{js})^2$. It suffices to bound this density by a polylogarithmic factor in p. Since

$$(\mu_{js} + Z_{js})^2 = \mathrm{Var}(Z_{js}) \cdot \left(\frac{\mu_{js}}{\sqrt{\mathrm{Var}(Z_{js})}} + \frac{Z_{js}}{\sqrt{\mathrm{Var}(Z_{js})}} \right)^2,$$

we can bound the density of $Y := \left(\frac{\mu_{js}}{\sqrt{\mathrm{Var}(Z_{js})}} + \frac{Z_{js}}{\sqrt{\mathrm{Var}(Z_{js})}} \right)^2$ using Lemma (3.7.7), and use this bound to get a corresponding bound on the density f_{js} since $\mathrm{Var}(Z_{js}) \to 1$. This variable is noncentral chi square with noncentrality parameter $\frac{\mu_{js}^2}{\mathrm{Var}(Z_{js})}$ and one degree of freedom. Hence by the Lemma, we have the bound,

$$f_Y(x) \leq 3 + \frac{3\mu_{js}^2}{2\mathrm{Var}(Z_{js})} \lesssim \mu_{js}^2.$$

Since μ_{js} is the signal from the alternative parameter $\theta \in \Theta_1(p, a, r, \beta)$, it is on the order of ρ, which is polylogarithmic in p. Hence f_Y is uniformly bounded by a polylogarithmic factor. Then

$$\begin{aligned}
f_{js}(z) &= \frac{d}{dx} \mathbb{P} \left(\mathrm{Var}(Z_{js}) \cdot Y \leq x \right) \big|_{x=z} \\
&= \frac{d}{dx} \mathbb{P} \left(Y \leq \frac{x}{\mathrm{Var}(Z_{js})} \right) \Bigg|_{x=z} \\
&= f_Y \left(\frac{z}{\mathrm{Var}(Z_{js})} \right) \cdot \frac{1}{\mathrm{Var}(Z_{js})} \\
&\lesssim \mathsf{PL}.
\end{aligned}$$

Plugging this bound into expression (3.88) and using $|\mathcal{T}_l'| \asymp \log \log n \in \mathsf{PL}$, we obtain the bound

$$\mathrm{Cov}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right\}, \mathbf{1}\left\{\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right\}\right) \lesssim \mathsf{PL}\cdot b^{-\frac{|k-l|}{4}}.$$

Hence the sum over all pairs (k,l) for which $|k-l|>m$ is

$$\sum_{|k-l|>m}\mathrm{Cov}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right\}, \mathbf{1}\left\{\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right\}\right) \lesssim \sum_{|k-l|>m}\mathsf{PL}\cdot b^{-\frac{|k-l|}{4}},$$

which can be upper bounded by

$$\begin{aligned}
&\leq \mathsf{PL}\cdot \sum_{k,l\leq \log n, |k-l|>m} (b^{-1/4})^{|k-l|}\\
&= \mathsf{PL}\cdot \sum_{k,l\leq \log n, |k-l|>m} (b^{-1/4})^{|k-l|-(m+1)+(m+1)}\\
&\lesssim \mathsf{PL}\cdot b^{-\frac{m+1}{4}}\cdot \sum_{k=1}^{\log n}\sum_{i=0}^{\infty}(b^{-1/4})^i\\
&\lesssim \mathsf{PL}\cdot b^{-\frac{m+1}{4}+1},
\end{aligned}$$

since $\sum_{i=0}^{\infty}(b^{-1/4})^i$ is a geometric series. Plugging in $b=p^a$, the above reads

$$\sum_{|k-l|>m}\mathrm{Cov}_\theta\left(\mathbf{1}\left\{\max_{t\in\mathcal{T}_k'} A_{jt} > y_q\right\}, \mathbf{1}\left\{\max_{s\in\mathcal{T}_l'} A_{js} > y_q\right\}\right) \lesssim \mathsf{PL}\cdot p^{\frac{a(3-m)}{2}}.$$

It is clear now that in order for the above to be $o(p^{-1})$, we need

$$\frac{a(3-m)}{2} < -1,$$

or equivalently, $m > 3+\frac{2}{a}$. Choosing m sufficiently large ensures (3.86), completing the proof.

□

Lemma 3.7.7. *The density $f_{\lambda,k}$ of a noncentral chi square variable with noncentrality parameter λ and k degrees of freedom satisfies the bound,*

$$f_{\lambda,k}(x) \leq 3k + \frac{3\lambda}{2}.$$

Proof. From Wikipedia, the density has the formula

$$f_{\lambda,k}(x) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!} f_{Y_{k+2i}}(x)$$

where Y_q is distributed as a central chi square with q degrees of freedom. Also from Wikipedia, the mode of a central chi square with q degrees of freedom is $(q-2) \vee 0$. Hence the above is bounded by

$$\begin{aligned}
&\leq \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!}(k+2i-2) \\
&= \sum_{i=0}^{k} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!}(k+2i-2) + \sum_{i=k+1}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!}(k+2i-2) \\
&\leq \sum_{i=0}^{k} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!} \cdot 3k + \sum_{i=k+1}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!} \cdot 3i \\
&\leq 3k + 3 \sum_{i=k+1}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{(i-1)!} \\
&\leq 3k + \frac{3\lambda}{2} \sum_{i=1}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^{i-1}}{(i-1)!} \\
&= 3k + \frac{3\lambda}{2},
\end{aligned}$$

where we have used the definition $e^x := \sum_{i=0}^{\infty} \frac{x^n}{n!}$. □

CHAPTER 4
SPARSE SIGNAL IDENTIFICATION

4.1 Introduction

Consider an observation Y that is the sum of a signal $X \sim P_\rho$ and an independent standard Gaussian perturbation,

$$Y = X + \varepsilon \in \mathbb{R}, \quad \varepsilon \sim N(0,1). \tag{4.1}$$

The family (P_ρ) is assumed to be sparse in the sense that the signal distribution P_ρ accumulates probability mass near the origin as the rate parameter $\rho \to 0$. Compared to the problem studied in the previous chapter, the testing problem we consider in this chapter is simpler in the sense that the observation in this model consists of independent samples from (4.1), whereas in the changepoint problem we observe a sequence of independent random vectors. Here, the signal X is non-null if it is non-negligibly far from zero in absolute value. In the changepoint problem, a row of the observation matrix is non-null if its mean vector has a changepoint.

In principle, statistical sparsity is well-defined for a sample size of $n = 1$, and is formulated by McCullagh and Polson (2018) in terms of weak-convergence of the measures (P_ρ) over a class of functions that are $O(x^2)$ at the origin. In the current chapter, we connect this definition of sparsity to the literature on hypothesis tests for the global null against sparse alternatives (see e.g. Donoho and Jin (2004), Cai and Wu (2014), Collier et al. (2017), Li and Fithian (2020) and references therein). For a given value of the sparsity rate $\rho > 0$, a batch of independent samples with distribution (4.1) is distinguishable from pure noise $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} N(0,1)$ as $n \to \infty$, for instance using a likelihood ratio test. Conversely, with few enough samples, the MLE for ρ, denoted by $\hat{\rho}_n$, is identically zero with non-trivial probability even when $\rho > 0$. An *identification boundary* formula determines the number of independent samples $n = b(\rho)$ needed for the maximum-likelihood estimate $\hat{\rho}_n$ to be positive in the sparse limit as $\rho \to 0$.

Organization of the chapter. In Section 4.2, we review a probabilistic definition of sparsity as a limiting property of a sequence of signal distributions (P_ρ). We next state the definition of the identification boundary and our main result, and connect them to the detection boundary literature for heavy-tailed signal distributions. Section 4.3 contains a brief discussion and illustration of the MLE for estimating null probabilities on a prostate dataset previously analyzed by Efron (2012). Appendix 4.4 contains proofs and numerical evidence for our results.

4.2 Sparse identification boundary

The sparse limit of a sequence (P_ρ) is characterized by a positive measure $H(\mathrm{d}x)$ on $\mathbb{R}\backslash\{0\}$ by means of weak convergence over a collection of suitably regular functions.

Definition 4.2.1 (McCullagh and Polson (2018))**.** *We say that a family* (P_ρ) *has a sparse limit with rate* ρ *and exceedance measure* $H(\mathrm{d}x)$ *if*

$$\lim_{\rho\to 0} \rho^{-1} \int w(x) P_\rho(\mathrm{d}x) = \int_{\mathbb{R}\setminus\{0\}} w(x) H(\mathrm{d}x) < \infty \tag{4.2}$$

for every bounded continuous function $w : \mathbb{R} \to \mathbb{R}$ *for which* $w(x)/x^2$ *is also bounded. If* $H(\mathrm{d}x)$ *satisfies*

$$\int_{\mathbb{R}\setminus\{0\}} (1 - e^{-x^2/2}) H(\mathrm{d}x) = 1, \tag{4.3}$$

then it is called a unit exceedance measure.

The inverse-power exceedance with index $d \in (0, 2)$, defined by

$$H_d(\mathrm{d}x) \coloneqq C_d \cdot \frac{\mathrm{d}x}{|x|^{d+1}}, \quad C_d \coloneqq \frac{d\, 2^{d/2-1}}{\Gamma(1 - d/2)}, \tag{4.4}$$

satisfies the definition (4.3) of a unit exceedance measure (McCullagh and Polson; 2018). For example, the t-distribution on $d \in (0, 2)$ degrees of freedom and scale parameter $\sigma > 0$, denoted $t_d(\sigma)$, with probability density

$$f_{d,\sigma}(x) = A_d \cdot \frac{\sigma^d}{(\sigma^2 + x^2/d)^{\frac{d+1}{2}}}, \quad A_d \coloneqq \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi}\,\Gamma(d/2)}, \tag{4.5}$$

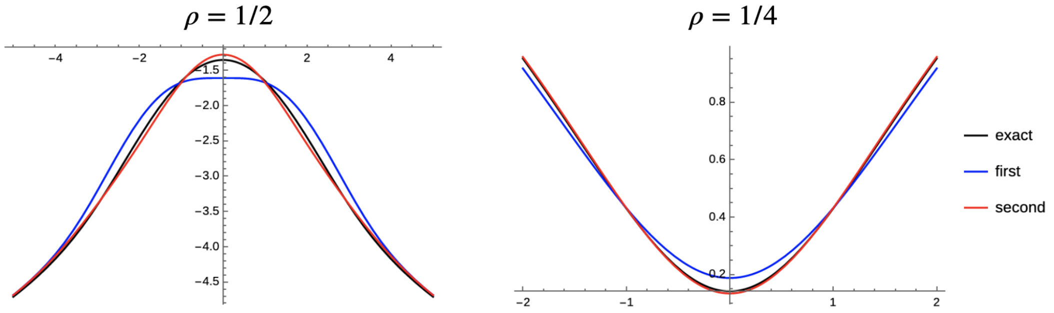
tends to the inverse-power exceedance $H_d(\mathrm{d}x)$ at rate

$$\rho = \frac{d^{\frac{d+1}{2}} A_d}{C_d} \cdot \sigma^d \qquad \text{as } \sigma \to 0. \tag{4.6}$$

4.2.1 Marginal distribution

For a symmetric signal distribution P_ρ tending to the unit inverse-power exceedance $H_d(\mathrm{d}x)$ with index d at rate ρ, the marginal distribution of the convolution (4.1), denoted by $m_\rho(y) \coloneqq \int \phi(y - x) P_\rho(\mathrm{d}x)$, is conveniently expressed in terms of ρ and the unit-exceedance $H_d(\mathrm{d}x)$. Letting $\phi(y)$ denote the standard normal density, we have

$$\begin{aligned} m_\rho(y) &= \phi(y) \left(\int e^{-x^2/2} P_\rho(\mathrm{d}x) + \int (e^{yx} - 1) e^{-x^2/2} P_\rho(\mathrm{d}x) \right) \\ &= \phi(y) \left(1 - \int (1 - e^{-x^2/2}) P_\rho(\mathrm{d}x) + \int (\cosh(xy) - 1) e^{-x^2/2} P_\rho(\mathrm{d}x) \right). \end{aligned}$$


$\rho = 1/2$

$\rho = 1/4$

Figure 4.1: The exact log marginal density, $\log m_\rho(y)$ where $m_\rho(y) := \int \phi(y-x)P_\rho(\mathrm{d}x)$, is plotted in black for the sparse Cauchy model at $\rho = 0.5$ in the left panel, along with the first and second order approximations in blue and red, respectively. On the right plot, $\log(-\log m_\rho(y))$ is plotted in black along with the corresponding sparse approximations in blue and red when $\rho = 0.25$.

by symmetry of P_ρ. Applying the definition of sparsity, the above is equal to

$$= \phi(y)\left(1 - \rho + \rho \int_{\mathbb{R}\setminus\{0\}} (\cosh(xy) - 1)e^{-x^2/2} H_d(\mathrm{d}x) + o(\rho)\right),$$

because $1 - e^{-x^2/2}$ and $\cosh(xy) - 1$ are both $O(x^2)$ at the origin. Denote by $\zeta_d(y)$ the integral factor appearing in the above expression for the marginal density,

$$\zeta_d(y) := \int_{\mathbb{R}\setminus\{0\}} (\cosh(xy) - 1)e^{-x^2/2} H_d(\mathrm{d}x).$$

The marginal distribution of Y is thus approximately distributed according to the following two-component mixture,

$$(1-\rho)\phi(y) + \rho\psi_d(y), \quad \psi_d(y) := \phi(y)\zeta_d(y). \tag{4.7}$$

The derivation leading to expression (4.7) implies the approximation error is at most $o(\rho)$, point-wise in y as $\rho \to 0$, but it can be substantially less in particular examples. For typical 'atom-and-slab' distributions considered in the detection boundary literature, our results in Section 4.2.3 suggest that the squared Hellinger distance between the exact marginal distribution of Y and its marginal approximation (4.7) is $o(\rho^2)$. Higher order approximations can be calculated (see Section 4.4.1), but the first order approximation suffices for practical purposes; in Figure 4.1 the first and second order approximations are compared for the sparse Cauchy model, obtained by taking $d = 1$ in expression (4.5).

Definition 4.2.2. *Suppose (P_ρ) has a sparse limit with inverse-power exceedance measure H_d and rate $\rho \to 0$. Let $\hat{\rho}_n$ denote the maximum likelihood estimator of the sparsity rate* *based on independent samples $Y_1, \ldots, Y_n$ from the marginal approximation* (4.7),

$$\hat{\rho}_n = \text{argmax}_{0 \le \rho \le 1} \prod_{i=1}^{n} \left((1-\rho)\phi(Y_i) + \rho\psi_d(Y_i) \right). \tag{4.8}$$

For $\gamma \in (0, 1]$, a function $b_\gamma(\rho)$ is called a γ-identification boundary for H_d if

$$\liminf_{\substack{\rho \to 0 \\ n = b_\gamma(\rho)}} P_\rho^{\otimes n}(\hat{\rho}_n > 0) \ge \gamma,$$

where $P_\rho^{\otimes n}$ refers to the product distribution of $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (1-\rho)\phi + \rho\psi_d$.

For notational convenience, we write $n = b_\gamma(\rho)$ to denote the number of samples needed to ensure the maximum-likelihood estimate $\hat{\rho}_n$ is positive with at least $\gamma > 0$ probability in the sparse limit as $\rho \to 0$. Although $b_\gamma(\rho)$ is a real-valued expression in what follows, this equality can be read as $n = \lceil b_\gamma(\rho) \rceil$, so as to ensure an integer valued sample size n.

4.2.2 Main result

Consider n independent observations from the two component mixture (4.7),

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (1-\rho)\phi + \rho\psi_d. \tag{4.9}$$

The identification boundary for the inverse-power family is established in the following result, proved in Appendix 4.4.

Theorem 4.2.1. *For any $\gamma > 0$, the formula*

$$b_\gamma(\rho) = -2\Gamma(1 - d/2)\log(1-\gamma) \cdot \frac{(\log \rho^{-1})^{d/2}}{\rho}$$

is a γ-identification boundary for the inverse-power exceedance $H_d(\mathrm{d}x)$ with index $d \in (0, 2)$.

The formula $\frac{(\log \rho^{-1})^{d/2}}{\rho}$ is critical in the sense that if n is of smaller order than this quantity as $\rho \to 0$, then no test based on n independent samples can reliably detect whether (4.9) is the generating process, or whether the data are purely noise $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(0, 1)$. We discuss this point with more detail in Section 4.4.4, where we review a bound on the total variation distance of two product distributions in terms of the squared Hellinger distance between the corresponding one-dimensional distributions. In this sense, establishing the identification boundary for testing between two product distributions reduces to computing the Hellinger distance between the one-dimensional distributions, which we calculate in Appendix 4.4 and summarize in the following lemma.

Lemma 4.2.1. *For $d \in (0,2)$, and ψ_d defined below expression (4.7), there exist constants $c, C > 0$ for which*

$$\frac{c\rho}{(\log \rho^{-1})^{d/2}} \leq H^2(\phi, (1-\rho)\phi + \rho\psi_d) \leq \frac{C\rho}{(\log \rho^{-1})^{d/2}} \qquad \text{as } \rho \to 0,$$

where the squared Hellinger distance between two probability distributions with densities f, g is defined $H^2(f,g) \coloneqq \frac{1}{2}\int(\sqrt{f} - \sqrt{g})^2$.

If we are given the power-index d, we may compute the MLE (4.8) and reject when $\hat{\rho}_n > 0$ to obtain an asymptotically powerful test for the global null. Type 1 error properties of this procedure are investigated numerically in Section 4.4.5. More simply, the max-test, implemented by comparing the maximum $|Y_i|$ to its quantiles under the global null, achieves the information theoretic limit for detecting inverse-power families, a fact pointed out by Li and Fithian (2020). We discuss a connection with this work in the next section.

4.2.3 Connection with detection boundaries

Li and Fithian (2020) study sparse alternatives to the global null that involve convolutions of scale families with the standard normal distribution

$$\begin{aligned}
&H_0^{(n)} : Y_i \overset{\text{iid}}{\sim} N(0,1) \\
&H_1^{(n)} : Y_i \overset{\text{iid}}{\sim} (1-n^{-\beta})N(0,1) + n^{-\beta}\left(G_n * N(0,1)\right), \quad \beta \in (0,1)
\end{aligned}$$

where G_n is a scale family parameterized by a real number $r > 0$, and $G_n * N(0,1)$ denotes the convolution with density $m(y) = \int \phi(y-x)G_n(\mathrm{d}x)$.

The *detection boundary* is a formula, involving the signal parameter r and the sparsity parameter β, that determines whether or not it is possible to distinguish between $H_0^{(n)}$ and $H_1^{(n)}$ with asymptotically negligible type I and II error as $n \to \infty$. When the signal distribution is sparse with inverse-power tails, e.g. Student-t with d degrees of freedom and scale n^{-r}, this formula is

$$\beta^*(r) = 1 - dr, \tag{4.10}$$

characterizing the global null testing problem as follows.

Corollary 4.2.1 (Li and Fithian (2020))**.** *Let G_n be the Student-t distribution on $d \in (0,2)$ degrees of freedom with scale parameter $\sigma = n^{-r}$ for some $r \in (0, 1/d)$.*

1. *If $\beta > \beta^*(r)$, then for any hypothesis testing function $\psi : \mathbb{R}^n \to \{0,1\}$, the sum of type 1 and 2 errors is bounded away from zero,*

$$\liminf_{n\to\infty} P_0^{(n)}(\psi(Y_1,\ldots,Y_n) = 1) + P_1^{(n)}(\psi(Y_1,\ldots,Y_n) = 0) > 0.$$

where $P_i^{(n)}$ is the distribution of $(Y_1,\ldots,Y_n)$ under $H_i^{(n)}$ for $i = 0,1$.

2. If $\beta < \beta^*(r)$, then the α-level max-test, defined

$$\psi_{\max}(Y_1, \ldots, Y_n) = 1\left\{\max_{i \le n} |Y_i| > t_\alpha \sqrt{2 \log n}\right\}, \tag{4.11}$$

where $t_\alpha = \min\left\{t : P_0^{(n)}\left(\max_{i\le n} |Y_i| > t\sqrt{2\log n}\right) \le \alpha\right\}$, is asymptotically powerful, i.e.

$$P_1^{(n)}(\psi_{\max} = 1) \to 1.$$

The following spike and slab mixture distribution is a sparse, one-dimensional analogue to the signal distribution giving rise to $H_1^{(n)}$,

$$X \sim (1 - \sigma^\beta)\delta_0 + \sigma^\beta t_d(\sigma^r), \quad \beta \in [0, 1), \quad r \in (0, 1/d], \tag{4.12}$$

where $t_d(\sigma^r)$ is the Student-t distribution on d degrees of freedom with scale parameter σ^r. Definition (4.2) does not distinguish between measures of the form (4.12) when $\beta = \beta^*(r)$. All such signal distributions, parameterized by the pairs $(\beta^*(r), r)$ for $r \in (0, 1/d)$, have the same approximate marginal distribution (4.7) to first order in sparsity. This equivalence is recorded in the following result.

Proposition 4.2.1. *Fix $d \in (0, 2)$ and suppose $r \in (0, 1/d]$. If $\beta = \beta^*(r)$, then as $\sigma \to 0$, the distribution of X under (4.12) has a sparse limit with rate and exceedance measure*

$$\rho = \frac{d^{\frac{d+1}{2}} A_d}{C_d} \cdot \sigma, \quad H_d(\mathrm{d}x) = C_d \cdot \frac{\mathrm{d}x}{|x|^{d+1}},$$

where the constants A_d, C_d are defined in expressions (4.4) *and* (4.5)*.*

Here, the t-distribution was used in the slab component of the signal distribution (4.12) for concreteness. Similar conclusions may be drawn for other spike-and-slab measures, where the Student-t density is replaced by an inverse-power density whose tail decays inverse-polynomially at the same rate.

The above result characterizes the limiting behavior of signal distributions parametrized by points $\{(\beta^*(r), r) : r \in (0, 1/d]\}$ on the detection boundary, taking $\sigma = n^{-1}$. The conclusion to draw from this comparison is that each signal distribution corresponding to a point on the detection boundary has the same sparse approximation to the marginal distribution,

$$Y_i \overset{\text{iid}}{\sim} (1 - \rho)\phi + \rho\psi_d,$$

where $\psi_d(y) = \phi(y) \int (\cosh(xy) - 1)e^{-x^2/2} H_d(\mathrm{d}x)$. This observation has implications for identifiability of the probability $\mathbb{P}(X = 0 \mid Y = y)$, the principal quantity to estimate for identifying active sites once the global null has been rejected. In particular, *non-atomic* distributions fall into this inverse-power sparse family, which lead to the same marginal distribution for the response variable Y_i and place zero posterior mass on the event $\{X = 0\}$. This implies the posterior probability $\mathbb{P}(X = 0 \mid Y)$ is unidentifiable, and thus the only estimable quantity is not directly interpretable as the posterior probability of a point null hypothesis. The plug-in estimate based on the marginal approximation (4.7) and the MLE (4.8) is discussed further in Section 4.3.

Second order approximation

Lemma 4.2.1 quantifies how close the standard normal distribution is to the first order approximation of $m_\rho(y) \coloneqq \int \phi(y - x) P_\rho(\mathrm{d}x)$. The distance between the first-order sparsity approximation (4.7) and the exact density $m_\rho(y)$ is less straightforward to characterize. Although all signal distributions of the form (4.12) with $\beta = \beta^*(r)$ tend to the inverse-power exceedance (4.4), some lead to a marginal distribution closer to the first-order approximation (4.7) than others. For instance, two signal distributions of the form (4.12) with different $\beta = \beta^*(r)$ values will have different second-order approximations for the resulting marginal distribution,

$$m_\rho(y) = (1 - \rho)\phi(y) + \rho\psi_d(y) + c\rho\sigma^{r(2-d)}(1 - y^2)\phi(y) + o(\rho\sigma^{r(2-d)}),$$

for some constant $c > 0$ depending only on d. The above equality follows from the derivation leading to expression (4.7) and Lemma 4.4.1 in Appendix 4.4.1. Given the above form of the second-order approximation, the approximation (4.7) can be assessed by computing the Hellinger distance between the first and second-order sparsity approximations. This calculation is summarized in Proposition 4.4.1, which is stated and proven in Appendix 4.4.2.

4.3 Discussion

In this chapter, we have considered the Gaussian convolutional model where the signal distribution has a heavy inverse-power tail. We derived first and second-order sparse approximations to the marginal density of the observation $Y = X + \varepsilon$ for the model (4.12) and quantified the closeness of these approximations via Hellinger distance. These results determine the number of independent samples required to detect a sparse inverse-power signal, and shed light on the degree to which signal distributions within the same inverse-power family can differ. Closely related is the detection boundary literature for heavy tailed signal distributions, and we have outlined an explicit connection in Section 4.2.3.

4.3.1 Implications for identifiability

The sparse framework of McCullagh and Polson (2018) highlights that the identification of atom and slab signal distributions is limited to the sparse family they belong to, which includes non-atomic distributions. This non-identifiability has implications for modeling signal activity, which becomes the key question once the global null has been rejected. In the setting studied here, we have proposed the MLE $\hat{\rho}_n$ for the sparsity parameter ρ as an asymptotically powerful method for detecting sparse signals tending to the inverse-power exceedance.

The estimate $\hat{\rho}_n$ not only provides a test for the global null, but also naturally leads to a method for assessing activity rates among sites once the global null has been rejected. The two-groups model, popularized by Efron et al. (2001), assumes a binary latent variable $H \in \{0, 1\}$, whose outcome determines the distribution from which the observation Y is drawn,

$$\begin{aligned} H &\sim \text{Bernoulli}(\rho) \\ Y \mid H &\sim (1 - H)\phi + H\psi_d. \end{aligned}$$

In this setting, the local false discovery rate is defined as the posterior probability that $H = 0$,

$$\text{lfdr}(y) := \mathbb{P}(H = 0 \mid Y = y) = \frac{1-\rho}{1-\rho+\rho\zeta_d(y)}. \tag{4.13}$$

Note that the event $H = 0$ need not coincide with $X = 0$ in the convolutional model, and indeed any estimate of the quantity (5.4) is conservative for the posterior probability $\mathbb{P}(X = 0 \mid Y = y)$. Without the assumption $\psi_d(0) = 0$ (see the 'zero-assumption' in Efron (2012), Patra and Sen (2016)), the two-groups model is unidentifiable. As a result, the right hand side of expression (5.4) has been widely accepted within the empirical Bayes community as the principal estimand for assessing the activity at the i^{th} site based on the observation Y_i. Our next result shows that the probability of the event $\{|X| \le \delta\}$ is identifiable in the sparse limit as long as $\delta \gg \rho^{1/2}$.

Proposition 4.3.1. *Let $\beta, \varepsilon \in (0, 1)$ be fixed as $\rho \to 0$, and $H(\mathrm{d}x)$ is a unit exceedance measure.*

1. *If $\delta \ge \rho^{\frac{1}{2}-\beta}$, then*

$$\lim_{\rho\to 0} \frac{P_\rho(|X| \le \delta)}{Q_\rho(|X| \le \delta)} = 1$$

   for any two sequences (P_ρ), (Q_ρ) that are sparse with rate ρ and exceedance measure $H(\mathrm{d}x)$.

2. *If $\delta \le \rho^{\frac{1}{2}+\beta}$, then there exist sparse sequences $(P_\rho), (Q_\rho)$ with rate ρ and unit exceedance $H(\mathrm{d}x)$ for which*

$$\lim_{\rho\to 0} \frac{P_\rho(|X| \le \delta)}{Q_\rho(|X| \le \delta)} < \varepsilon.$$

One conclusion to draw from this result is that it is misguided to test a composite null

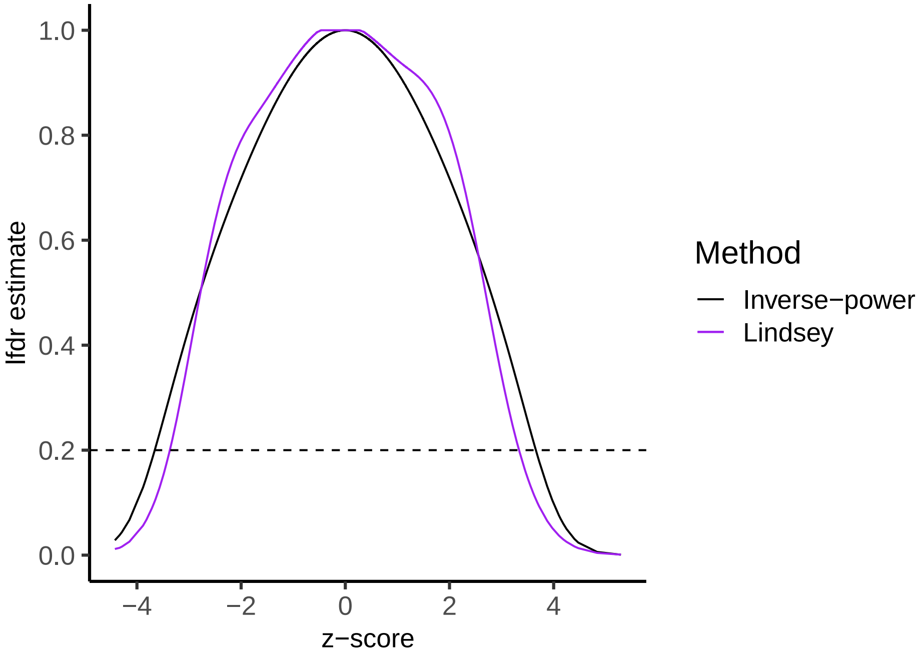


Figure 4.2: We plot the estimated lfdr curves for the prostate cancer dataset with a horizontal dashed line at the $\alpha = 0.2$ cut-off. The MLE plug-in estimate is plotted in black, with Lindsey's method overlaid in purple.

$|X| \leq \delta$ if $\delta \ll \rho^{1/2}$. In practice, ρ can be estimated (e.g. using $\hat{\rho}_n$) and this boundary can be assessed using the estimator for ρ.

4.3.2 Prostate cancer dataset

The prostate cancer dataset analyzed in this section is taken from Chapter 5 of Efron (2012). Gene expression levels are measured at $n = 6033$ genomic sites on 52 cases and 50 controls. The difference in average gene expressions is standardized, and after a pre-processing step, the observations are treated as z-scores $\{Z_i\}_{i=1}^n$ arriving from i.i.d. from the two-groups model:

$$\begin{aligned} H &\sim \text{Bernoulli}(\rho) \\ Z \mid H &\sim \begin{cases} f_0 & \text{if } H = 0 \\ f_1 & \text{if } H = 1. \end{cases} \end{aligned} \tag{4.14}$$

In our analysis we will treat f_0 as the standard normal density, and f_1 as the alternative component ψ_d in the first-order approximation (4.7). In Figure 5.10, we visually compare the estimates of (5.4) based on Lindsey's method (Lindsey (1974a), Lindsey (1974b), Efron

(2012)) with the plug-in estimate based on the (non-zero) maximum likelihood estimates

$$(\hat{\rho}_n, \hat{d}) = (0.092, 1.662).$$

The method based on the inverse-power exceedance is more conservative in the tails, as the estimated lfdr curve is greater than that of Lindsey's method. At the 20% cut-off, there are only 26 estimated lfdr values falling below $\alpha = 0.2$, whereas Lindsey's method makes 54 discoveries. The average value of the lfdr estimates among all z-scores is $0.908 = 1 - \hat{\rho}_n$ for the inverse-power method, compared to 0.929 for Lindsey's method, which reports an estimate of 0.932 for the null proportion.

4.4 Proofs

Proof of Theorem 4.2.1. The log likelihood is

$$\ell(\rho) = \sum_{i=1}^{n} \log \phi(Y_i) + \log\left(1 + \rho\left(\zeta_d(Y_i) - 1\right)\right),$$

so the first and second derivatives are given by

$$\ell'(\rho) = \sum_{i=1}^{n} \frac{\zeta_d(Y_i) - 1}{1 + \rho(\zeta_d(Y_i) - 1)}$$
$$\ell''(\rho) = \sum_{i=1}^{n} \frac{-(\zeta_d(Y_i) - 1)^2}{(1 + \rho(\zeta_d(Y_i) - 1))^2} < 0.$$

The maximum is achieved at $\hat{\rho}_n = 0$ when $\ell'(0) \leq 0$, so that $\hat{\rho}_n > 0$ is equivalent to

$$\ell'(0) = \sum_{i=1}^{n} \zeta_d(Y_i) - n > 0.$$

The event $\sum_{i=1}^{n} \zeta_d(Y_i) > n$ is implied by $\max_{i \leq n} \zeta_d(Y_i) > n$, so the probability can be lower bounded,

$$\begin{aligned}
\mathbb{P}(\hat{\rho}_n > 0) &\geq \mathbb{P}\left(\max_{i \leq n} \zeta_d(Y_i) > n\right) \\
&\geq \mathbb{P}\left(\max_{i \leq n} \frac{ce^{Y_i^2/2}}{|Y_i|^{d+1}} > n\right) \\
&\geq \mathbb{P}\left(\max_{i \leq n} \frac{ce^{Y_i^2/2}}{|Y_i|^{d+1}} > n, \max_{i \leq n} Y_i > \sqrt{2(1+\delta)\log \rho^{-1}}\right)
\end{aligned}$$

for some $c > 0$ by Lemma 4.4.2, as long as n is larger than a universal constant. Now since $n \asymp \frac{(\log \rho^{-1})^{d/2}}{\rho}$, we have

$$\frac{e^{Y_i^2/2}}{|Y_i|^{d+1}} 1_{\{\max Y_i > \sqrt{2(1+\delta)\log \rho^{-1}}\}} \geq \frac{\rho^{-(1+\delta)}}{(\log \rho^{-1})^{(d+1)/2}} = \omega(n),$$

as $\rho \to 0$, which implies that $\liminf_{\rho \to 0} \mathbb{P}(\hat{\rho}_n > 0)$ is lower bounded for any $\delta > 0$ by

$$\liminf_{\rho \to 0} \mathbb{P}(\hat{\rho}_n > 0) \geq \liminf_{\rho \to 0} \mathbb{P}\left(\max_{i \leq n} Y_i > \sqrt{2(1+\delta)\log \rho^{-1}}\right). \tag{4.15}$$

Since the observations are iid, we have as $\rho \to 0$ (and thus $n = b_\gamma(\rho) \to \infty$),

$$\mathbb{P}\left(\max_{i\leq n} Y_i \leq \sqrt{2(1+\delta)\log\rho^{-1}}\right) \sim \left(1 - \mathbb{P}(Y_1 > \sqrt{2(1+\delta)\log\rho^{-1}})\right)^n \tag{4.16}$$

The exceedance probability is at least

$$\begin{aligned}\mathbb{P}(Y_1 > \sqrt{2(1+\delta)\log\rho^{-1}}) &= (1-\rho)\bar{\Phi}(\sqrt{2(1+\delta)\log\rho^{-1}}) + \rho\int_{\sqrt{2(1+\delta)\log\rho^{-1}}}^{\infty}\phi(y)\zeta_d(y)\mathrm{d}y \\ &\geq \rho\int_{\sqrt{2(1+\delta)\log\rho^{-1}}}^{\infty}\phi(y)\zeta_d(y)\mathrm{d}y.\end{aligned}$$

By Lemma 4.4.2, dominated convergence implies that as $\rho \to 0$ we have

$$\begin{aligned}\int_0^\infty \rho\phi(y)\zeta_d(y)1_{\left\{y>\sqrt{2(1+\delta)\log\rho^{-1}}\right\}}\mathrm{d}y &\sim \rho\int_{\sqrt{2(1+\delta)\log\rho^{-1}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\cdot C_d\sqrt{2\pi}\frac{e^{y^2/2}}{|y|^{d+1}}\mathrm{d}y \\ &= \frac{C_d}{d}\frac{\rho}{(2(1+\delta)\log\rho^{-1})^{d/2}}.\end{aligned}$$

Plugging this into (4.16), we have shown

$$\limsup_{\rho\to 0}\mathbb{P}\left(\max_{i\leq n} Y_i \leq \sqrt{2(1+\delta)\log\rho^{-1}}\right) \leq \limsup_{\rho\to 0}\exp\left(-\frac{C_d n\rho}{d(2(1+\delta)\log\rho^{-1})^{d/2}}\right).$$

Together with (4.15), which holds for every $\delta > 0$, this implies

$$\liminf_{\rho\to 0}\mathbb{P}(\hat{\rho}_n > 0) \geq 1 - \lim_{\rho\to 0}\exp\left(-\frac{C_d n\rho}{d2^{d/2}(\log\rho^{-1})^{d/2}}\right) = \gamma.$$

□

Proof of Proposition 4.2.1. Let $w : \mathbb{R} \to \mathbb{R}$ be a function for which $\frac{w(x)}{x^2}$ be continuous and bounded as a function of x, and let P_ρ denote any signal distribution of the form (4.12) with $\beta = \beta^*(r)$. Then

$$\begin{aligned}\int w(x)P_\rho(\mathrm{d}x) &= (1-\sigma^{1-dr})w(0) + \sigma^{1-dr}\int w(x)A_d\cdot\frac{\sigma^{dr}}{(\sigma^{2r}+x^2/d)^{\frac{d+1}{2}}}\mathrm{d}x \\ &= \frac{d^{\frac{d+1}{2}}A_d}{C_d}\cdot\sigma\int w(x)\cdot\frac{C_d}{d^{\frac{d+1}{2}}(\sigma^{2r}+x^2/d)^{\frac{d+1}{2}}}\mathrm{d}x.\end{aligned}$$

It follows that with $\rho = \frac{d^{\frac{d+1}{2}} A_d}{C_d} \cdot \sigma$, we have

$$\rho^{-1} \int w(x) P_\rho(\mathrm{d}x) \to \int_{\mathbb{R}\setminus\{0\}} w(x) H_d(\mathrm{d}x) \quad \text{as } \rho \to 0.$$

□

4.4.1 Second order sparsity approximation

For the convolutional model (4.1) where the signal distribution P_ρ tends to a unit exceedance measure H at rate ρ, the marginal density of Y is

$$m_\rho(y) = \phi(y) \left(1 - \int (1 - e^{-x^2/2}) P_\rho(\mathrm{d}x) + \int (\cosh(xy) - 1) e^{-x^2/2} P_\rho(\mathrm{d}x) \right).$$

When (P_ρ) is the mixture,

$$P_\rho = (1 - \sigma^{1-dr})\delta_0 + \sigma^{1-dr} t_d(\sigma^r), \quad r \in (0, 1/d],$$

the following lemma characterizes the second order-expansion to the terms appearing in the marginal density of Y.

Lemma 4.4.1. *Let $P_\rho = (1 - \sigma^{1-dr})\delta_0 + \sigma^{1-dr} t_d(\sigma^r)$. Then*

$$\int (1 - e^{-x^2/2}) P_\rho(\mathrm{d}x) = \rho - \rho\sigma^{r(2-d)} \cdot \frac{C_d}{A_d} \frac{(d+1) d^{\frac{1-d}{2}}}{4} + o(\rho\sigma^{r(2-d)})$$

$$\int (\cosh(xy) - 1) e^{-x^2/2} P_\rho(\mathrm{d}x) = \rho\zeta_d(y) - \rho\sigma^{r(2-d)} \cdot \frac{C_d}{A_d} \frac{(d+1) d^{\frac{1-d}{2}}}{4} y^2 + o(\rho\sigma^{r(2-d)}),$$

as $\sigma \to 0$, where $C_d, A_d > 0$ are defined in expressions (4.4) *and* (4.5), *and $\rho = \frac{d^{\frac{d+1}{2}} A_d}{C_d} \cdot \sigma$.*

Proof. For the first integral,

$$\begin{aligned}
\int (1 - e^{-x^2/2}) P_\rho(\mathrm{d}x) &= \sigma^{1-dr} A_d \int (1 - e^{-x^2/2}) \cdot \frac{\sigma^{dr}}{(\sigma^{2r} + x^2/d)^{\frac{d+1}{2}}} \mathrm{d}x \\
&= \sigma \cdot \frac{d^{\frac{d+1}{2}} A_d}{C_d} \int (1 - e^{-x^2/2}) \cdot \frac{C_d}{d^{\frac{d+1}{2}} (\sigma^{2r} + x^2/d)^{\frac{d+1}{2}}} \mathrm{d}x \\
&= \rho \int (1 - e^{-x^2/2}) \cdot \frac{C_d}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \mathrm{d}x \\
&= \rho \int (1 - e^{-x^2/2}) \cdot C_d \left(\frac{1}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} - \frac{1}{|x|^{d+1}} + \frac{1}{|x|^{d+1}} \right) \mathrm{d}x.
\end{aligned}$$

Since H_d is a unit exceedance measure, the above is equal to

$$\begin{aligned} &= \rho - \rho \int (1 - e^{-x^2/2}) \cdot C_d \left(\frac{1}{|x|^{d+1}} - \frac{1}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \right) \mathrm{d}x \\ &= \rho - \rho \int (1 - e^{-x^2/2}) \cdot \frac{C_d}{|x|^{d+1}} \left(\frac{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}} - |x|^{d+1}}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \right) \mathrm{d}x. \end{aligned}$$

Now since $(\sigma^{2r} d + x^2)^{\frac{d+1}{2}} - |x|^{d+1} \sim \sigma^{2r} d \cdot \frac{d+1}{2} |x|^{d-1}$ as $\sigma \to 0$, and since

$$d^{\frac{d+1}{2}} \int \frac{1 - e^{-x^2/2}}{x^2} \cdot \frac{A_d \sigma^{rd}}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \mathrm{d}x \to \frac{1}{2},$$

because $t_d(\sigma^r) \to \delta_0$, we have

$$\int (1 - e^{-x^2/2}) \cdot \frac{C_d}{|x|^{d+1}} \left(\frac{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}} - |x|^{d+1}}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \right) \mathrm{d}x \sim \frac{d^{\frac{1-d}{2}} (d+1) C_d}{4 A_d} \sigma^{2r - rd},$$

as $\sigma \to 0$. It now follows that

$$\int (1 - e^{-x^2/2}) P_\rho(\mathrm{d}x) = \rho - \rho \sigma^{r(2-d)} \frac{d^{\frac{1-d}{2}} (d+1) C_d}{4 A_d} + o(\rho \sigma^{r(2-d)}) \quad \text{as } \sigma \to 0.$$

The second integral is similar but we record the derivation here for completeness. By definition,

$$\int (\cosh(xy) - 1) e^{-x^2/2} P_\rho(\mathrm{d}x) = \sigma^{1-dr} A_d \int (\cosh(xy) - 1) \cdot \frac{\sigma^{dr}}{(\sigma^{2r} + x^2/d)^{\frac{d+1}{2}}} \mathrm{d}x.$$

Following the above sequence of manipulations, we find that the above is equal to

$$= \rho \zeta_d(y) - \rho \int (\cosh(xy) - 1) e^{-x^2/2} \cdot \frac{C_d}{|x|^{d+1}} \left(\frac{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}} - |x|^{d+1}}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \right) \mathrm{d}x.$$

Now using the first order Taylor expansion $(\sigma^{2r} d + x^2)^{\frac{d+1}{2}} - |x|^{d+1} \sim \sigma^{2r} d \cdot \frac{d+1}{2} |x|^{d-1}$ as $\sigma \to 0$ and the fact that

$$d^{\frac{d+1}{2}} \int \frac{(\cosh(xy) - 1) e^{-x^2/2}}{x^2} \frac{A_d \sigma^{rd}}{(\sigma^{2r} d + x^2)^{\frac{d+1}{2}}} \mathrm{d}x \to \frac{y^2}{2},$$

we have

$$\int (\cosh(xy) - 1)e^{-x^2/2} P_\rho(\mathrm{d}x) = \rho\zeta_d(y) - \rho\sigma^{r(2-d)} \frac{d^{\frac{1-d}{2}}(d+1)C_d}{4A_d} y^2 + o(\rho\sigma^{r(2-d)}),$$

as $\sigma \to 0$. □

The form of the second order sparsity approximation follows from this lemma, recorded in the following corollary.

Corollary 4.4.1. *If $Y = X + \varepsilon$ where $X \sim (1-\sigma^{1-dr})\delta_0 + \sigma^{1-dr} t_d(\sigma^r)$ for some $r \in (0, 1/d]$, then the marginal density of Y satisfies*

$$m_\rho(y) = (1-\rho)\phi(y) + \rho\psi_d(y) + \rho\sigma^{r(2-d)} \cdot \frac{C_d}{A_d} \frac{d^{\frac{1-d}{2}}(d+1)}{4}(1-y^2) + o(\rho\sigma^{r(2-d)}),$$

as $\sigma \to 0$, where $\rho = \dfrac{d^{\frac{d+1}{2}} A_d}{C_d} \cdot \sigma$.

In Figure 4.1, the first and second order approximations are plotted on the log-scale against the exact marginal density of $Y = X + \varepsilon$ for an inverse-power signal with $d = 1$. For moderate values of ρ, the difference from the second order approximation to the exact density is much smaller than the difference between the first and second order approximations. Higher order terms are straightforward to derive and bear close connections with the even-degree Hermite polynomials, but we do not pursue them here. For practical purposes, the first and second order sparsity expansions suffice for approximating the marginal density.

Proof of Proposition 4.3.1. For the first part, since

$$\frac{P_\rho(|X| \le \delta)}{Q_\rho(|X| \le \delta)} = \frac{1 - P_\rho(|X| > \delta)}{1 - Q_\rho(|X| > \delta)},$$

it suffices to show that $P_\rho(|X| > \delta) \to 0$ for any P_ρ which is sparse at rate ρ and exceedance H. Since $\rho \to 0$, by Markov's inequality,

$$P_\rho(|X| > \delta) = \mathbb{P}_\rho(X^2 \wedge 1 > \delta^2) \le \frac{\mathbb{E}_\rho(X^2 \wedge 1)}{\delta^2},$$

where we have assumed wlog that $\delta < 1$. Now since $w(x) = x^2 \wedge 1$ is continuous and bounded,

$$\frac{\mathbb{E}_\rho(X^2 \wedge 1)}{\delta^2} = \frac{\rho \int (x^2 \wedge 1) H(\mathrm{d}x) + o(\rho)}{\delta^2} \lesssim \rho^{2\beta} \to 0,$$

since $H(\mathrm{d}x)$ is a unit exceedance measure, and $\delta \ge \rho^{\frac{1}{2}-\beta}$.

For the second part, let G_ρ be any sparse measure tending to $H(\mathrm{d}x)$ at rate ρ, i.e.

$$\rho^{-1} \int w(x) G_\rho(\mathrm{d}x) \to \int w(x) H(\mathrm{d}x)$$

for every $w \in \mathcal{W} \coloneqq \{w \text{ continuous, bounded, and } w(x) = O(x^2) \text{ near the origin}\}$. Let

$$\begin{aligned} P_\rho &= (1-\varepsilon/2)\delta_0 + (\varepsilon/2)G_\rho \\ Q_\rho &= (1-\varepsilon/2)\delta_{\pm\rho^{\frac{1}{2}+\beta/2}} + (\varepsilon/2)G_\rho \end{aligned}$$

where $\delta_{\pm\rho^{\frac{1}{2}+\beta/2}}$ puts 1/2 probability on each of $\rho^{\frac{1}{2}+\beta/2}$ and $-\rho^{\frac{1}{2}+\beta/2}$. For any $w \in \mathcal{W}$, we have

$$\rho^{-1} \int w(x) P_{2\rho/\varepsilon}(\mathrm{d}x) = (2\rho/\varepsilon)^{-1} \int w(x) G_{2\rho/\varepsilon}(\mathrm{d}x) \to \int w(x) H(\mathrm{d}x),$$

as $\rho \to 0$. Thus $P_{2\rho/\varepsilon}$ is sparse with rate ρ and exceedance $H(\mathrm{d}x)$. Similarly,

$$\begin{aligned} \rho^{-1} \int w(x) Q_{2\rho/\varepsilon}(\mathrm{d}x) &= \rho^{-1} \left[(1-\varepsilon/2) \cdot \frac{w(\rho^{\frac{1}{2}+\beta/2}) + w(-\rho^{\frac{1}{2}+\beta/2})}{2} + (\varepsilon/2) \int w(x) G_{2\rho/\varepsilon}(\mathrm{d}x) \right] \\ &= \rho^{-1} \left[o(\rho) + (2/\varepsilon)^{-1} \int w(x) G_{2\rho/\varepsilon}(\mathrm{d}x) \right] \\ &\qquad\qquad\qquad\qquad (w(x) = O(x^2) \text{ as } x \to 0) \\ &= (2\rho/\varepsilon)^{-1} \int w(x) G_{2\rho/\varepsilon}(\mathrm{d}x) + o(1) \\ &\to \int w(x) H(\mathrm{d}x), \end{aligned}$$

so that $Q_{2\rho/\varepsilon}$ is sparse with rate ρ and exceedance measure $H(\mathrm{d}x)$. Now since $\delta < \rho^{\frac{1}{2}+\beta} < \rho^{\frac{1}{2}+\beta/2}$,

$$\begin{aligned} P_{2\rho/\varepsilon}([-\delta, \delta]) &\geq 1 - \varepsilon/2 \\ Q_{2\rho/\varepsilon}([-\delta, \delta]) &\leq \varepsilon/2, \end{aligned}$$

so that the ratio is small,

$$\frac{Q_{2\rho/\varepsilon}([-\delta, \delta])}{P_{2\rho/\varepsilon}([-\delta, \delta])} \leq \frac{\varepsilon/2}{1-\varepsilon/2} < \varepsilon. \qquad (\varepsilon < 1)$$

□

4.4.2 Probability metric calculations

Proof of Lemma 4.2.1. Let P_0 denote ϕ and P_1 denote $(1-\rho)\phi+\rho\psi_d$. First we show a lower bound,

$$\begin{aligned}
H^2(P_0, P_1) &= \mathbb{E}_0\left(\sqrt{\frac{P_1}{P_0}(Y)} - 1\right)^2 \\
&= \mathbb{E}_0\left(\sqrt{\frac{(1-\rho)\phi + \rho\psi_d}{\phi}(Y)} - 1\right)^2 \\
&= \mathbb{E}_0\left(\sqrt{1 + \rho\left(\zeta_d(Y) - 1\right)} - 1\right)^2 \\
&\geq \mathbb{E}_0\left(\sqrt{1 + \rho\left(\zeta_d(Y) - 1\right)} - 1\right)^2 1_{\{\zeta_d(Y)-1\geq\rho^{-1}\}}.
\end{aligned}$$

Using the inequality $(\sqrt{1+t}-1)^2 \geq (\sqrt{2}-1)^2(t\wedge t^2)$, the above is lower bounded,

$$\begin{aligned}
&\geq (\sqrt{2}-1)^2\rho\mathbb{E}_0\left(\zeta_d(Y) - 1\right) 1_{\{\zeta_d(Y)-1\geq\rho^{-1}\}} \\
&\geq (\sqrt{2}-1)^2\rho\mathbb{E}_0\left(\frac{e^{Y^2/2}}{8Y^{1+d}} - 1\right) 1_{\{\zeta_d(Y)-1\geq\rho^{-1}\}},
\end{aligned}$$

where we have used Lemma 4.4.2 for the above inequality. Since $\rho \to 0$, the above is eventually greater than

$$\geq (\sqrt{2}-1)^2\rho\mathbb{E}_0\left(\frac{e^{Y^2/2}}{16Y^{1+d}}\right) 1_{\{\zeta_d(Y)\geq 2\rho^{-1}\}},$$

as $\rho \to 0$. Integrating against the standard normal density gives

$$\begin{aligned}
H^2(P_0, P_1) &\geq \frac{(\sqrt{2}-1)^2}{16}\rho\int_{\{y:\zeta_d(y)\geq 2\rho^{-1}\}} \frac{e^{y^2/2}}{y^{1+d}}\phi(y)\mathrm{d}y \\
&\geq \frac{(\sqrt{2}-1)^2}{16\sqrt{2\pi}}\rho\int_{\{y:\zeta_d(y)\geq 2\rho^{-1}\}} \frac{1}{y^{1+d}}\mathrm{d}y \\
&\geq \frac{(\sqrt{2}-1)^2}{16\sqrt{2\pi}}\rho\int_{\{y:e^{y^2/2}y^{-(1+d)}\geq 16\rho^{-1}\}} \frac{1}{y^{1+d}}\mathrm{d}y \\
&\geq \frac{(\sqrt{2}-1)^2}{16\sqrt{2\pi}}\rho\int_{\sqrt{4\log\rho^{-1}}}^{\infty} \frac{1}{y^{1+d}}\mathrm{d}y.
\end{aligned}$$

Evaluating the integral gives

$$H^2(P_0,P_1) \gtrsim \frac{\rho}{(\log\rho^{-1})^{d/2}}, \quad \text{as } \rho\to 0. \tag{4.17}$$

For the upper bound, note

$$\begin{aligned} H^2(P_0,P_1) &= \mathbb{E}_0\left(\sqrt{1+\rho\,(\zeta_d(Y)-1)}-1\right)^2 \\ &\leq \rho^2 + \mathbb{E}_0\left(\sqrt{1+\rho\zeta_d(Y)}-1\right)^2 1_{\{\zeta_d(Y)>1\}}, \end{aligned}$$

since when $\zeta_d(Y)\leq 1$, it holds that $\left(\sqrt{1+\rho(\zeta_d(Y)-1)}-1\right)^2 \leq (1-\sqrt{1-\rho})^2 \leq \rho^2$. Using the inequality $(\sqrt{1+t}-1)^2 \leq t\wedge t^2$, the above implies

$$H^2(P_0,P_1) \leq \rho^2 + \mathbb{E}_0(\rho\zeta_d(Y))1_{\{\zeta_d(Y)>\rho^{-1}\}} + \mathbb{E}_0(\rho\zeta_d(Y))^2 1_{\{\zeta_d(Y)\leq\rho^{-1}\}} \tag{4.18}$$

By Lemma 4.4.2, the second term in (4.18) is

$$\begin{aligned} \rho\mathbb{E}_0\zeta_d(Y)1_{\{\zeta_d(Y)>\rho^{-1}\}} &\leq 2\rho\mathbb{E}_0\frac{e^{Y^2/2}}{Y^{1+d}}1_{\{\zeta_d(Y)>\rho^{-1}\}} \\ &\leq 2\rho\mathbb{E}_0\frac{e^{Y^2/2}}{Y^{1+d}}1_{\{|Y|>\sqrt{2\log\rho^{-1}}\}}, \end{aligned}$$

where the second inequality follows since on the event $\zeta_d(Y) > \rho^{-1}$, we have $e^{Y^2/2} > \zeta_d(Y) > \rho^{-1}$. Integrating gives

$$\rho\mathbb{E}_0\zeta_d(Y)1_{\{\zeta_d(Y)>\rho^{-1}\}} \leq \frac{4\rho}{\sqrt{2\pi}}\int_{\sqrt{2\log\rho^{-1}}}^{\infty}\frac{1}{y^{1+d}}\mathrm{d}y \lesssim \frac{\rho}{(\log\rho^{-1})^{d/2}}, \quad \text{as } \rho\to 0. \tag{4.19}$$

By Lemma 4.4.4, the third term in (4.18) is

$$\rho^2\mathbb{E}_0\zeta_d(Y)^2 1_{\{\zeta_d(Y)\leq\rho^{-1}\}} \leq \rho^2\mathbb{E}_0\left(C+|Y|+\frac{ce^{Y^2/2}}{|Y|^{1+d}}\cdot 1_{\{|Y|>1\}}\right)^2 1_{\{\zeta_d(Y)\leq\rho^{-1}\}},$$

For any $a,b,c>0$, one has $(a+b+c)^2 \leq 4(a^2+b^2+c^2)$, so the above is bounded by

$$\leq 4\rho^2\left[C^2+1+c^2\cdot\underbrace{\mathbb{E}_0\left(\frac{e^{Y^2}}{|Y|^{2(1+d)}}\cdot 1_{\{|Y|>1,\zeta_d(Y)\leq\rho^{-1}\}}\right)}_{(*)}\right]. \tag{4.20}$$

Now by Lemma 4.4.5,

$$\begin{aligned}1_{\{\zeta_d(Y)\le\rho^{-1}\}} &= 1_{\{\zeta_d(Y)\le\rho^{-1}\}\cap\{|Y|\le\sqrt{2\log\rho^{-1}}\}} + 1_{\{\zeta_d(Y)\le\rho^{-1}\}\cap\{|Y|>\sqrt{2\log\rho^{-1}}\}}\\ &\le 1_{\{|Y|\le\sqrt{2\log\rho^{-1}}\}} + 1_{\{\sqrt{2\log\rho^{-1}}<|Y|\le\sqrt{2(1+c_\rho)\log\rho^{-1}}\}},\end{aligned}$$

where $c_\rho = \frac{1+d}{2}\cdot\frac{\log\log\rho^{-1}}{\log\rho^{-1}}$ as $\rho\to 0$. This implies the last term in (4.20) is bounded

$$\begin{aligned}(*) &\le \mathbb{E}_0\left(\frac{e^{Y^2}}{|Y|^{2(1+d)}}\left(1_{\{1\le|Y|\le\sqrt{2\log\rho^{-1}}\}} + 1_{\{\sqrt{2\log\rho^{-1}}<|Y|\le\sqrt{2(1+c_\rho)\log\rho^{-1}}\}}\right)\right)\\ &= \frac{2}{\sqrt{2\pi}}\left[\int_1^{\sqrt{2\log\rho^{-1}}}\frac{e^{y^2/2}}{y^{2(1+d)}}\mathrm{d}y + \int_{\sqrt{2\log\rho^{-1}}}^{\sqrt{2(1+c_\rho)\log\rho^{-1}}}\frac{e^{y^2/2}}{y^{2(1+d)}}\mathrm{d}y\right]\\ &\le \frac{2}{\sqrt{2\pi}}\cdot\sqrt{2\log\rho^{-1}}\cdot\frac{\rho^{-1}}{(2\log\rho^{-1})^{1+d}} + \frac{2}{\sqrt{2\pi}}\sqrt{2\log\rho^{-1}}\cdot(\sqrt{1+c_\rho}-1)\cdot\frac{\rho^{-(1+c_\rho)}}{(2\log\rho^{-1})^{1+d}}\\ &\le \frac{2}{\sqrt{2\pi}}\left[\frac{\rho^{-1}}{(\log\rho^{-1})^{d+1/2}} + c_\rho\cdot\frac{\rho^{-(1+c_\rho)}}{(\log\rho^{-1})^{d+1/2}}\right].\end{aligned}$$

Now since $\rho^{-c_\rho} = (\rho^{-1})^{c_\rho} = (\log\rho^{-1})^{\frac{1+d}{2}}$, the above becomes

$$\lesssim \frac{\rho^{-1}}{(\log\rho^{-1})^{d+1/2}} + \frac{\rho^{-1}\cdot(\log\rho^{-1})^{\frac{1+d}{2}}}{(\log\rho^{-1})^{d+1/2}} \le \frac{2\rho^{-1}}{(\log\rho^{-1})^{d/2}},$$

since $d+1/2\ge d/2$ for any $d>0$. Plugging this into (4.20) gives

$$\rho^2\mathbb{E}_0\zeta_d(Y)^2 1_{\{\zeta_d(Y)\le\rho^{-1}\}} \lesssim \rho^2 + \rho^2\cdot\frac{\rho^{-1}}{(\log\rho^{-1})^{d/2}} \lesssim \frac{\rho}{(\log\rho^{-1})^{d/2}}.$$

Combining this with (4.18) and (4.19) completes the proof of the upper bound. Together with the lower bound (4.17), this completes the proof of the lemma.

□

Proposition 4.4.1. *Let $r\in(0,1/d]$ where $d\in(0,2)$ and denote by P_ρ and Q_ρ the first and second order sparsity approximations to the signal distribution $(1-\sigma^{1-dr})\delta_0+\sigma^{1-dr}t_d(\sigma^r)$, with densities*

$$\begin{aligned}P_\rho(\mathrm{d}y) &= (1-\rho+\rho\zeta_d(y))\,\phi(y)\mathrm{d}y\\ Q_\rho(\mathrm{d}y) &= \left(1-\rho+\rho\zeta_d(y)+c\rho\sigma^{r(2-d)}(1-y^2)\right)\phi(y)\mathrm{d}y,\end{aligned}$$

where $c := \frac{C_d(d+1)d^{\frac{1-d}{2}}}{4A_d}$ *and* ρ *is defined in Lemma 4.2.1. Then the Hellinger distance is*

$$H^2(P_\rho, Q_\rho) = \widetilde{O}\left(\rho^{2(1+r(2-d))}\right),$$

where $\widetilde{O}$ *means up to multiplicative constants and poly-logarithmic factors in* ρ^{-1}.

Note that for $r > 0$, the squared Hellinger distance between the first and second-order approximations is $O(\rho^2)$, which is smaller than that between the zero and first-order approximations by a factor of ρ (Lemma 4.2.1).

Proof of Proposition 4.4.1. Let p_ρ and q_ρ denote the densities of P_ρ and Q_ρ respectively. The squared Hellinger distance is

$$H^2(P_\rho, Q_\rho) = \mathbb{E}\left(\sqrt{\frac{q_\rho}{p_\rho}(Y)} - 1\right)^2,$$

where the expectation is taken with respect to the first order approximation P_ρ. The difference between the first and second order sparsity approximations to $m_\rho(y)$ is

$$q_\rho(y) - p_\rho(y) = c\rho\sigma^{r(2-d)}(1-y^2)\phi(y),$$

so the Hellinger distance can be written

$$\begin{aligned} H^2(P_\rho, Q_\rho) &= \mathbb{E}\left(\sqrt{1 + \frac{q_\rho - p_\rho}{p_\rho}(Y)} - 1\right)^2 \\ &= \mathbb{E}\left(\sqrt{1 + c\rho\sigma^{r(2-d)} \cdot \frac{(1-Y^2)}{1-\rho+\rho\zeta_d(Y)}} - 1\right)^2 1_{\{\zeta_d(Y) \le \rho^{-1}\}} \qquad (4.21) \\ &+ \mathbb{E}\left(\sqrt{1 + c\rho\sigma^{r(2-d)} \cdot \frac{(1-Y^2)}{1-\rho+\rho\zeta_d(Y)}} - 1\right)^2 1_{\{\zeta_d(Y) > \rho^{-1}\}}. \qquad (4.22) \end{aligned}$$

Using $\sqrt{1+\varepsilon} - 1 \sim \varepsilon/2$ as $\varepsilon \to 0$, the first piece (4.21) is

$$\lesssim \mathbb{E}\left(\rho\sigma^{r(2-d)} \log \rho^{-1} 1_{\{\zeta_d(Y) \le \rho^{-1}\}}\right)^2 \asymp \rho^{2+2r(2-d)}(\log \rho^{-1})^2,$$

where we have used Lemma 4.4.4 to bound Y^2 by the order of $\log \rho^{-1}$ on the event $\zeta_d(y) \le \rho^{-1}$. On the event $\zeta_d(Y) > \rho^{-1}$ we have $|Y| \to \infty$ as $\rho \to 0$, so Lemma 4.4.2 implies the second piece (4.22) is bounded

$$\lesssim \mathbb{E}\left(\sigma^{r(2-d)}|Y|^{2+d+1}e^{-Y^2/2}\right)^2 1_{\{\zeta_d(Y) > \rho^{-1}\}} \lesssim \rho^{2+2r(2-d)}(\log \rho^{-1})^{3+d},$$

since $|y|^{3+d}e^{-y^2/2}$ is decreasing in $|y|$ as $|y| \to \infty$. Thus we have shown that

$$H^2(P_\rho, Q_\rho) \lesssim \rho^{2(1+r(2-d))}(\log \rho^{-1})^{3+d} \quad \text{as } \rho \to 0,$$

so the claim holds up to poly-logarithmic factors in ρ^{-1}.

□

4.4.3 Technical lemmas

Lemma 4.4.2. *Let $d \in (0,2)$ and $\zeta_d(y) := \int (\cosh(xy) - 1)e^{-x^2/2}H_d(\mathrm{d}x)$. Then for any $y \in \mathbb{R}$ with $|y|$ larger than some universal constant, there exist $c, C > 0$ satisfying*

$$\frac{ce^{y^2/2}}{|y|^{1+d}} \le \zeta_d(y) \le \frac{Ce^{y^2/2}}{|y|^{1+d}}.$$

As $|y| \to \infty$, we have $\zeta_d(y) \sim C_d\sqrt{2\pi} \cdot \frac{e^{y^2/2}}{|y|^{d+1}}$, where $C_d = \frac{d2^{d/2-1}}{\Gamma(1-d/2)}$.

Proof. For the lower bound, write

$$\begin{aligned}
\zeta_d(y) &= 2\int_0^\infty (\cosh(yx) - 1)e^{-x^2/2}H_d(\mathrm{d}x) \\
&= 2\int_0^\infty \left(\frac{e^{|yx|} + e^{-|yx|}}{2} - 1\right) e^{-x^2/2} \cdot \frac{d2^{d/2-1}}{\Gamma(1-d/2)|x|^{d+1}}\mathrm{d}x \\
&= 2C_d\sqrt{2\pi}\int_0^\infty \left(\frac{e^{|yx|} + e^{-|yx|}}{2} - 1\right) e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}|x|^{d+1}}\mathrm{d}x,
\end{aligned}$$

where C_d is defined in (4.4). Now since $\frac{e^{|yx|}}{2} - 1 \ge \frac{e^{|yx|}}{4}$ is equivalent to $x \ge \frac{\log 4}{|y|}$, the above is lower bounded,

$$\begin{aligned}
&\ge 2C_d\sqrt{2\pi}\int_{\frac{\log 4}{|y|}}^\infty \frac{e^{|y|x}}{4} \cdot e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}|x|^{d+1}}\mathrm{d}x \\
&= \frac{C_d\sqrt{2\pi}}{2}\int_{\frac{\log 4}{|y|}}^\infty e^{|y|x - x^2/2} \cdot \frac{1}{\sqrt{2\pi}|x|^{d+1}}\mathrm{d}x \\
&= \frac{C_d\sqrt{2\pi}}{2}e^{y^2/2}\int_{\frac{\log 4}{|y|}}^\infty \phi(x - |y|) \cdot \frac{1}{|x|^{d+1}}\mathrm{d}x.
\end{aligned}$$

Recognizing the integral as an expectation, the above can be rewritten,

$$= \frac{C_d\sqrt{2\pi}}{2}e^{y^2/2}\mathbb{E}\left(\frac{1}{||y| + Z|^{d+1}} \cdot 1_{\left\{Z + |y| \ge \frac{\log 4}{|y|}\right\}}\right),$$

where the expectation is over $Z \sim N(0,1)$. Factoring out $|y|^{-(d+1)}$ gives

$$= \frac{C_d\sqrt{2\pi}}{2}\frac{e^{y^2/2}}{|y|^{d+1}}\mathbb{E}\left(\frac{1}{|1+Z/|y||^{d+1}}\cdot 1_{\left\{Z+|y|\geq \frac{\log 4}{|y|}\right\}}\right)$$
$$\geq \frac{C_d\sqrt{2\pi}}{2}\frac{e^{y^2/2}}{|y|^{d+1}}\mathbb{E}\left(\frac{1}{|1+Z/|y||^{d+1}}\cdot 1_{\{Z>0\}}\right),$$

for any $|y| > 1$. If $|y| \to \infty$, then by Fatou's Lemma,

$$\liminf_{\rho\to 0}\mathbb{E}\left(\frac{1}{(1+Z/|y|)^{d+1}}1_{\{Z>0\}}\right) \geq \mathbb{P}(Z>0) = \frac{1}{2}.$$

It follows that for $|y|$ larger than some universal constant,

$$\mathbb{E}\left(\frac{1}{|1+Z/|y||^{d+1}}\cdot 1_{\{Z>0\}}\right) \geq \frac{1}{4},$$

which implies the lower bound,

$$\zeta_d(y) \geq \frac{ce^{y^2/2}}{|y|^{d+1}}, \quad c := \frac{C_d\sqrt{2\pi}}{8}.$$

The upper bound is implied by Lemma 4.4.4. The asymptotic statement follows from Laplace approximation and the definition of $\zeta_d(y)$; the ratio $\zeta_d(y)/(C_d\sqrt{2\pi}\cdot\frac{e^{y^2/2}}{|y|^{d+1}})$ is plotted for three values $d = 0.5, 1, 1.5$ in Figure 4.3.

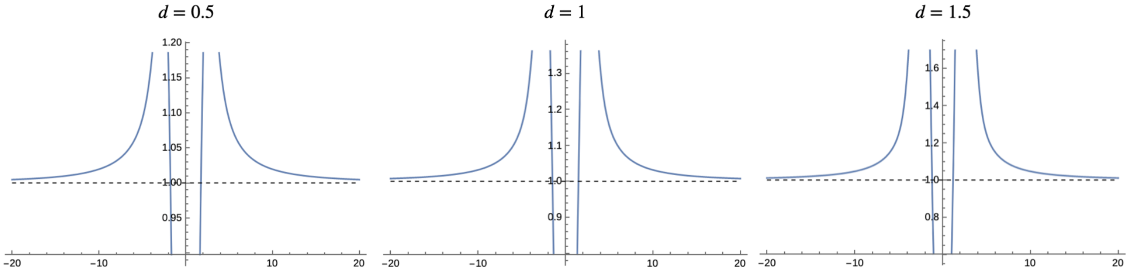


Figure 4.3: The ratio $\frac{\zeta_d(y)}{C_d\sqrt{2\pi}e^{y^2/2}/|y|^{d+1}}$ is plotted for three values of $d \in \{0.5, 1, 1.5\}$ with a dotted horizontal line at 1, approached by all three curves as $|y|$ gets large.

$\square$

Lemma 4.4.3. *Let $d \in (0,2)$ and $\zeta_d(y) := \int(\cosh(xy)-1)e^{-x^2/2}H_d(\mathrm{d}x)$, where H_d is the unit exceedance measure defined by (4.4), and suppose that $Y \sim N(0,1)$. Then there exist* *constants $c_1, c_2 > 0$ such that for all $x > 0$ larger than some universal constant,*

$$\frac{c_1}{x(\log x)^{d/2+1}} \leq \mathbb{P}(\zeta_d(Y) > x) \leq \frac{c_2}{x(\log x)^{d/2+1}}$$

Proof. By Lemma 4.4.2, for x larger than a universal constant, we have for some constants $c, C > 0$

$$\mathbb{P}\left(\frac{ce^{Y^2/2}}{|Y|^{d+1}} > x\right) \leq \mathbb{P}(\zeta_d(Y) > x) \leq \mathbb{P}\left(\frac{Ce^{Y^2/2}}{|Y|^{d+1}} > x\right).$$

Now note that

$$\frac{Ce^{Y^2/2}}{|Y|^{d+1}} > x \iff |Y| > \sqrt{2\log(x|Y|^{d+1}/C)},$$

which implies that $|Y| > \sqrt{2\log(x/C)}$ for large enough x, and thus

$$|Y| > \sqrt{2\left(\log x + \frac{d+1}{2}\log\left(\frac{2}{C}\log\frac{x}{C}\right)\right)}.$$

It follows from Mill's ratio that

$$\begin{aligned}\mathbb{P}(\zeta_d(Y) > x) &\leq \mathbb{P}\left(\frac{Ce^{Y^2/2}}{|Y|^{d+1}} > x\right)\\ &\leq 2\mathbb{P}\left(Y > \sqrt{2\log x + (d+1)\log\left(\frac{2}{C}\log\frac{x}{C}\right)}\right)\\ &\leq \frac{c_2}{\sqrt{\log x}} \cdot \exp\left(-\log x - \frac{d+1}{2}\log\log x\right),\end{aligned}$$

for some constant $c_2 > 0$. Simplifying gives

$$\mathbb{P}(\zeta_d(Y) > x) \leq \frac{c_2}{x(\log x)^{d/2+1}}.$$

For the lower bound, note that

$$\frac{ce^{Y^2/2}}{|Y|^{d+1}} > x \iff \frac{|Y|^2}{2} - \log|Y|^{d+1} > \log(x/c), \tag{4.23}$$

and the function $g(y) \coloneqq y^2/2 - \log(y^{d+1})$ is strictly increasing in y for $y > \sqrt{d+1}$. Thus, $g(|Y|) > \log(x/c)$ is implied by $|Y| > y^*$ for some $y^* > \sqrt{d+1}$. Now notice that for

$y^* := \sqrt{2\log(x/c) + (d+1)\log(c^*\log x)}$, we have

$$\begin{aligned} g(y^*) &= \log(x/c) + \frac{d+1}{2}\log(c^*\log x) - \frac{d+1}{2}\log\left(2\log(x/c) + (d+1)\log(c^*\log x)\right) \\ &= \log(x/c) + \frac{d+1}{2}\log\left(\frac{c^*\log x}{2\log(x/c) + (d+1)\log(c^*\log x)}\right) \\ &> \log(x/c) \end{aligned}$$

for e.g. $c^* = 3$, as $x \to \infty$. It follows that the condition (4.23) is implied by the event $|Y| > \sqrt{2\log(x/c) + (d+1)\log(3\log x)}$. Together with Mill's ratio, this gives

$$\begin{aligned} \mathbb{P}(\zeta_d(Y) > x) &\geq \mathbb{P}\left(\frac{ce^{Y^2/2}}{|Y|^{d+1}} > x\right) \\ &\geq 2\mathbb{P}\left(Y > \sqrt{2\log(x/c) + (d+1)\log(3\log x)}\right) \\ &\geq \frac{c_1}{\sqrt{\log x}} \cdot \exp\left(-\log x - \frac{d+1}{2}\log\log x\right), \end{aligned}$$

for some constant $c_1 > 0$. Simplifying gives

$$\mathbb{P}(\zeta_d(Y) > x) \geq \frac{c_1}{x(\log x)^{d/2+1}},$$

completing the proof. □

Lemma 4.4.4. *Let $d \in (0,2)$ and $\zeta_d(y) := \int(\cosh(xy) - 1)e^{-x^2/2}H_d(\mathrm{d}x)$. Then there exist constants $c, C > 0$ such that for any $y \in \mathbb{R}$,*

$$\zeta_d(y) \leq C(1 + |y|^d) + \frac{ce^{y^2/2}}{|y|^{1+d}} \cdot 1_{\{|y|>1\}}. \tag{4.24}$$

Proof. The function $\zeta_d(y)$ can be upper bounded,

$$\begin{aligned} \zeta_d(y) &= 2\int_0^\infty (\cosh(xy) - 1)e^{-x^2/2}H_d(\mathrm{d}x) \\ &\leq 2\int_0^{|y|^{-1}} \frac{(|y|x)^2}{2}e^{-x^2/2}H_d(\mathrm{d}x) + 2\int_{|y|^{-1}}^\infty e^{|y|x - x^2/2}H_d(\mathrm{d}x) \\ &= \int_0^{|y|^{-1}} (|y|x)^2 e^{-x^2/2}H_d(\mathrm{d}x) + 2\sqrt{2\pi}e^{y^2/2}\int_{|y|^{-1}}^\infty \phi(x - |y|)H_d(\mathrm{d}x). \end{aligned}$$

If $|y| \leq 1$, then $x > |y|^{-1}$ implies $\phi(x - |y|) \cdot \frac{1}{|x|^{d+1}} \leq \phi(0) \cdot |y|^{d+1} \leq 1$. The above then implies

$$\begin{aligned}\zeta_d(y)1_{\{|y|\leq 1\}} &\leq y^2\int_0^{|y|^{-1}} x^2e^{-x^2/2}H_d(\mathrm{d}x) + 2\sqrt{2\pi}e^{1/2}C_d\\ &= C_d\left(y^2\int_0^{|y|^{-1}} x^{1-d}e^{-x^2/2}\mathrm{d}x + 2\sqrt{2\pi}e^{1/2}\right)\\ &\leq C_d\left(y^2\int_0^{|y|^{-1}} x^{1-d}\mathrm{d}x + 2\sqrt{2\pi}e^{1/2}\right).\end{aligned}$$

Evaluating the integral gives

$$\zeta_d(y)1_{\{|y|\leq 1\}} \leq \frac{C_d}{2-d}\left(y^2(|y|^{-1})^{2-d} + 2\sqrt{2\pi}e^{1/2}\right) = \frac{C_d}{2-d}\left(|y|^d + 2\sqrt{2\pi}e^{1/2}\right). \tag{4.25}$$

If $|y| > 1$, then the above calculations give

$$\begin{aligned}\zeta_d(y)1_{\{|y|>1\}} &\leq \frac{C_d}{2-d}\cdot|y|^d + 2\sqrt{2\pi}e^{y^2/2}\int_{|y|^{-1}}^{\infty}\phi(x-|y|)H_d(\mathrm{d}x)\\ &= \frac{C_d}{2-d}\cdot|y|^d + 2C_d\sqrt{2\pi}e^{y^2/2}\int_{|y|^{-1}}^{\infty}\phi(x-|y|)\frac{1}{|x|^{d+1}}\mathrm{d}x\\ &= \frac{C_d}{2-d}\cdot|y|^d + 2C_d\sqrt{2\pi}\cdot\frac{e^{y^2/2}}{|y|^{d+1}}\int_{|y|^{-1}}^{\infty}\phi(x-|y|)\left|\frac{y}{x}\right|^{d+1}\mathrm{d}x.\end{aligned}$$

Substituting $z = x - |y| \Rightarrow \mathrm{d}x = \mathrm{d}z$, the above becomes

$$= \frac{C_d}{2-d}\cdot|y|^d + 2C_d\sqrt{2\pi}\cdot\frac{e^{y^2/2}}{|y|^{d+1}}\int_{|y|^{-1}-|y|}^{\infty}\phi(z)\left|\frac{1}{1+z/|y|}\right|^{d+1}\mathrm{d}z. \tag{4.26}$$

The integral is bounded,

$$\begin{aligned}\int_{|y|^{-1}-|y|}^{\infty}\phi(z)\left|\frac{1}{1+z/|y|}\right|^{d+1}\mathrm{d}z &\leq \int_{|y|^{-1}-|y|}^{0}\phi(z)\left|\frac{1}{1+z/|y|}\right|^{d+1}\mathrm{d}z + \frac{1}{2}\\ &= \int_0^{|y|-|y|^{-1}}\phi(z)\left|\frac{1}{1-z/|y|}\right|^{d+1}\mathrm{d}z + \frac{1}{2}.\end{aligned} \tag{4.27}$$

If $1 < |y| < 5$, then (4.27) is bounded by a constant,

$$\leq (|y| - |y|^{-1})\phi(0)5^{2(d+1)} + \frac{1}{2} \leq (5-1/5)\phi(0)5^6 + \frac{1}{2}.$$

If $|y| > 5$, then $\sqrt{4(d+1)\log|y|} < |y| - |y|^{-1}$, and (4.27) can be split

$$\begin{aligned}
&= \int_0^{\sqrt{4(d+1)\log|y|}} \phi(z)\frac{1}{(1-z/|y|)^{d+1}}\mathrm{d}z + \int_{\sqrt{4(d+1)\log|y|}}^{|y|-|y|^{-1}} \phi(z)\frac{1}{(1-z/|y|)^{d+1}}\mathrm{d}z + \frac{1}{2} \\
&\le \frac{1}{(1-\sqrt{4(d+1)\log|y|}/|y|)^{d+1}} + \int_{\sqrt{4(d+1)\log|y|}}^{|y|-|y|^{-1}} \phi(z)\frac{1}{(1-z/|y|)^{d+1}}\mathrm{d}z + \frac{1}{2} \\
&\le \frac{1}{(1-\sqrt{4(d+1)\log|y|}/|y|)^{d+1}} + y^{2(d+1)}\bar{\Phi}\left(\sqrt{4(d+1)\log|y|}\right) + \frac{1}{2} \\
&\le \frac{1}{(1-\sqrt{4(d+1)\log|y|}/|y|)^{d+1}} + \frac{1}{\sqrt{4(d+1)\log|y|}} + \frac{1}{2}. \qquad \text{(Mill's ratio)}
\end{aligned}$$

Since $\frac{\sqrt{\log|y|}}{|y|}$ is decreasing in $|y|$ when $|y| > 5$, and the above is also bounded by a constant,

$$\le \left(\frac{1}{1-\sqrt{12\log 5/5}}\right)^3 + \frac{1}{\sqrt{12\log 5}} + \frac{1}{2}.$$

It follows that (4.27) is bounded by a constant whenever $|y| > 1$, so that (4.26) yields the bound

$$\zeta_d(y)1_{\{|y|>1\}} \le \left(\frac{C_d}{2-d}\cdot|y|^d + 2C_d\sqrt{2\pi}C'\cdot\frac{e^{y^2/2}}{|y|^{d+1}}\right)\cdot 1_{\{|y|>1\}},$$

for some $C' > 0$. Combined with (4.25), we have shown

$$\begin{aligned}
\zeta_d(y)1_{\{|y|\le1\}} + \zeta_d(y)1_{\{|y|>1\}} &\le \frac{C_d}{2-d}\left(|y|^d + 2\sqrt{2\pi e}\right) + 2C_d\sqrt{2\pi}C'\cdot\frac{e^{y^2/2}}{|y|^{d+1}}\cdot 1_{\{|y|>1\}} \\
&= \frac{2C_d\sqrt{2\pi e}}{2-d} + \frac{C_d}{2-d}\cdot|y|^d + 2C_d\sqrt{2\pi}C'\cdot\frac{e^{y^2/2}}{|y|^{d+1}}\cdot 1_{\{|y|>1\}} \\
&\le \frac{2C_d\sqrt{2\pi e}}{2-d}\left(1+|y|^d\right) + 2C_d\sqrt{2\pi}C'\cdot\frac{e^{y^2/2}}{|y|^{d+1}}\cdot 1_{\{|y|>1\}},
\end{aligned}$$

so the inequality (4.24) is satisfied with $C := \frac{2C_d\sqrt{2\pi e}}{2-d}$ and $c := 2C_d\sqrt{2\pi}C'$, where

$$C' = \left[\left(\frac{1}{1-\sqrt{12\log 5/5}}\right)^3 + \frac{1}{\sqrt{12\log 5}} + \frac{1}{2}\right] \vee \left[(5-1/5)\phi(0)5^6 + \frac{1}{2}\right].$$

□

Lemma 4.4.5. *Let $d \in (0,2)$ and $\zeta_d(y) := \int(\cosh(xy)-1)e^{-x^2/2}H_d(\mathrm{d}x)$, and suppose $\rho > 0$*

tends to zero. Then for any $y \in \mathbb{R}$ such that $|y| > \sqrt{2\log\rho^{-1}}$,

$$\zeta_d(y) \leq \rho^{-1} \Rightarrow |y| \leq \sqrt{2(1+c_\rho)\log(8\rho^{-1})},$$

where $c_\rho \sim \frac{(1+d)\log(2\log\rho^{-1})}{2\log\rho^{-1}}$ as $\rho \to 0$.

Proof. When $|y| > \sqrt{2\log\rho^{-1}}$, we have $\zeta_d(y) \geq \frac{e^{y^2/2}}{8|y|^{1+d}}$, so that $\zeta_d(y) \leq \rho^{-1}$ implies $\frac{e^{y^2/2}}{8|y|^{1+d}} \leq \rho^{-1}$, or equivalently

$$y^2\left(1 - \frac{2(1+d)\log|y|}{y^2}\right) \leq 2\log(8\rho^{-1}).$$

Now since $\frac{\log|y|}{y^2}$ is decreasing in $|y|$ as soon as $|y|$ exceeds a universal constant, the above implies

$$y^2\left(1 - \frac{(1+d)\log(2\log\rho^{-1})}{2\log\rho^{-1}}\right) \leq 2\log(8\rho^{-1}).$$

Rearranging the above inequality gives the desired result. $\square$

4.4.4 Hellinger distance and hypothesis testing

The total variation distance between

$$\begin{aligned} P_0^{(n)} &:= \otimes_{i=1}^n N(0,1), \\ P_\rho^{(n)} &:= \otimes_{i=1}^n \left((1-\rho)N(0,1) + \rho\psi_d\right) \end{aligned}$$

satisfies the relationship

$$\inf_{T:\mathbb{R}^n\to\{0,1\}} \left\{P_0^{(n)}(T=1) + P_\rho^{(n)}(T=0)\right\} = 1 - \mathsf{TV}(P_0^{(n)}, P_\rho^{(n)}),$$

where the infimum is taken over all tests T which take as input the observations $Y_1, \ldots, Y_n$ and return a decision to either reject or accept the global null. By the NP lemma (Neyman and Pearson; 1933), the infimum is achieved by the likelihood ratio test,

$$\text{Reject if} \quad \frac{P_\rho^{(n)}}{P_0^{(n)}}(Y_1, \ldots, Y_n) > 1.$$

Failure of the likelihood ratio test to distinguish between $P_0^{(n)}$ and $P_\rho^{(n)}$ defines hardness in an information theoretic sense. The Hellinger distance between two distributions P, Q with densities p, q is defined

$$H^2(P, Q) \coloneqq \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2,$$

and the total variation distance can be expressed

$$\mathsf{TV}(P, Q) \coloneqq \frac{1}{2} \int |p - q|.$$

They have the following relationship.

Lemma 4.4.6. *For the Hellinger and total variation distance defined above,*

$$H^2(P, Q) \leq \mathit{TV}(P, Q) \leq \sqrt{2} H(P, Q).$$

It follows that the total variation distance tends to zero if and only if the squared Hellinger distance does as well. The total variation between two simple hypotheses characterizes the error of the likelihood ratio test between them, and the relationship between total variation and Hellinger distance means it is sufficient to analyze the Hellinger distance when determining whether or not this error goes to zero as the number of samples increases. Hellinger distance is convenient to analyze for independent samples due to a tensorization property, stated below.

Lemma 4.4.7. *If* $P = \otimes_{i=1}^n P_i$ *and* $Q = \otimes_{i=1}^n Q_i$*, then*

$$H^2(P, Q) = 1 - \prod_{i=1}^n (1 - H^2(P_i, Q_i)).$$

Proofs of these two lemmas can be found in the notes by Duchi (2016). Lemma 4.4.7 implies that if $P_0^{(n)}$ and $P_\rho^{(n)}$ are each product distribution as in

$$\begin{aligned} P_0^{(n)} &: Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \phi \\ P_\rho^{(n)} &: Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (1 - \rho)\phi + \rho\psi_d, \end{aligned}$$

and if $H^2(\phi, (1 - \rho)\phi + \psi_d) \to 0$ as $\rho \to 0$ and $n \to \infty$, then

$$H^2(P_0^{(n)}, P_\rho^{(n)}) \sim 1 - \exp\left(-nH^2(\phi, (1 - \rho)\phi + \psi_d)\right).$$

By Lemma 4.2.1, this implies that the Hellinger distance tends to zero (resp. 1) depending on the convergence to zero (resp. divergence) of the quantity

$$nH^2(\phi, (1 - \rho)\phi + \psi_d) = \frac{n\rho}{(\log \rho^{-1})^{d/2}}.$$

4.4.5 Type 1 error of the MLE $\hat{\rho}_n$

Recall that $\hat{\rho}_n$ is the maximum likelihood estimator of the sparsity rate based on independent samples $Y_1, \ldots, Y_n$ from the marginal approximation (4.7),

$$\hat{\rho}_n = \operatorname{argmax}_{0 \le \rho \le 1} \prod_{i=1}^{n} \left((1-\rho)\phi(Y_i) + \rho\psi_d(Y_i)\right).$$

By the calculation in the proof of Theorem 4.2.1, the type 1 error is equivalent to

$$\mathbb{P}_0(\hat{\rho}_n > 0) = \mathbb{P}_0\left(\sum_{i=1}^{n} \zeta_d(Y_i) > n\right) = \mathbb{P}_0\left(\frac{1}{n}\sum_{i=1}^{n} \zeta_d(Y_i) > 1\right),$$

where $\mathbb{P}_0$ denotes $Y_1, \ldots, Y_n \sim N(0,1)$ independently. By Lemma 4.4.3, the tail behavior of $\zeta_d(Y_i)$ when $Y_i \sim N(0,1)$ is

$$\mathbb{P}_0(\zeta_d(Y_i) > z) \asymp \frac{1}{z(\log z)^{d/2+1}} \text{ as } z \to \infty.$$

This fact is used to show that type 1 error of the MLE tends to zero as $n \to \infty$ at an inverse logarithmic rate, stated and proven below.

Theorem 4.4.1 (Type 1 consistency of the MLE)**.** *Let* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(0,1)$*, and let* $\hat{\rho}_n$ *denote the MLE* (4.8) *of* ρ*. Then*

$$\mathbb{P}(\hat{\rho}_n > 0) = O\left(\frac{1}{\log n}\right)$$

as $n \to \infty$.

Proof. Define $X_i := \zeta_d(Y_i)$ and $Z_i := X_i \cdot 1_{\{X \le L_n\}}$, where $L_n := \frac{n}{(\log n)^{d/2}}$. By the calculation in the proof of Theorem 4.2.1, the type 1 error is equivalent to

$$\mathbb{P}(\hat{\rho}_n > 0) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \zeta_d(Y_i) > 1\right) = \mathbb{P}(\bar{X} > 1),$$

where $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $\bar{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i$ and note that

$$\begin{aligned}\mathbb{P}(\bar{X} > 1) &= \mathbb{P}(\bar{X} > 1, \bar{Z} > 1) + \mathbb{P}(\bar{X} > 1, \bar{Z} \le 1) \\ &\le \mathbb{P}(\bar{Z} > 1) + \mathbb{P}(\bar{Z} \ne \bar{X}).\end{aligned}$$

By the union bound and Lemma 4.4.3,

$$\mathbb{P}(\bar{Z} \ne \bar{X}) \le \mathbb{P}\left(\max_{i \le n} X_i > L_n\right) \le n\mathbb{P}(\zeta_d(Y_1) > L_n) \lesssim \frac{(\log n)^{d/2}}{(\log n)^{d/2+1}} = \frac{1}{\log n}.$$

By Chebyshev,

$$\mathbb{P}(\bar{Z} > 1) = \mathbb{P}(\bar{Z} - \mathbb{E}\bar{Z} > 1 - \mathbb{E}\bar{Z}) \leq \frac{\mathrm{Var}(\bar{Z})}{(1 - \mathbb{E}\bar{Z})^2} \leq \frac{n^{-1}\mathbb{E}(Z_1)^2}{(1 - \mathbb{E}Z_1)^2}.$$

We claim the first and second moments satisfy

$$1 - \mathbb{E}(Z_1) \gtrsim \frac{1}{(\log n)^{d/2}}, \tag{4.28}$$

$$\mathbb{E}(Z_1^2) \lesssim \frac{n}{(\log n)^{d+1}}. \tag{4.29}$$

which would imply $\mathbb{P}(\bar{Z} > 1) \lesssim \frac{1}{\log n}$, completing the proof.

Proof of (4.28)**.** Let f denote the density of X_1, and notice $\mathbb{E}(Z_1) + \mathbb{E}(X_1 \cdot 1_{\{X_1 > L_n\}}) = \mathbb{E}X_1 = 1$. The first moment of the truncated variable Z_1 is thus

$$\begin{aligned}
\mathbb{E}(Z_1) &= 1 - \mathbb{E}X_1 \cdot 1_{\{X_1 > L_n\}} \\
&= 1 - \int_{L_n}^{\infty} x f(x) \mathrm{d}x \\
&= 1 - \left(\int_{L_n}^{\infty} (1 - F(x)) \mathrm{d}x + \Big[- x(1 - F(x)) \Big]_{L_n}^{\infty} \right).
\end{aligned}$$

Since $L_n \to \infty$, Lemma 4.4.3 implies there is some $c_1 > 0$ for which the above is bounded by

$$\begin{aligned}
&\leq 1 - \int_{L_n}^{\infty} \frac{c_1}{x(\log x)^{d/2+1}} \mathrm{d}x \\
&= 1 - \left[-\frac{2}{d} \frac{c_1}{(\log x)^{d/2}} \right]_{L_n}^{\infty} = 1 - \frac{2c_1}{d(\log L_n)^{d/2}} \sim 1 - \frac{2c_1}{d(\log n)^{d/2}}.
\end{aligned}$$

Proof of (4.29)**.** The second moment is

$$\begin{aligned}
\mathbb{E}(Z_1^2) &= \frac{1}{F(L_n)} \int_0^{L_n} x^2 f(x) \mathrm{d}x \\
&= \frac{1}{F(L_n)} \left[2 \int_0^{L_n} x(1 - F(x)) \mathrm{d}x + L_n^2 (1 - F(L_n)) \right] \\
&= \frac{1}{F(L_n)} \left[2 \int_0^{L_n^{1/2}} x(1 - F(x)) \mathrm{d}x + 2 \int_{L_n^{1/2}}^{L_n} x(1 - F(x)) \mathrm{d}x + L_n^2 (1 - F(L_n)) \right]
\end{aligned}$$

By Lemma 4.4.3, there is some constant $c_2 > 0$ for which the above is bounded by

$$\leq \frac{1}{F(L_n)}\left[2\int_0^{L_n^{1/2}} L_n^{1/2}(1-F(x))\mathrm{d}x + 2\int_{L_n^{1/2}}^{L_n} \frac{c_2 x}{x(\log x)^{d/2+1}}\mathrm{d}x + L_n^2 \cdot \frac{c_2}{L_n(\log L_n)^{d/2+1}}\right]$$

$$\lesssim L_n^{1/2}\int_0^{\infty}(1-F(x))\mathrm{d}x + \frac{L_n - L_n^{1/2}}{(\log L_n)^{d/2+1}} + \frac{L_n}{(\log L_n)^{d/2+1}} \asymp \frac{n}{(\log n)^{d+1}}.$$

□

Numerical evidence suggests the type 1 error decays inverse-logarithmically in the sample size, with exponent differing by power index d (Figure 4.4).

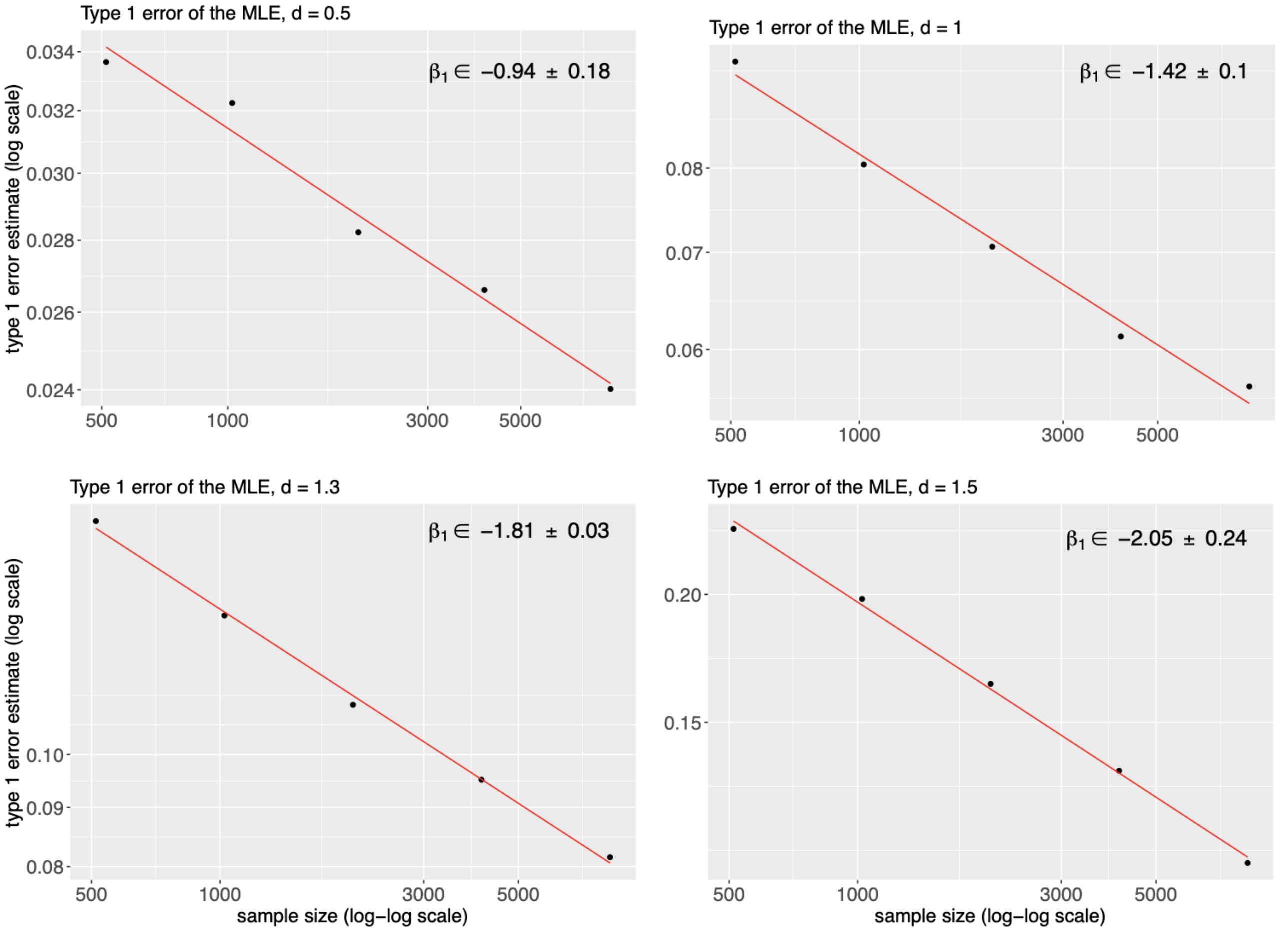


Figure 4.4: The type 1 error $\mathbb{P}_0(\hat{\rho}_n > 0)$ for sample sizes $n \in \{2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}\}$ and power index $d \in \{1/2, 1, 3/2\}$ are estimated using $N = 10^5$ many monte carlo replications. In each replication, we check whether $\hat{\rho}_n > 0$ based on iid samples $X_1, \ldots, X_n \sim N(0,1)$. The log of the error estimate of $\mathbb{P}_0(\hat{\rho}_n > 0)$ is regressed on $\log\log n$, and the fitted slope is an estimate of the exponent in the poly-logarithmically decaying error probability.

CHAPTER 5
FREQUENTIST LOCAL FALSE DISCOVERY RATES

5.1 Introduction

Suppose that we are testing a scientific hypothesis, and observe a t-statistic equal to 3.5. How confident can we be in rejecting the corresponding null hypothesis? One way to quantify the evidence is to calculate a p-value, say $p = 0.001$ if there are 50 degrees of freedom. A common mistake is to interpret 0.001 as the probability that the null is true in light of the data, but calculating this probability would require two other pieces of information: the prior probability that the null is true, and the distribution of the test statistic under the alternative. In many situations, different observers disagree about these quantities; moreover, to speculate about them is to stipulate that the truth value of a hypothesis is a random variable, which many scientists find unintuitive (Goodman (1999), Savage (1972)).

When analyzing many experiments at once, empirical Bayes or hierarchical Bayes methods are useful for estimating a prior distribution for effect sizes (see Efron (2019) for a recent overview). Examples from the psychology literature include aggregate-analyses of growth mindset interventions for academic achievement, and of "nudge" effects in behavioral psychology (Tipton et al.; 2022). Meta-analyses such as the ones performed by Macnamara and Burgoyne (2022) and Mertens et al. (2022a) assume exchangeability of effect sizes within a relevant sub-area of the psychology literature. From a frequentist standpoint, this implies a constant overall effect common to each experiment, despite variation among contexts in which effects are studied (Gelman; 2015). From a subjective Bayesian standpoint, the exchangeability assumption becomes less plausible as we look further into the details of different experiments.

Questions of relevance prevent us from accepting the results of a simple Bayesian analysis as an accurate reflection of our subjective posterior beliefs. Nevertheless, empirical Bayes methods are appealing from a pragmatic perspective. If we restrain ourselves from endlessly pulling these Bayesian threads, and instead perform a simple empirical Bayes analysis, what interpretation does the result have? In this chapter, we present a frequentist interpretation of what we have estimated, *without* assuming that the parameters are exchangeable, or even random.

Rather than attempt to calculate a subjective probability, we instead ask the question: of all the studies in this part of the literature with t-statistics close to 3.5, what fraction had true null hypotheses? This question makes sense to ask without Bayesian assumptions, and can be meaningfully answered in terms of relative frequencies within a large collection of hypotheses (see Von Mises (1964) for an entirely frequency-based theory of probability[1]).

[1] Von Mises' theory starts with the concept of a 'collective' (Von Mises; 1919), which formalizes the notion of repeated and unrelated trials. Although unsuccessful as the starting point for a mathematical theory of probability (Neyman; 1957), the concept gave rise to a lively discussion about the meaning of randomness and its role in scientific discovery; see e.g. Popper (2005) (Chapter 8, section 50), Church (1940), Wald (1939), Ville (1939), de Finetti (1938), Reichenbach (1949), Reichenbach (1971), Spielman (1976), Lindley (1966), Dawid (1985a), Dawid (1985b), Bienvenu et al. (2009) and references therein.

Moreover, if scientists in a given research field agree to report this number, say 10%, then approximately 10% of the ones who report this number would make false discoveries. An analogy can be made to *calibration* in forecasting (Dawid; 1982), which we discuss in Sections 5.2.3 and 5.6.

The local false discovery rate (lfdr) is equal to a posterior probability when effects are random, but is more generally interpreted as the expected proportion of nulls among hypotheses with similar test-statistics. This concept inspires a new frequentist type 1 error criterion for evaluating multiple testing procedures, called the boundary false discovery rate (bFDR), which is a local counterpart to the FDR. Simply put, the bFDR of a multiple testing procedure is the rate of false discoveries around its rejection threshold. Controlling bFDR at 20% means that we incur no loss by making one additional false discovery for every four additional true discoveries, whereas classical FDR control implies a different trade-off that depends on the unknown data generating process (Soloff et al.; 2022). To illustrate these ideas, we apply them to two real datasets: an aggregate of nudges in behavioral psychology and a genomic dataset previously analyzed by Efron et al. (2011). We apply these concepts to answer our original question, as presented in the context of nudge experiments aggregated by Mertens et al. (2022a). How confident can we be in a psychology researcher's claim about the effect of a nudge, given that its p-value is 0.001?

5.1.1 *Example: Mertens et al. (2022a) aggregate analysis of nudges*

The concept of "nudging" is described by Thaler and Sunstein (2009) as a way of gently influencing people towards making decisions for the benefit of themselves and of their communities without necessarily restricting their options. In an attempt to evaluate the effectiveness of psychological nudging on human behavior, Mertens et al. (2022a) collected data from 447 nudge experiments in the behavioral psychology literature. This aggregation included a wide range of nudges, such as making healthy options more visible on a breakfast menu, or adding extra text to a prompt to encourage organ donor registration. The goal of the meta-analysis by Mertens et al. (2022a) was to analyze these nudges along with hundreds of others to determine whether or not nudges are effective overall. The formulation of this question and the authors' conclusion was the subject of some debate, see e.g. Maier et al. (2022), Mertens et al. (2022b), Szaszi et al. (2022).

To understand the degree to which false discoveries are present in the aggregated dataset, we estimated the false discovery rate (FDR) using the Storey estimator (Storey; 2002) for the proportion of true nulls, restricting attention to just the $m = 261$ many p-values falling below the 5% two-sided significance level. This restriction is a way to work around the publication bias present in scientific journals; although ineffective nudges may be underrepresented among published studies in the aggregated dataset, the null hypotheses whose p-values fall within the significance region are less prone to censorship (Hung and Fithian (2020), Jaljuli et al. (2022)).

The Storey estimator of the null proportion within the significance region is around 28%, suggesting that roughly over a quarter of results reported below the 2.5% one-sided significance level are false discoveries. To mitigate the high rate of false claims, we ran the

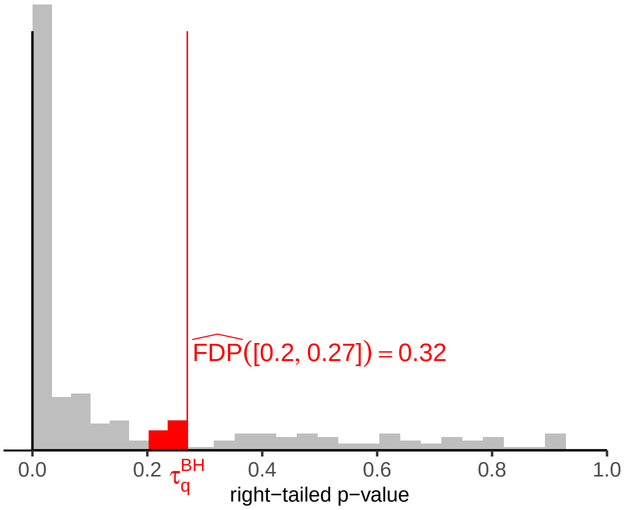


Figure 5.1: One-sided p-values from the nudge dataset falling below 0.025 are re-normalized by a factor of 40 (the reciprocal of the 2.5% one-sided significance threshold) and plotted in the above histogram. This pre-processing step returns the truncated observations to the original $(0, 1)$ scale and restores validity of the p-values. The BH threshold targeting an overall FDR of $q = 0.10$ is around $\tau_q^{\mathrm{BH}} = 0.27$, below which there are 202 rejections. The estimated FDP towards the edge of the BH set is around 32% within the interval $[0.2, 0.27]$.

Storey-adjusted BH procedure (Storey et al.; 2004) targeting a 10% FDR, yielding a more stringent rejection threshold, as shown in Figure 5.1, below which only 202 of the p-values fell. Upon inspecting the histogram left of the BH threshold, it is worth noting that the rate of false discoveries varies a lot across subsets of the rejection set. As Figure 5.1 shows, the estimated proportion of false discoveries (FDP) exceeds 10% for a subset of rejections near the rejection threshold; the method of estimating the FDP is illustrated more explicitly in Figure 5.2 using the plot of order statistics. The message here is that BH is overly liberal in its last few rejections, which are of low quality as judged on the basis of the order statistics.

Continuing the thought experiment further, we could imagine shrinking a subset of rejections down to a single one, and asking about whether a particular nudge was effective or not. At this point, if we consider a Bayesian model for the hypotheses,

$$H_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ nudge has the desired effect} \\ 0 & \text{if the } i^{\text{th}} \text{ nudge has no effect or backfires,} \end{cases}$$

we can get an approximate answer to this question by estimating the posterior probability that the hypothesis is null, conditional on the observation. In this model, the truth statuses

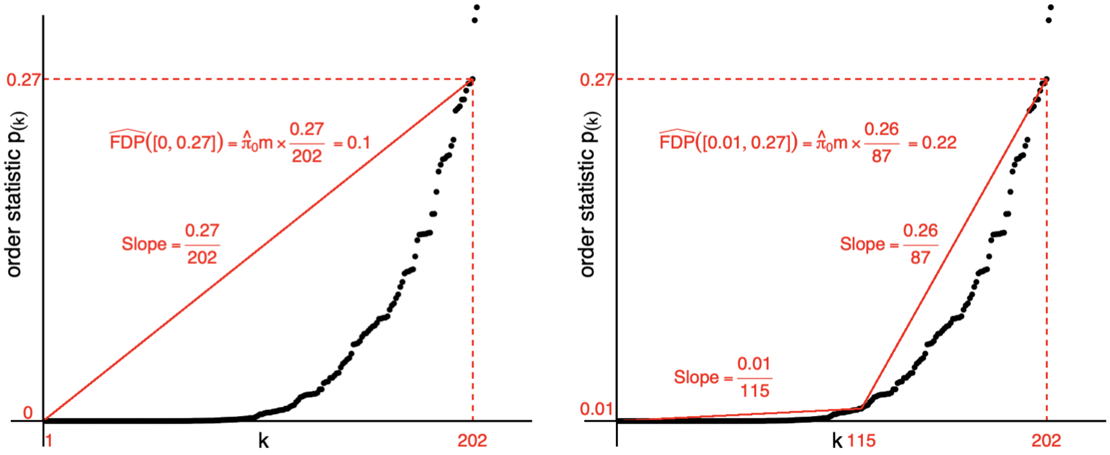


Figure 5.2: The order statistics are plotted against their rank, and the FDP over the entire set of rejections is estimated at 0.1 (left), whereas within the second half of the rejections the estimate is 0.22 (right). The FDP estimate over a subset of the rejection region is obtained via $\frac{V[a,b]}{R[a,b]} \approx \frac{m\hat{\pi}_0 \times (b-a)}{R[a,b]}$, where $[a,b] \subset [0, 0.27]$, $V[a,b]$ is the number of nulls in $[a,b]$, and $R[a,b]$ is the number of p-values in $[a,b]$. The estimate is proportional to the slope of the secant over the subset, and increases as the subset approaches the rejection threshold.

of hypotheses are Bernoulli random variables,

$$H_i \overset{\text{iid}}{\sim} \text{Bern}(1 - \pi_0), \tag{5.1}$$

$$p_i \mid H_i \sim \begin{cases} f_0, & \text{if } H_i = 0 \\ f_1 & \text{if } H_i = 1, \end{cases}$$

for $i = 1, \ldots, m$ where $\pi_0 \in [0,1]$ is the probability that a hypothesis is null, f_0 and f_1 are the null and alternative densities. The *local false discovery rate* (lfdr, Efron et al.; 2001), defined

$$\text{lfdr}(t) = \mathbb{P}(H_i = 0 \mid p_i = t) = \frac{\pi_0 f_0(t)}{f(t)}, \quad f := \pi_0 f_0 + (1 - \pi_0) f_1, \tag{5.2}$$

is a fundamental quantity in the Bayesian hypothesis testing paradigm, characterizing optimal rejection regions under various conditions (see, e.g. Sun and Cai (2007), Xie et al. (2011), and Heller and Rosset (2021)).

Since $\text{lfdr}(t)$ is a posterior probability, its meaning appears to hinge on the Bayesian assumption (5.1). The i.i.d. assumption implies that effects are identically distributed, but do we really believe, for example, that the breakfast and organ donation nudges are drawn from a common 'nudge distribution'? Regardless of whether we are willing to affirm exchangeability of nudge effects, we noticed a subset of low-quality BH rejections in the nudge data just by examining the plot of order statistics (Figure 5.2). From a frequentist perspective, evaluating the effectiveness of individual nudges remains a relevant question. If we were to view the truth status of each hypothesis as fixed and non-random, then lfdr is identically zero or one. As demonstrated with the Mertens et al. (2022a) dataset, local false discovery rates can still be estimated, but their precise meaning in the absence of assumption (5.1) remains to be made clear.

Organization of the chapter. In Section 5.2 we state a definition of the lfdr, offering epistemic and decision theoretic interpretations within the frequentist model. In Section 5.3, we define an error criterion for multiple testing called the boundary false discovery rate (bFDR) and describe a method that provably controls its bFDR in the frequentist model. In Section 5.4, we discuss connections between the lfdr at the rejection threshold and the bFDR of the procedure. In Section 5.5, we illustrate some of the ideas in this chapter on two real datasets. Section 5.6 concludes the chapter with a brief discussion, and Section 5.7 contains all proofs not provided in the exposition.

FDR	bFDR	lfdr	clfdr
false discovery rate	boundary FDR	local false discovery rate	compound local false discovery rate
$\mathbb{P}(H_{(I)} = 0)$	$\mathbb{P}(H_{(R)} = 0)$	$\bar{\pi}_0 \bar{f}_0(t)/\bar{f}(t)$	$\mathbb{P}(H_{(k)} = 0 \mid \{p_1, \ldots, p_m\})$

Table 5.1: The abbreviations FDR, bFDR, lfdr, and clfdr are displayed in the above table with their mathematical definitions in the bottom row. Here the compound lfdr is defined for the k^{th} smallest p-value $p_{(k)}$, and is equivalent (up to a random permutation) to the clfdr definition given in Section 5.4 (See Proposition 5.4.1).

5.2 A frequentist local false discovery rate

Let $H_1, \ldots, H_m \in \{0, 1\}$ be fixed, and suppose each p-value is independently distributed according to a probability measure on $[0, 1]$,

$$p_i \overset{\text{ind}}{\sim} P^{(i)} \quad \text{for } i = 1, \ldots, m. \tag{5.3}$$

A typical instance of this setup is for $H_i = 0$ to indicate that $P^{(i)}$ is within a certain class of probability measures, for instance the uniform distribution or the set of super-uniform distributions on $[0, 1]$. We impose no formal restrictions on the relationship between H_i and $P^{(i)}$, since our goal in this section is to provide an interpretation for the local fdr while being agnostic to the model specification.

The local false discovery rate (lfdr) at a point t in the sample space is defined as the probability, conditional on *some* p-value having been realized at t, that the hypothesis corresponding to *that* p-value is null,

$$\text{lfdr}(t) := \mathbb{P}(H_J = 0 \mid p_J = t \text{ for some } J \in [m]). \tag{5.4}$$

The hypothesis H_J corresponds to an observation at t, and $\mathrm{lfdr}(t)$ models our uncertainty about whether or not this hypothesis is null. If each $P^{(i)}$ has a Lebesgue continuous density $f^{(i)}$ on $[0, 1]$, then by independence, on the event that J exists, it is unique. Otherwise if multiple p-values are realized at t, then p_J may be understood as a uniform draw from the set of p-values tied at t. Here the randomness in the index J, representing *which* p-value is realized at t, substitutes for the random truth value of each H_j.

At first glance, one might be concerned about conditioning on the event that a continuous random variable is exactly equal to a point, since this is a zero probability occurrence. The conditional probability can be understood as a limit of probabilities,

$$\mathrm{lfdr}(t) = \lim_{\varepsilon \to 0} \mathbb{P}(H_J = 0 \mid p_J \in N_\varepsilon(t) \text{ for some } J \in [m]),$$

for a collection of neighborhoods $N_\varepsilon(t)$ shrinking to $\{t\}$ as $\varepsilon \to 0$. If each p_i has a continuous density on $[0, 1]$, then the manner in which the neighborhood $N_\varepsilon(t)$ shrinks to $\{t\}$ is arbitrary, always yielding the same limit. If there are multiple p-values in $N_\varepsilon(t)$, we interpret the selected p_J in (5.4) as a uniform draw from among them. In the $\varepsilon \to 0$ limit, the conditional probability is proportional to the ratio between the average null density and the overall average density of the p-values.

Theorem 5.2.1. *Fix $H_1, \ldots, H_m \in \{0, 1\}$ and suppose $p_1, \ldots, p_m$ are generated independently from (5.3). If each $P^{(i)}$ has a continuous density $f^{(i)}$ on $[0, 1]$, we have*

$$\mathrm{lfdr}(t) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}, \tag{5.5}$$

where $\bar{f}_0 = \frac{1}{m\bar{\pi}_0} \sum_{i:H_i=0} f^{(i)}$ and $\bar{f} = \frac{1}{m} \sum_{i=1}^m f^{(i)}$ are the null average and overall average densities, and $\bar{\pi}_0 = \frac{1}{m} \sum_{i=1}^m (1 - H_i)$ is the proportion of nulls. If $(H_i, p_i)_{i=1}^m$ are generated independently from the two-groups model (5.1), then

$$\mathrm{lfdr}(t) = \frac{\pi_0 f_0(t)}{f(t)},$$

where $f = \pi_0 f_0 + (1 - \pi_0) f_1$ is the marginal density of each observation.

Definition (5.4) recovers the Bayes lfdr within the two-groups model. It is non-trivial in the frequentist model because by conditioning on t being in the *set* of observations, and not on the observations themselves, there is still uncertainty that can be measured on the basis of the order statistics. For example, the hypotheses corresponding to the smallest few p-values are a random subset of hypotheses, even if the effects are fixed.

Sun and Cai (2007) showed formally that lfdr is the right quantity to look at for deciding whether or not to reject a hypothesis in the i.i.d. Bayes model. This conclusion continues to hold in the frequentist model, where optimality is defined with respect to a weighted combination of type 1 and 2 errors, a claim that we formalize in the next section.

5.2.1 A compound decisions perspective

Consider data $p = (p_1, \ldots, p_m)$ drawn independently from model (5.3) with fixed effects $H = (H_1, \ldots, H_m) \in \{0,1\}^m$, where each $P^{(i)}$ has a density $f^{(i)}$. In this setting, a multiple testing procedure can be represented by a set of decisions $\delta(p) = (\delta_1(p), \ldots, \delta_m(p)) \in \{0,1\}^m$ with compound loss,

$$L(H, \delta) = \frac{1}{m} \sum_{i=1}^{m} \ell(H_i, \delta_i), \tag{5.6}$$

for some non-negative function $\ell : \{0,1\}^2 \to \mathbb{R}_+$. Robbins (1951) was the first to formulate the compound decision problem and observe that if the decision rule is separable, i.e. $\delta_i(p) = \mathfrak{d}(p_i)$ for some function $\mathfrak{d} : [0,1] \to \{0,1\}$, then the compound risk could be understood as the Bayes risk for a single problem instance drawn uniformly at random. In the current setting, this insight is instantiated by letting $(H_I, p_I) \in \{0,1\} \times \mathbb{R}$, where $I \sim \text{Uniform}\{1, \ldots, m\}$, and rewriting the compound risk as an expectation with respect to the index I,

$$\mathbb{E}L(H, \delta) = \mathbb{E}\ell(H_I, \mathfrak{d}(p_I)). \tag{5.7}$$

Equation (5.7) is a form of "the fundamental theorem of compound decisions" (Zhang; 2003) and connects the compound decision problem to Bayes optimality theory, the key observation being that (H_I, p_I) is drawn from a two-groups model:

$$\begin{aligned} H_I &\sim \text{Bern}(1 - \bar{\pi}_0), \\ p_I \mid H_I = h &\sim \begin{cases} \bar{f}_0 & \text{if } h = 0 \\ \bar{f}_1 & \text{if } h = 1 \end{cases} \end{aligned} \tag{5.8}$$

where $\bar{f}_0 := \frac{1}{m_0} \sum_{i:H_i=0} f^{(i)}$ and $\bar{f}_1 := \frac{1}{m_1} \sum_{i:H_i=1} f^{(i)}$ denote the average null and alternative densities. Although the fixed effects model (5.3) allows for m potentially different distributions, applying equation (5.7) brings us back to a two-groups model, where many different densities have been aggregated into $\bar{f}_0$ and $\bar{f}_1$. This reduction is the natural next step to take in the development of compound decision theory for hypothesis testing (Yekutieli and Weinstein; 2019). We call (5.8) the *oracle two-groups model* to emphasize that it depends on unknown parameters such as $\bar{f}_1$ and m_0, and that it holds in the frequentist setting where the truth status of each hypothesis is fixed.

This reformulation of the compound risk shows that minimizing the average case loss (5.6) is equivalent to achieving the Bayes risk in the oracle two-groups model. Taking the component-wise loss to be $\ell_\lambda(H_i, \delta_i) := 1_{\{H_i=1,\delta_i=0\}} + \lambda \cdot 1_{\{H_i=0,\delta_i=1\}}$, we obtain the weighted misclassification risk,

$$\mathbb{E}L_\lambda(H, \delta) = \mathbb{E}\ell_\lambda(H_I, \mathfrak{d}(p_I)) = \mathbb{P}(H_I = 1, \mathfrak{d}(p_I) = 0) + \lambda \cdot \mathbb{P}(H_I = 0, \mathfrak{d}(p_I) = 1),$$

for a parameter $\lambda > 0$ specifying the relative cost in making a type I error compared to a type II error. The smallest risk among all separable rules is attained by the Bayes rule, which is characterized by the local false discovery rate within this univariate model,

$$\mathbb{P}(H_I = 0 \mid p_I = t) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}, \tag{5.9}$$

where $\bar{f} \coloneqq \bar{\pi}_0 \bar{f}_0 + (1 - \bar{\pi}_0)\bar{f}_1$ is the average density of the observations. Yekutieli and Weinstein (2019) applied this theory to the present context, noting that the procedure

$$\mathfrak{d}^*(p_i) = \begin{cases} 1 & \text{if } \operatorname{lfdr}(p_i) \leq \frac{1}{1+\lambda} \\ 0 & \text{otherwise,} \end{cases} \tag{5.10}$$

optimizes a trade-off between marginal false discovery/non-discovery rates that is equivalent to minimizing the weighted misclassification risk.

In light of the decision rule (5.10), it is tempting to take the value of lfdr at an observed test-statistic p_i as a measure of our confidence in the i^{th} hypothesis being null. However, as $H_i \in \{0, 1\}$ is non-random, $\operatorname{lfdr}(p_i)$ is no longer interpretable as a posterior probability,

$$\operatorname{lfdr}(p_i) \neq \mathbb{P}(H_i = 0 \mid p_i) \in \{0, 1\}.$$

To develop a more intuitive justification of viewing $\operatorname{lfdr}(p_i)$ as a measure of confidence about whether or not a hypothesis is null, we first look more closely at the meaning of $\operatorname{lfdr}(t)$ for a fixed input t: in a sequence of independent realizations of the vector $(p_1, \ldots, p_m)$, the lfdr at t is the long-run proportion of realizations where some entry was realized near t, that the corresponding hypothesis was null.

The frequentist 'parallel-universe' interpretation of lfdr is admittedly a stretch of imagination, relying on conceptual replicates of the entire multiple testing experiment which are structurally identical yet in some sense, independent; further, the utility of this interpretation in making inferences about an underlying probability structure is not clear if we only observe a single multiple testing experiment (Dawid; 1985b). In the next section, we identify conditions under which the value $\operatorname{lfdr}(t)$ may be more simply understood as the local frequency of nulls near t within a single multiple testing experiment,

$$\operatorname{lfdr}(t) \approx \frac{\#\{i : H_i = 0, p_i \approx t\}}{\#\{i : p_i \approx t\}}.$$

5.2.2 Local frequency of nulls

Although the reduction to an oracle two-groups model is useful for seeing why the lfdr is a fundamental quantity for testing many fixed effects, it is not necessary to reference a Bayes two-groups model at all in order to make sense of the lfdr. It can be understood on its own as the marginal FDR in an interval shrinking to a point. Subset versions of the FDP and mFDR are defined

$$\mathrm{FDP}(A) = \frac{V(A)}{1 \vee R(A)}, \quad \mathrm{mFDR}(A) = \frac{\mathbb{E}V(A)}{\mathbb{E}R(A)}$$

where $V(A) \coloneqq \#\{i : H_i = 0, p_i \in A\}$ and $R(A) \coloneqq \#\{i : p_i \in A\}$ for a subset $A \subset [0, 1]$, and the interval pFDR is defined

$$\mathrm{pFDR}([s, t]) = \mathbb{E}\left(\mathrm{FDP}([s, t]) \mid p_i \in [s, t] \text{ for some } i\right).$$

Under various conditions, the interval versions of pFDR and mFDR tend to lfdr as the interval shrinks to a point.

Theorem 5.2.2. *Suppose $p_i \sim P^{(i)}$ for $i = 1, \ldots, m$. If each $P^{(i)}$ has a continuous density $f^{(i)}$, then for each $t \in (0, 1)$,*

$$\lim_{\varepsilon \to 0} \mathrm{mFDR}([t - \varepsilon, t]) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}.$$

If we further assume that $(p_i)_{i=1}^m$ are independent, then for each $t \in (0, 1)$

$$\lim_{\varepsilon \to 0} \mathrm{pFDR}([t - \varepsilon, t]) = \frac{\bar{\pi}_0 \bar{f}_0(t)}{\bar{f}(t)}. \tag{5.11}$$

In the absence of the continuous density assumption, i.e. if some $P^{(i)}$ has an atom at t, then

$$\lim_{\varepsilon \to 0} \mathrm{mFDR}([t - \varepsilon, t]) = \frac{\bar{\pi}_0 \bar{P}_0(\{t\})}{\bar{P}(\{t\})},$$

where $\bar{P}_0 \coloneqq \frac{1}{m_0} \sum_{i:H_i=0} P^{(i)}$ and $\bar{P} \coloneqq \frac{1}{m} \sum_{i=1}^m P^{(i)}$ are the average null and overall average probability measures.

Equation (5.11) suggests the following interpretation of the formula for the lfdr: it is roughly equal to the proportion of null hypotheses whose p-values fell near t. Other conceptions of the lfdr such as (5.9) and (5.4) require us to envision independent replications of the entire multiple testing experiment, conditioning on those in which a selected test-statistic is realized near a point. The interpretation in the current section demands less from our imagination, making reference to a single collection of hypotheses and their test statistics.

As the compound decision theory demonstrates, lfdr converts p-values to the scale for trading off type 1 and 2 errors, leading to procedures of the form:

$$\text{Reject } H_i = 0 \text{ if } \mathrm{lfdr}(p_i) \le \alpha, \quad \alpha \in (0, 1). \tag{5.12}$$

In large samples, the least promising rejection has $\mathrm{lfdr}(p_i) \approx \alpha$. What does it mean to say there is a $\mathrm{lfdr}(p_i) \approx \alpha$ chance that $H_i = 0$ when effects are fixed? Equation (5.11) illuminates another function of lfdr, which is to provide calibrated forecasts[2].

To illustrate the idea, suppose the local weatherman predicts a 35% chance of rain tomorrow. How confident can we be that it will rain? If we attempt to decide for ourselves what the "true" probability is, we will quickly find ourselves in a thicket of questions about what we should condition on: our own state of knowledge, the position and momentum of every particle in the atmosphere, and so on. A simpler, and less assumption-laden answer to this question is to evaluate the weatherman's track record: in the past, when he has predicted a 35% chance of rain, how frequently has it actually rained? In our case, we may reframe the question as: among the instances in which $\text{lfdr}(p_i) \approx \alpha$, how often was the corresponding hypothesis null?

5.2.3 lfdr is calibrated

Given access to the lfdr function, we could process the p-values by applying lfdr to each observation, resulting in $\text{lfdr}(p_i)$ for $i = 1, \dots, m$. In large samples, this procedure has the following property: approximately an α proportion of the p-values with $\text{lfdr}(p_i) \approx \alpha$ are true nulls,

$$\alpha \approx \frac{\#\{i : H_i = 0, \text{lfdr}(p_i) \approx \alpha\}}{\#\{i : \text{lfdr}(p_i) \approx \alpha\}}. \tag{5.13}$$

Our next result formalizes this intuition, establishing that the difference between the left and right hand sides of expression (5.13) tends to zero in probability under mild regularity conditions.

Proposition 5.2.1. *Suppose $p_i \sim f^{(i)}$ are independently generated for $i = 1, \dots, m$ where $H_i = 0 \Rightarrow f^{(i)}$ is the Uniform$(0,1)$ distribution, $\bar{f}$ is decreasing and differentiable, and $\varepsilon_m \to 0$ is a sequence for which $m_0 \varepsilon_m \to \infty$. If $(\bar{f}^{-1})'(\bar{\pi}_0/\alpha) > 0$ is bounded away from zero, then*

$$\text{FDP}\Big(\{i \leq m : |\text{lfdr}(p_i) - \alpha| \leq \varepsilon_m\}\Big) \xrightarrow{\mathbb{P}} \alpha \textit{ as } m \to \infty,$$

for any $\alpha \in$ range(lfdr).

It is straightforward to show that an analagous result holds in the Bayes model, and a short proof of strong consistency in the fixed-effects model can be found in Appendix 5.7, which may be of interest in online settings where one is concerned about the possibility of deviations occurring infinitely often within a single sequence of realizations $(H_i, p_i)_{i=1}^{\infty}$. Proposition 5.2.1 is related to a general calibration theorem in Dawid (1982), who studied calibration of probability forecasts based on previous outcomes until the current point in time. In our setting, the outcomes $H_1, H_2, \dots \in \{0, 1\}$ are non-random and unobserved.

[2] see e.g. Dawid (1982), Dawid (1985a), Dawid and Vovk (1999), Gneiting et al. (2007), Gupta et al. (2020) and references therein for various notions of calibration.

To accommodate this setup, we say that *a function* $g : [0,1] \to [0,1]$ *provides calibrated null probabilities if*

$$\mathbb{P}(H_J = 0 \mid g(p_J) = \alpha \text{ for some } J \in [m]) = \alpha, \tag{5.14}$$

for any $\alpha \in \text{range}(g)$, where, if there are ties among values of $g(p_i)$ at α, then $g(p_J)$ is defined as a uniform draw from among them. The following result connects definition (5.4) of the lfdr to the concept of calibration in forecasting.

Proposition 5.2.2. *Suppose* $p_i \sim f^{(i)}$ *are independent for* $i = 1, \ldots, m$, *and each* $f^{(i)}$ *is a continuous density. Then* lfdr : $[0,1] \to [0,1]$ *provides calibrated null probabilities, i.e.*

$$\mathbb{P}(H_J = 0 \mid \text{lfdr}(p_J) = \alpha \textit{ for some } J \in [m]) = \alpha,$$

for any $\alpha \in \text{range}(\text{lfdr})$.

Using lfdr is not the only way to calibrate; for example, forecasting with the constant function $g \equiv \bar{\pi}_0$ is also valid in the sense of (5.14). However, lfdr-based methods are sharp, in the sense that any function providing calibrated null probabilities is equivalent to the lfdr but conditioned on coarser information. This establishes lfdr as the finest calibration function, a result that follows directly from Proposition 1 of Gupta et al. (2020).

Proposition 5.2.3 (Corollary of Proposition 1 in Gupta et al. (2020))**.** *Suppose* $p_1, \ldots, p_m$ *are generated independently from the model* (5.3) *where each* $f^{(i)}$ *is a continuous density. If* $g : [0,1] \to [0,1]$ *is calibrated in the sense of* (5.14), *then*

$$g(t) = \mathbb{E}(\text{lfdr}(p_I) \mid p_I \in A_t) \quad I \sim \text{Uniform}\{1, \ldots, m\}$$

for some subset $A_t \subset [0,1]$.

In finite samples, we cannot estimate lfdr perfectly. For p-values, it is commonly assumed in the multiple testing literature that a smaller value of p_i indicates more evidence against $H_i = 0$ (see, e.g. Genovese and Wasserman (2004) and Strimmer (2008)), which is encoded in the current setting by the following monotonicity assumption.

$$H_i = 1 \text{ implies that } f^{(i)} \text{ is decreasing.} \tag{MA}$$

This assumption reduces the downstream multiple testing problem to choosing a threshold $\hat{\tau}$ based on $p_1, \ldots, p_m$, and rejecting $H_i = 0$ when $p_i \leq \hat{\tau}$. In the next section, we discuss the type 1 error of a rejection threshold $\hat{\tau}$ in terms of the rate of false discoveries at the threshold, which we call the *boundary false discovery rate.*

5.3 Multiple testing: FDR at the boundary

To evaluate multiple testing procedures, it is natural to ask whether all the rejections are individually defensible, not just whether the list of all rejections is defensible as a whole.

In a Bayesian model, this question can naturally be formulated in terms of the maximum *a posteriori* null probability over all the rejections. Soloff et al. (2022) define the max-lfdr for a multiple testing procedure as the expectation of this maximum, thereby evaluating a procedure $\mathcal{R} = \{i : \text{reject } H_i = 0\}$ according to its *least promising* rejection,

$$\text{max-lfdr}(\mathcal{R}) = \mathbb{E}\left[\max_{i \in \mathcal{R}} \mathbb{P}(H_i = 0 \mid p_i)\right].$$

In a frequentist analysis under the fixed effects model (5.3), however, it is less obvious how to formalize what we mean by the "least promising rejection." In particular, because the null probability for each hypothesis is either one or zero, the maximum is always one whenever we make any false rejections at all.

Instead, consider the null event for the hypothesis associated with the largest p-value within the rejection region, and denote by $H_{(k)} \in \{0, 1\}$ the hypothesis corresponding to $p_{(k)}$, the k^{th} smallest p-value. For a procedure $\mathcal{R}$ whose rejection region $[0, \hat{\tau}]$ contains the R smallest p-values, the *boundary false discovery rate* (bFDR) is defined as the probability that the last rejection is a false discovery,

$$\text{bFDR}(\mathcal{R}) := \mathbb{P}(H_{(R)} = 0), \tag{5.15}$$

where $H_{(0)} := 1$ indicates the event where no rejections are made. In the case of discrete p-value distributions (Section 5.3.3), ties among p-values at the boundary point $\{\hat{\tau}\}$ are understood to be broken uniformly at random. The boundary FDR recovers the max-lfdr within a Bayes two-groups model when the alternative density is decreasing, and avoids a trivial reduction to the FWER when the hypotheses are fixed.

Definition (5.15) is at first glance puzzling; if the hypotheses $H_1, \ldots, H_m \in \{0, 1\}$ are fixed to begin with, then how can we speak of the probability that one of them is null? The reason is that we are not asking about a fixed hypothesis, but a random one depending on the data. $H_{(R)}$ is the hypothesis corresponding to $p_{(R)}$, which depends on the particular realization of order statistics.

5.3.1 Comparison with FDR

By comparison, the usual FDR measures the null probability of a uniformly selected rejection:

$$\text{FDR}(\mathcal{R}) = \mathbb{P}(H_{(I)} = 0), \quad I \sim \text{Uniform}\{1, \ldots, R\}.$$

Under Assumption (MA), the boundary rejection has the greatest null probability of any rejection. As a result, the boundary FDR is larger than the FDR. While one might therefore be tempted to conclude that bFDR control is an inherently more conservative goal than FDR control, in practice this may or may not be the case, because we should use a larger numerical threshold when controlling the bFDR than when controlling the FDR. For example, an analyst who equates $\lambda = 4$ type II errors with a single type I error would want to control bFDR at level $1/(1 + \lambda) = 0.2$. The same analyst would *not* be satisfied with a method

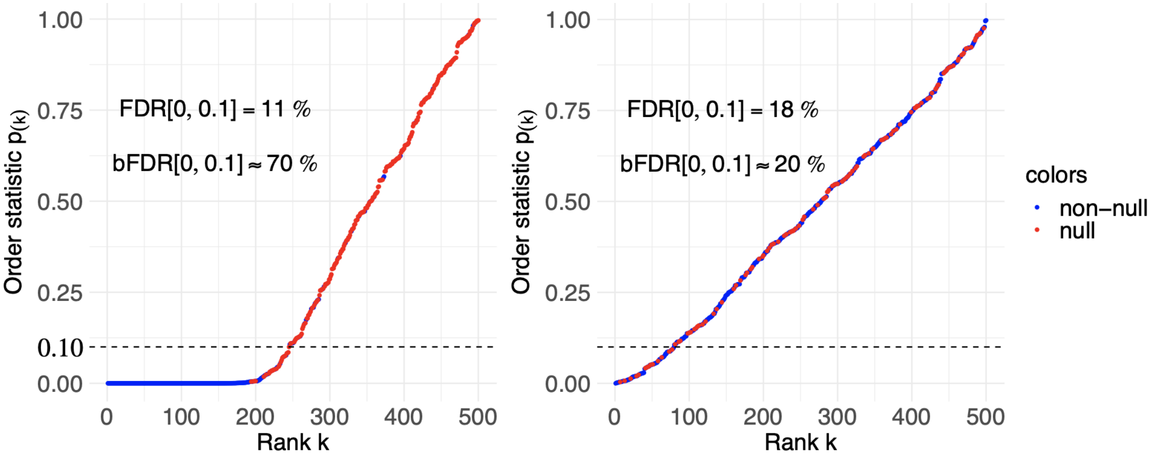


Figure 5.3: Left: order statistics plot for p-values generated with $m = 500, \bar{\pi}_0 = 0.5$ where null p-values (red) are i.i.d. Uniform$(0, 1)$ and alternative p-values (blue) are i.i.d. Beta$(0.05, 1)$. Right: order statistics plot for p-values generated with $m = 500, \bar{\pi}_0 = 0.25$ where alternative p-values are i.i.d. Beta$(0.8, 1)$.

whose FDR is 0.2, since the cost of the false discoveries would on average exactly cancel out the benefits of the true discoveries, yielding no net benefit relative to a trivial procedure that never rejects anything.[3] Instead, such an analyst would always aim to control FDR at some level smaller than 0.2, for example 0.1 so that they achieve some net benefit from the experiment. As a result, no sensible analyst would ever be interested in bFDR control and FDR control at the same level. Since bFDR control and FDR control would never be carried out at the same level, it is unclear which is more conservative in any given case.

A synthetic example is shown in Figure 5.3, where there is a bigger difference between the FDR and the boundary FDR when the non-null signal distribution is more concentrated near the origin. In the right panel of this example, where the null and non-null p-values are similarly distributed, the bFDR and FDR are also quite similar, as the rate of nulls is more or less constant throughout the interval $[0, 0.1]$. The bFDR and FDR estimates for the interval $[0, t]$ are plotted for t between 0 and 1 in Figure 5.4 for the same experiment with a larger sample size.

Figure 5.5 displays the plot of order statistics from the nudge data, where the FDR is estimated at 10%, while the estimated boundary FDR is above 50%. In other words, the last BH rejection is estimated to be more likely null than non-null. An analyst who equates 1 false positive with 9 missed discoveries would not find that last rejection worthwhile to follow up on. Instead, the analyst could control the rate of false discoveries among *additional* rejections,

[3] To illustrate this point, consider the weighted classification risk from Section 5.2.1, which can be redefined (up to additive and multiplicative constants) as

$$L_\lambda(H, \delta) \coloneqq \lambda V - (R - V),$$

where V is the number of false positives among the R discoveries. A procedure targeting a false discovery rate $V/R = 1/(1 + \lambda)$ achieves the same loss as a trivial procedure that simply sets $V = R = 0$.

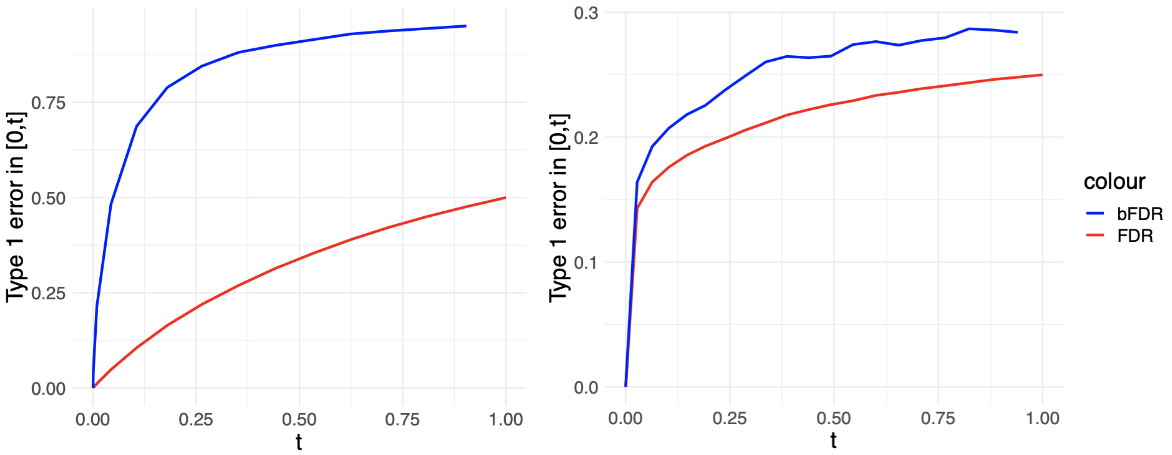


Figure 5.4: The FDR and bFDR of the interval $[0, t]$ are estimated for $t \in \{1/20, 2/20, \dots, 19/20\}$. Left: FDR and bFDR curves for p-values generated with $m = 10^5, \bar{\pi}_0 = 0.5$ where null p-values are i.i.d. $\text{Uniform}(0, 1)$ and alternative p-values are i.i.d. $\text{Beta}(0.05, 1)$. Right: FDR and bFDR curves for p-values generated with $m = 10^5, \bar{\pi}_0 = 0.25$ where alternative p-values are i.i.d. $\text{Beta}(0.8, 1)$.

using the observed data to estimate the null probability of these potential discoveries. In the next section, we describe a method that provably controls its bFDR when the p-values are independent and uniform under the null.

5.3.2 Controlling the boundary FDR

Within the i.i.d. Bayes model, Soloff et al. (2022) proposed the Support Line (SL) method for controlling the max-lfdr under a monotonicity constraint. The procedure run at level α rejects the R_α smallest p-values, where

$$R_\alpha \coloneqq \underset{k=0,\dots,m}{\operatorname{argmax}} \left\{ \frac{\alpha k}{m} - p_{(k)} \right\}, \quad p_{(0)} \coloneqq 0. \tag{5.16}$$

In the frequentist model, SL satisfies a bound on its boundary FDR under mild conditions.

Theorem 5.3.1 (Corollary of Lemma 2 in Soloff et al. (2022))**.** *If* $p_1, \dots, p_m$ *are independent and* $H_i = 0$ *implies* $p_i \sim \text{Uniform}(0, 1)$*, then*

$$\text{bFDR}(\mathcal{R}_\alpha) = \bar{\pi}_0 \alpha,$$

where $\mathcal{R}_\alpha \coloneqq \{i : p_i \leq p_{(R_\alpha)}\}$ *is defined by (5.16).*

Proof of Theorem 5.3.1. The event $\{H_{(R_\alpha)} = 0\}$ can be written as a disjoint union,

$$\{H_{(R_\alpha)} = 0\} = \bigcup_{i:H_i=0} \{p_{(R_\alpha)} = p_i\} \Rightarrow \mathbb{P}(H_{(R_\alpha)} = 0) = \sum_{i:H_i=0} \mathbb{P}(p_{(R_\alpha)} = p_i) = m_0 \cdot \frac{\alpha}{m}.$$

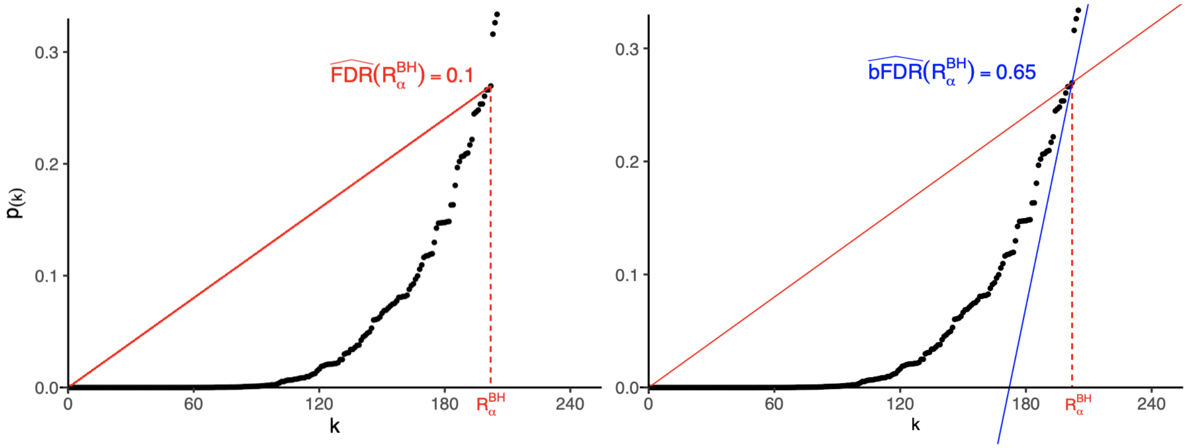


Figure 5.5: The left plot illustrates the Storey-adjusted BH procedure targeting FDR at level 10% on the one sided p-values from the Mertens et al. (2022a) dataset. The right plot shows the estimate of the boundary FDR of the Storey-BH rejection set, which is proportional to the slope of a line supporting the plot at the largest p-value within the rejection region.

The last equality follows from Lemma 2 of Soloff et al. (2022), which states that for any configuration of $p_1, \ldots, p_{m-1}$, the probability that a null p-value p_m achieves the optimum in (5.16) is equal to $\frac{\alpha}{m}$. □

Remark 5.3.1. *An alternative proof of the fact "$H_i = 0 \Rightarrow \mathbb{P}(p_{(R_\alpha)} = p_i) = \frac{\alpha}{m}$" can be found in Appendix 5.7, the technical key for which is a telescoping sum argument. The claim that the BH procedure controls its FDR can also be understood from this perspective and a proof is provided in Appendix 5.7.*

Under Assumption (MA), the rejection most likely to be null is the last one $H_{(R_\alpha)}$, and thus the null probability for each element the rejection set $\mathcal{R}_\alpha$ is controlled, in the sense that for any $K \leq R_\alpha$ which is a function of P_m,

$$\begin{aligned}\mathbb{P}(H_{(K)} = 0) &= \sum_{i:H_i=0} \mathbb{P}(\mathrm{rank}(p_i) = K) \\ &\leq \sum_{i:H_i=0} \mathbb{P}(\mathrm{rank}(p_i) = R_\alpha) = \mathbb{P}(H_{(R_\alpha)} = 0) = \bar{\pi}_0\alpha,\end{aligned}$$

where the inequality follows from Lemma 5.7.3 and the assumption that nulls are independent and uniformly distributed. The monotonicity assumption holds for example when the p-values are computed for one or two-sided tests in a Gaussian sequence model.

Remark 5.3.2. *In the Bayesian two-groups model, the max-lfdr coincides with the boundary FDR when the alternative density is decreasing. In the frequentist model, the analogous assumption is* (MA)*, which ensures that the boundary FDR coincides the with null probability for the least promising rejection. Without this assumption, it is not necessarily the case that* *the least promising rejection occurs at the boundary, as it may occur within the interior of the rejection set.*

For example, if the null p-values are uniformly distributed and the alternative p-values are distributed according to a common density with exactly one mode at $\tau/2$ for some $\tau \in (0,1)$, then the bFDR *of the procedure: "reject $H_i = 0$ when $p_i \leq \tau$", corresponds to the rejection made closest to $\tau/2 \in (0,\tau)$, which may not be the largest p-value in the rejection region $[0,\tau]$.*

If it was known that the alternative density is not monotone, then it would not make sense to only consider procedures of the form "reject $H_i = 0$ when $p_i \leq \tau$" for thresholds $\tau \in [0,1]$. In this work however, we only consider such procedures, as (MA) *is a natural shape constraint for the probability density function of a p-value corresponding to a false null hypothesis.*

Generalizing beyond the uniform null assumption, a sufficient condition for conservative boundary FDR control in Theorem 5.3.1 is for each null density to be bounded,

$$H_i = 0 \Rightarrow f^{(i)}(t) \leq 1 \quad \text{for all } t \in [0,\alpha]. \tag{5.17}$$

This condition is distinct from requiring the nulls be super-uniformly distributed, an assumption commonly made in the multiple testing literature which is insufficient for SL to control its boundary FDR[4]. However, a typical example in which super-uniformly distributed p-values arise is one-sided Gaussian location testing,

$$X_i \sim N(\theta_i, 1), \quad i = 1, \ldots, m$$

where $H_i = 0 \Rightarrow \theta_i \leq 0$. In this case, the probability density function for $p_i = 1 - \Phi(X_i)$ satisfies (5.17) under the null, for any $\alpha \leq 1/2$. This observation extends to one-parameter exponential families with continuous densities.

Proposition 5.3.1. *Let $(g_\theta)_{\theta\in\mathbb{R}}$ denote an exponential family of continuous distributions on $\mathbb{R}$ with densities*

$$g_\theta(z) = \exp(\theta z - A(\theta))g_0(z), \quad \theta, z \in \mathbb{R},$$

with corresponding cdfs (G_θ). For one-sided testing of the hypotheses $H_i : \theta_i \leq \theta_0$, let $\alpha^* = 1 - G_{\theta_0}(\mathbb{E}_{\theta_0} Z)$ be the upper quantile of the mean under θ_0. Then the null density of the one-sided p-value $p = 1 - G_{\theta_0}(Z)$ is bounded by 1 on $[0,\alpha^*]$, for all $\theta \leq \theta_0$.

5.3.3 Discrete-uniform null distribution

Other common settings in which super-uniformly distributed p-values arise are in permutation testing and variable selection (see, e.g. Barber and Candès (2015), Edgington and

[4] Failure of bFDR control for the SL procedure under the super-uniform null assumption is shown to be possible in Appendix 5.7.

Onghena (2007), Nair and Janson (2023) and references therein), where null p-values are distributed uniformly over a grid $\left\{\frac{1}{L}, \frac{2}{L} \ldots, \frac{L-1}{L}, 1\right\}$ for some positive integer L. The p-values can be converted to continuous random variables by taking

$$\tilde{p}_i \sim \text{Uniform}\left(\frac{\ell-1}{L}, \frac{\ell}{L}\right), \quad i = 1, \ldots, m$$

when $p_i = \frac{\ell}{L}$ for some $\ell \in \{1, \ldots, L\}$. Running SL on the randomized p-values $(\tilde{p}_i)$ yields exact bFDR control, but the rejection set also depends on the particular realization of the transformation $p_i \mapsto \tilde{p}_i$. The results that follow in this section are asymptotic, and regard the boundary FDR of the SL procedure run directly on the discrete p-values. If either $L \to \infty$ with m fixed, or $m \to \infty$ with L fixed, then the boundary FDR is controlled asymptotically (Theorems 5.3.2 and 5.3.3). Things are less clear when m/L converges to a constant. In this case, numerical evidence (see Appendix 5.7, Figure 5.12) suggests the bFDR guarantee may be violated even as m, L get large, with $L \approx 2m$. Lemma 5.3.1 sheds some light on the interplay between m and L and is important for proving our next result.

Theorem 5.3.2. *Let $p_1, \ldots, p_m$ be independently drawn from the fixed effects model* (5.3) *where $H_i = 0$ implies $p_i \sim$* Uniform $\left\{\frac{1}{L}, \frac{2}{L} \ldots, \frac{L-1}{L}, 1\right\}$*, and each $P^{(i)}$ has the same support. For any fixed $\alpha \in [0, 1]$ and m, we have*

$$\text{bFDR}(\mathcal{R}_\alpha) \leq \bar{\pi}_0 \alpha + O\left(\frac{m^2}{L}\right),$$

as $L \to \infty$.

The proof of Theorem 5.3.2 is similar to that of Theorem 5.3.1, both of which start by splitting the boundary FDR into a symmetric contribution from each null,

$$\text{bFDR}(\mathcal{R}_\alpha) = m_0 \mathbb{P}(\text{rank}(p_m) = R),$$

breaking ties uniformly at random, and assuming without loss of generality that $H_m = 0$. As $L \to \infty$, the probability that p_m is the realized boundary rejection is asymptotically no larger than α/m.

Lemma 5.3.1. *Let $p_1, \ldots, p_{m-1} \in [0, 1]$ be deterministic (non-random) variables, and suppose that $p_m \sim$* Uniform $\left\{\frac{1}{L}, \frac{2}{L}, \ldots, \frac{L-1}{L}, 1\right\}$*. Then as $L \to \infty$,*

$$\mathbb{P}(\text{rank}(p_m) = R) \leq \frac{\alpha}{m} + O\left(\frac{m}{L}\right),$$

as $L \to \infty$, where $\text{rank}(p_m) := \#\{i : p_i \leq p_m\}$ *is the rank of p_m among the full list $p_1, \ldots, p_m$, breaking ties at random.*

The assumption that $L \to \infty$ in the previous lemma is needed to rule out small cases where it is possible for the inequality to be violated by a factor of 2.[5] Our next result characterizes the boundary FDR of the SL_α procedure as $m \to \infty$, keeping L fixed.

Theorem 5.3.3. *Let $p_1, \ldots, p_m$ be drawn independently from the fixed effects model (5.3), where $H_i = 0$ implies $p_i \sim \text{Uniform}\left\{\frac{1}{L}, \frac{2}{L}, \ldots, \frac{L-1}{L}, 1\right\}$. Suppose as $m \to \infty$ that the average probability mass function $\bar{f}$ converges to a limiting pmf f^*, supported on $\left\{\frac{1}{L}, \frac{2}{L}, \ldots, \frac{L-1}{L}, 1\right\}$, and that $\frac{m_0}{m} \to \pi_0^* \in (0,1)$. Further assume that there is a unique maximizer ℓ^* of the population objective,*

$$\ell^* := \underset{\ell=0,\ldots,L}{\operatorname{argmax}} \left\{ \alpha \sum_{k=0}^{\ell} f^*(k/L) - \ell/L \right\}.$$

Then for fixed L, we have

$$\lim_{m\to\infty} \mathrm{bFDR}(\mathcal{R}_\alpha) = \frac{\pi_0^*}{Lf^*(\ell^*/L)} \cdot 1_{\{\ell^*>0\}}.$$

Remark 5.3.3. *It follows as a corollary of Theorem 5.3.3 that $\mathrm{bFDR}(\mathcal{R}_\alpha) \le \pi_0^*\alpha$ asymptotically (as $m \to \infty$, keeping L fixed) because ℓ^* must occur at an $\ell \le L$ for which the discrete difference sequence is non-negative,*

$$\alpha f^*(\ell^*/L) - 1/L \ge 0,$$

which implies that $\frac{\pi_0^*/L}{f^*(\ell^*/L)} \le \pi_0^*\alpha$.

5.4 Compound local false discovery rate

The SL procedure can be understood as estimating the lfdr,

$$\widehat{\mathrm{lfdr}}(t) := \frac{1}{\hat{f}_m(t)},$$

rejecting $H_i = 0$ when $\widehat{\mathrm{lfdr}}(p_i) \le \alpha$, where $\hat{f}_m$ is Grenander's estimator for a decreasing density $\bar{f}$ (Grenander (1956), Strimmer (2008)). This procedure approximates a simple rule in the oracle two-groups model (Section 5.2.1):

$$\text{Reject } H_i = 0 \text{ if } \mathrm{lfdr}(p_i) \le \bar{\pi}_0\alpha,$$

which is *simple* in the sense that the i^{th} decision depends only on p_i. Suppose now that the i^{th} decision is allowed to depend on all entries of p in a symmetric way. Formally, the

[5] A counterexample is obtained in Appendix 5.7 by setting $p_1, \ldots, p_{m-1}$ to specific values for which $\mathbb{P}(p_{(R)} = p_m) > \frac{2\alpha}{m}$ when $L = 9, m = 6, \alpha = 1/2$.

decision rule $\delta(p) \coloneqq (\delta_1(p), \ldots, \delta_m(p))$ is is permutation invariant (PI) if

$$\delta_\pi(p) = \delta(p_\pi) \quad \text{for any } \pi \in \mathcal{S}_m,$$

where $\mathcal{S}_m$ is the set of permutations on $[m]$, and $v_\pi \coloneqq (v_{\pi(1)}, \ldots, v_{\pi(m)})$ denotes the vector $v \in \mathbb{R}^m$ permuted by π. The SL procedure is permutation invariant, and a Storey-modified version approaches the best simple rule in an i.i.d. Bayes two-groups model with decreasing alternative density (Soloff et al.; 2022).

In the frequentist model, the best PI rule depends on the underlying densities $f^{(1)}, \ldots, f^{(m)}$ in a complex way, and can be quite different from the best simple rule when the number of tests is small. When there are many independent test statistics (e.g. $m \geq 1000$), the best PI rule for minimizing a weighted classification risk is roughly equivalent to the best simple rule under regularity conditions, a claim we formalize in the next section.

5.4.1 A compound decisions perspective

Weinstein (2021) details a general framework for analyzing permutation-invariant decision problems, recasting the aggregated risk function as a Bayes risk. The best rule in the induced Bayesian model determines the functional form of the best compound decision rule, subject to the PI constraint. The argument for the weighted classification risk in the current setting goes as follows.

For fixed $H_1, \ldots, H_m \in \{0, 1\}$, a random permutation induces an exchangeable Bayesian two-groups model $(\widetilde{H}_i, \widetilde{p}_i)_{i=1}^m$, where

$$\begin{gathered} \pi \sim \text{Uniform}(\mathcal{S}_m) \\ \widetilde{H} \coloneqq H_\pi, \quad \widetilde{p} \coloneqq p_\pi. \end{gathered}$$

The weighted classification risk of a PI rule in this model coincides with its frequentist compound risk, yielding an instance of the fundamental theorem of compound decisions (Zhang; 2003),

$$\mathbb{E}L_\lambda(H, \delta(p)) = \mathbb{E}L_\lambda(\widetilde{H}, \delta(\widetilde{p})),$$

from which it follows that the best PI decision rule is given by

$$\delta_i^*(p) \coloneqq 1_{\left\{\text{clfdr}_i(p) \leq \frac{1}{1+\lambda}\right\}}, \quad \text{clfdr}_i(u) \coloneqq \mathbb{P}(\widetilde{H}_i = 0 \mid \widetilde{p} = u), \tag{5.18}$$

where $u \in [0, 1]^m$. More explicitly, the compound local false discovery rate (clfdr) is equal to

$$\text{clfdr}_i(p) = \frac{\sum_{\pi \in \mathcal{S}_m : H_{\pi(i)}=0} \prod_{j=1}^m f^{(\pi(j))}(p_j)}{\sum_{\pi \in \mathcal{S}_m} \prod_{j=1}^m f^{(\pi(j))}(p_j)}, \quad i = 1, \ldots, m$$

and is *compound* in the sense that $\operatorname{clfdr}(p_i) \coloneqq \operatorname{clfdr}_i(p)$ depends on all entries in p. This quantity was first identified by Weinstein (2021) as the optimal thresholding statistic for minimizing a PI loss function. Our next result gives an interpretation of clfdr in the frequentist model.

Proposition 5.4.1. *If $p_1, \dots, p_m$ are independently generated from the fixed effects model* (5.3)*, then for any K that is a function of the order statistics,*

$$\mathbb{P}(H_{(K)} = 0 \mid p_{(1)}, \dots, p_{(m)}) = \operatorname{clfdr}(p_{(K)}),$$

where $\operatorname{clfdr}(p_{(K)}) \coloneqq \operatorname{clfdr}_i(p)$ *on the event* $\{p_i = p_{(K)}\}$.

In words, the compound lfdr for the K^{th} smallest p-value is equal to the probability that its hypothesis is null, conditional on observing just the order statistics, i.e. the data without order. Given the true densities, the clfdr can typically only be computed in small problems (e.g. $m \leq 20$), but can be approximated in larger problems (e.g. $m \approx 1000$) using a method for approximating the permanent of a matrix with positive entries (McCullagh; 2014).

The compound lfdr provides an alternative to the lfdr (5.5) for assigning a confidence score to each of the observed p-values, and matches the oracle restricted to PI decision rules. By contrast, the simple lfdr matches the oracle restricted to a separable rule, and is recovered in this model by conditioning on just one observation,

$$\operatorname{lfdr}(p_i) = \mathbb{P}(\widetilde{H}_i = 0 \mid \widetilde{p}_i), \quad i = 1, \dots, m.$$

Under sufficiently regular conditions, the best simple decisions are asymptotically efficient (Greenshtein and Ritov; 2009), coinciding with the best permutation-invariant decisions.

Proposition 5.4.2 (Corollary of Theorem 3.1 in Greenshtein and Ritov (2009))**.** *Suppose p-values are drawn independently from the fixed effects model* (5.3) *with continuous densities $f^{(i)} = f_0$ when $H_i = 0$ and $f^{(i)} = f_1$ when $H_i = 1$. If $\frac{m_0}{m} \to \pi_0 \in (0,1)$ as $m \to \infty$, and* $\operatorname{Var}\left(\frac{f_1}{f_0}(p_1)\right) \vee \operatorname{Var}\left(\frac{f_0}{f_1}(p_2)\right) < \infty$ *when $p_1 \sim f_0$ and $p_2 \sim f_1$, then we have for each $i = 1, 2, \dots$*

$$\frac{\operatorname{clfdr}_i(p)}{\operatorname{lfdr}(p_i)} \xrightarrow{\mathbb{P}} 1$$

as $m \to \infty$.

The variance condition in the above result holds for instance when f_0 is continuous uniform and f_1 is the $\operatorname{Beta}(a, b)$ density with $a, b \in (1/2, 2)$. Figure 5.6 shows a scatter plot comparing the compound lfdr[6] to the simple lfdr for several realizations of $m = 1000$

[6] Since the expression for the compound lfdr involves a sum over all permutations, it is hard to compute for problems with m larger than ten or so. In these plots, the clfdr is approximated numerically by iteratively dividing the matrix $M \in \mathbb{R}^{m \times m}$ of density evaluations $M_{ij} \coloneqq f^{(i)}(p_j)$ by its row and column sums to obtain a doubly stochastic matrix (Sinkhorn and Knopp; 1967), and applying a deterministic asymptotic formula for the permanent of a doubly stochastic matrix (McCullagh; 2014).

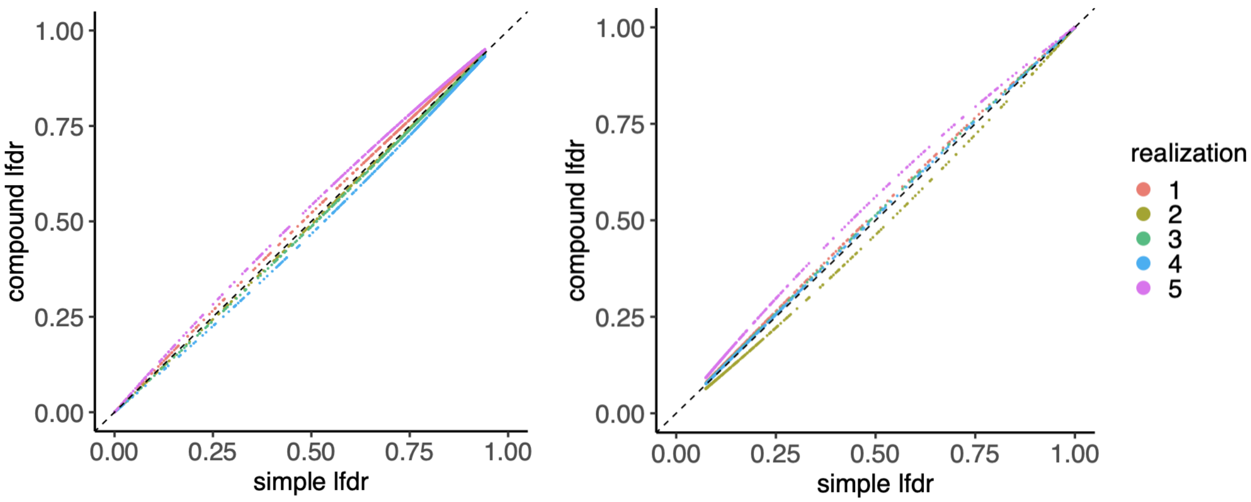


Figure 5.6: For each of 5 realizations of $p_1, \ldots, p_m$, with $m = 1000$, $\bar{\pi}_0 = 0.8$, $f_0 = 1_{[0,1]}$ and either $f_1 = \text{Beta}(1/4, 1)$ (left) or $f_1 = \text{Beta}(1, 50)$ (right), the $\text{clfdr}(p)$ is approximated numerically and the points $(\text{lfdr}(p_i), \text{clfdr}_i(p))$ are plotted with the diagonal $y = x$ shown as a dashed line.

p-values with $\bar{\pi}_0 = 0.8$, uniform nulls, and a common alternative, either $f_1 = \text{Beta}(1/4, 1)$ (left) or $f_1 = \text{Beta}(1, 50)$ (right).

5.4.2 Connection with boundary FDR

As a local analogue to the frequentist FDR, the boundary FDR bears a close connection to the concept of lfdr and the definition presented in Section 5.2. In the frequentist model, both the lfdr and the boundary FDR can be understood as the expectation of a random coordinate of $\text{clfdr}(p)$.

Proposition 5.4.3. *Suppose that $t \in \{p_1, \ldots, p_m\}$ for independent and continuously distributed p-values drawn from the fixed effects model* (5.3)*. If $J \in [m]$ denotes the index for which $p_J = t$, then*

$$\text{lfdr}(t) = \mathbb{E}\left(\text{clfdr}(p_J) \mid t \in \{p_1, \ldots, p_m\}\right).$$

where $\text{clfdr}(p_J) \coloneqq \text{clfdr}_i(p)$ *on the event $\{J = i\}$. If R is a function of the order statistics,*

$$\text{bFDR}(\mathcal{R}) = \mathbb{E}\left(\text{clfdr}(p_{(R)})\right),$$

where $\mathcal{R} = \{i : p_i \leq p_{(R)}\}$, and $\text{clfdr}(p_{(R)}) \coloneqq \text{clfdr}_i(p)$ *on the event $\{p_{(R)} = p_i\}$.*

For the SL_α set, the bFDR agrees asymptotically with lfdr at the threshold $\hat{\tau}_\alpha = p_{(R_\alpha)}$, which concentrates around a population threshold τ_α^* that is the solution to $\text{lfdr}(\tau_\alpha^*) = \bar{\pi}_0\alpha$. Lemmas 5.7.5 and 5.7.6 show that the difference $\hat{\tau}_\alpha - \tau_\alpha^*$ scales as $m^{-1/3}$ with high probability, up to a log factor in m. Under some additional regularity conditions, this implies that $\mathrm{lfdr}(\hat{\tau}_\alpha)$ converges in probability to the boundary FDR for the SL_α procedure at the same rate.

Theorem 5.4.1. *Suppose (p_i) are generated independently from the fixed effects model (5.3), where each $P^{(i)}$ has a continuous density $f^{(i)}$ that is uniform when $H_i = 0$, and that $\bar{f}$ has a unique solution τ_α^* to the equation $\bar{f}(\tau_\alpha^*) = \alpha^{-1}$. If $\bar{f}$ is decreasing, and for some constants $\delta, J > 0$ we have $J \leq |\bar{f}'(t)| \leq J^{-1}$ for all t with $|t - \tau_\alpha| \leq \varepsilon$, where $\varepsilon := \left(\frac{48}{\alpha J^2}\right)^{1/3} m^{-1/3} \log(2m/\delta)$, then for a constant $C > 0$ depending on α, J and δ,*

$$\mathbb{P}\left(|\mathrm{lfdr}(\hat{\tau}_\alpha) - \mathrm{bFDR}(\mathcal{R}_\alpha)| > Cm^{-1/3}\log(m/\delta)\right) \leq \delta,$$

where $\mathcal{R}_\alpha := \{i : p_i \leq p_{(R_\alpha)}\}$ is the SL_α rejection set defined by (5.16).

Remark 5.4.1. *The non-standard rate $m^{-1/3}$ can be understood by balancing the mean and variance of the objective near τ_α^* as follows. The procedure (5.16) is equivalent to computing,*

$$\hat{\tau}_\alpha = \underset{t \in [0,1]}{\operatorname{argmax}}\ U_m(t)$$
$$U_m(t) := F_m(t) - F_m(\tau_\alpha^*) - \alpha^{-1}(t - \tau_\alpha^*)$$

where F_m is the empirical cdf of the p-values. Parametrizing $t - \tau_\alpha^* = m^{-a}h$ for some $h, a > 0$, the mean and standard deviation of $U_m(t)$ are approximately

$$-\frac{|\bar{f}'(\tau_\alpha^*)|}{2}(t - \tau_\alpha^*)^2 \asymp -m^{-2a}h^2, \qquad \sqrt{\frac{\alpha^{-1}(t - \tau_\alpha^*)}{m}} \asymp m^{-\frac{a+1}{2}}h.$$

The objective $U_m(t)$ is positive with non-negligible probability when the second term is of larger order than the first, the tipping point occurring when $a = 1/3$. Fixing $h > 0$ and $a = 1/3$, the random process $m^{2/3}U_m(t)$ converges to a Brownian motion with parabolic drift,

$$m^{2/3}U_m(t) \xrightarrow{d} -\frac{|\bar{f}'(\tau_\alpha^*)|}{2}h^2 + N(0, h/\alpha),$$

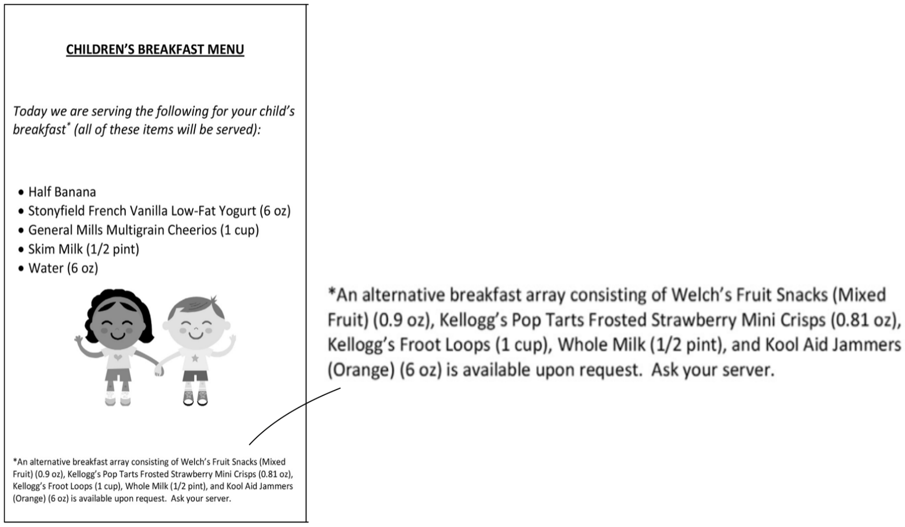
the maximizer of which is characterized by a known distribution (Chernoff; 1964). A more detailed derivation is supplied in Lemmas 5.7.5 and 5.7.6, stated and proved in Appendix 5.7.

5.5 Applications

We illustrate some of the key ideas from this chapter on two real datasets. We first continue our analysis of the aggregate nudge data from Section 5.1.1, and then compare some methods for estimating local fdr on a gene-expression microarray dataset for prostate cancer taken from Chapter 6 of Efron (2012).

5.5.1 Mertens et al. (2022a) nudge meta-analysis

In one of the studies from the Mertens et al. (2022a) dataset, parents were split into two groups and shown a menu from which each parent would select one of two breakfast options for their child. The first group was shown a menu with the healthy option displayed in large font in the center of the menu. A less healthy breakfast option was only shown in a footnote, illustrated in Figure 5.7. The second group saw the same menu but with the placement of the two options flipped. The researchers then recorded the proportion of parents within each group that selected the healthy option, and found that the first group was 79% more likely to choose the healthy option, with a large and positive t-statistic ($t = 6.2$, Loeb et al. (2017)), indicating that the researcher's nudge was effective.



CHILDREN'S BREAKFAST MENU

Today we are serving the following for your child's breakfast (all of these items will be served):*

- Half Banana
- Stonyfield French Vanilla Low-Fat Yogurt (6 oz)
- General Mills Multigrain Cheerios (1 cup)
- Skim Milk (1/2 pint)
- Water (6 oz)

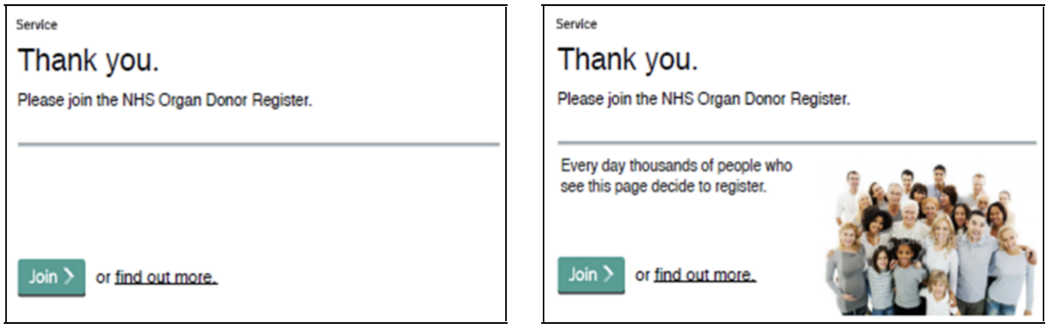
*An alternative breakfast array consisting of Welch's Fruit Snacks (Mixed Fruit) (0.9 oz), Kellogg's Pop Tarts Frosted Strawberry Mini Crisps (0.81 oz), Kellogg's Froot Loops (1 cup), Whole Milk (1/2 pint), and Kool Aid Jammers (Orange) (6 oz) is available upon request. Ask your server.

*An alternative breakfast array consisting of Welch's Fruit Snacks (Mixed Fruit) (0.9 oz), Kellogg's Pop Tarts Frosted Strawberry Mini Crisps (0.81 oz), Kellogg's Froot Loops (1 cup), Whole Milk (1/2 pint), and Kool Aid Jammers (Orange) (6 oz) is available upon request. Ask your server.

Figure 5.7: A breakfast menu from the Loeb et al. (2017) study with the healthy option set as default.

Another study in the Mertens et al. (2022a) meta-analysis supplied data collected by researchers from the NHS Organ Donor Registry in the U.K. who hoped to devise nudges that could increase the number of organ donors. Drivers were split into groups and, after having their license renewed online, were prompted to become organ donors. One group was shown a panel which simply offered the option to join, while another group was shown the same panel, but with an additional image of smiling organ donors and extra text suggesting that organ donation is a social norm. These two prompts are displayed in Figure 5.8.

The researchers measured the proportion in each group that joined the organ donor registry, and found that drivers in the group that saw the second panel in Figure 5.8 were slightly less likely to join, with an estimated decrease of 5% in the sign-up rate relative to the control group, and a corresponding t-statistic of -1.7. With a negative t-statistic, there is no evidence to suggest that the additional image promoted organ donor registration. By contrast, the nudge in the same organ donation study that used another nudge based on a concept of reciprocity produced a much stronger effect. The reciprocity nudge increased the sign-up rate by 35% relative to the control group, with an associated t-statistic of 13.1.



Service

Thank you.

Please join the NHS Organ Donor Register.

Service

Thank you.

Please join the NHS Organ Donor Register.

Every day thousands of people who see this page decide to register.

Figure 5.8: The original prompt from the organ donation study by Sallis et al. (2018) is shown on the left and the 'social-norm' nudge on the right.

The organ donation nudges and the breakfast nudge are just a few examples that demonstrate the heterogeneity in context and effectiveness of the nudges analyzed in the Mertens et al. (2022a) dataset. In classical meta-analysis, the studies to be aggregated typically contain estimates of a common effect. For instance, Peto (1996) was interested in estimating the effect of the drug tamoxifen on survival rates among breast cancer patients using data from trials conducted in Europe and North America for different lengths of the treatment, ranging from two to five to ten years. In this setting, the effect of the drug tamoxifen is assumed to be constant across locations. It is reasonable to aggregate the evidence with the hope to determine whether tamoxifen is effective, and to quantify the overall effect of each treatment. These estimates might then be used to inform prescriptions for the duration of tamoxifen treatment on future breast cancer patients.

Estimation of the overall effect is a well-defined statistical task, but the estimand is not meaningful in the nudge setting. What does it mean to aggregate a breakfast nudge with an organ donation nudge, and what use is there for an estimate of their average effect? Some of the nudges worked, some of them didn't and some of them may well have backfired; scientifically interesting departures from the global null can be consistent with a negligible overall effect (Szaszi et al.; 2022). Instead of estimating the average effect of all nudges being studied in this area of the psychology literature, we might rather ask which of the nudges worked.

Mertens et al. (2022a) compiled data from 447 nudge experiments and estimated a positive overall nudge effect, deemed statistically significant. Maier et al. (2022) responded saying there was "no evidence for nudging after adjusting for publication bias". In the examples presented above, evidence for the effectiveness of particular nudges ranged widely. For instance, the reciprocity nudge in the organ donation study had a t-statistic of 13.1, whereas the social norm nudge was insignificant. The breakfast nudge was moderately significant, with a t-statistic near 6. Instead of asking whether nudges are effective on average, we ought to focus on identifying the promising ones with some control over our type 1 error. Test statistics as large as 13 and 6 clearly indicate evidence of an effect, but it is less clear what to make of other studies in the nudge dataset with t-statistics around, say 3 or 4. In the next section, we estimate a scale of statistical significance tailored to assessing nudge effects,

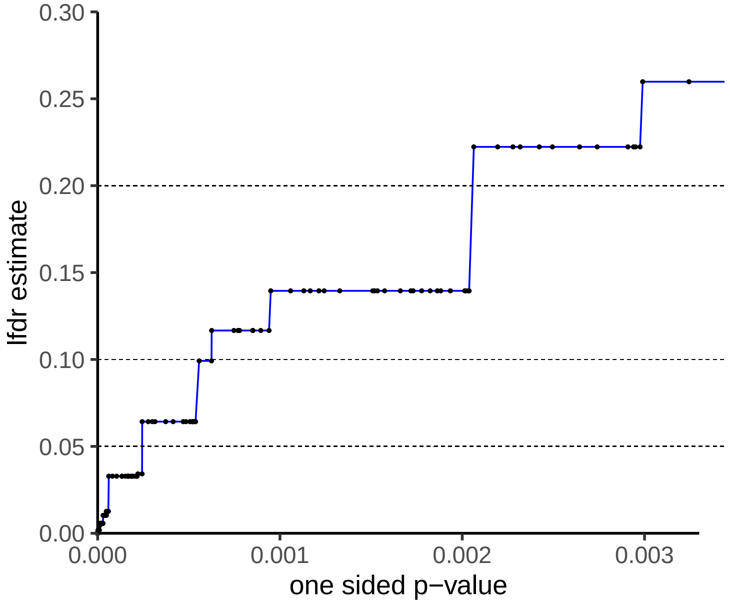


Figure 5.9: A plot is shown of the estimated local false discovery rate for the nudge data. Dashed horizontal lines indicate common cut-offs for the estimated lfdr, and the corresponding one sided p-value cut-offs are roughly $0.00024, 0.00063$, and 0.0020 (see also Table 5.2). Black dots indicate the observed (un-adjusted) p-values and their estimated lfdr values.

based on estimates of the lfdr using the Mertens et al. (2022a) dataset.

Which nudges were effective?

We estimate a more stringent threshold, forty times smaller than the standard 0.025 one-sided p-value cut-off, at which approximately 10% of the findings in this sub-area of the psychology literature with p-values near 0.025/40 are false positives. To target a 5% false discovery rate near the rejection threshold, the standard p-value cut-off should instead be reduced by a factor of about a hundred. For a 20% false discovery rate at the margin, we find that the cut-off should be reduced by a factor of about twelve, which roughly agrees with a more general proposal to reduce the cut-off from 0.05 to 0.005 (see e.g. Greenwald et al. (1996) and Benjamin et al. (2018)). The estimated cut-offs are summarized in Table 5.2.

lfdr cut-off	0.05	0.10	0.20	0.30
one sided p-value cut-off	2.44×10^{-4}	6.25×10^{-4}	2.04×10^{-3}	3.70×10^{-3}

Table 5.2: Estimated significance cut-offs for one-sided p-values in the nudge literature based on the Mertens et al. (2022a) dataset. See also the plot in Figure 5.9.

To estimate these significance thresholds for the nudge publications, we first estimate the base rate of nulls being published in this area of the literature at the 5% two-sided significance level. As explained in Section 5.1.1, restricting our attention to one-sided p-values below 2.5% is a way to work around the issue of publication bias, since these nulls are less likely to be censored by journals adhering to the classical 5% significance cut-off. If the null p-values are super-uniformly distributed, then those falling below the significance cut-off are super-uniform over the interval $(0, 0.025)$, and thus multiplying each p-value by 40 restores validity.

After this adjustment, we treat the scaled p-values as a fresh dataset of p-values. Using Storey's method, a conservative approximation to the rate of false discoveries among significant studies is roughly 28%. Incorporating this estimate of the null proportion, the SL procedure targeting 10% boundary FDR yields a rejection threshold of 6.25×10^{-4}. Running the SL procedure implicitly estimates the lfdr at each observed p-value, based on the Grenander estimator. These estimates are plotted in Figure 5.9 for the nudge data.

5.5.2 Prostate cancer dataset

The prostate cancer dataset, taken from Chapter 6 of Efron (2012), is a case-control study with gene-activity levels measured for 52 cases and 50 controls at 6033 genomic sites. At each site, a difference is computed between the average gene expression levels for cases and controls. In the analysis by Efron (2012), the goal is to identify a subset of the genes relevant towards understanding prostate cancer. The difference at the i^{th} genetic site is divided by a pooled standard error to obtain a t-score T_i, and these are transformed into z-scores via $Z_i = \bar{\Phi}^{-1}(1 - F_0(T_i))$, where F_0 is the Student-t cdf. After this pre-processing step, the (Z_i) are viewed as independent draws from a two-groups model,

$$\begin{aligned} H &\sim \text{Bernoulli}(1-\pi_0) \\ Z \mid H &\sim \begin{cases} f_0 & \text{if } H = 0 \\ f_1 & \text{if } H = 1, \end{cases} \end{aligned} \tag{5.19}$$

where the 'zero-assumption' $f_1(0) = 0$ is made to ensure identifiability of the model.

Remark 5.5.1. *The 'zero assumption' made by Efron is necessary for identifiability of the two-groups model (Patra and Sen; 2016), but the event $H_i = 0$ in the two groups model* (5.19) *doesn't have a clear connection with the event $\theta_i = 0$ in convolutional model $Z_i = \theta_i + \varepsilon_i$, the latter of which may occur with zero probability when θ_i is random. In the identified model, the estimand $\mathbb{P}(H_i = 0 \mid Z_i)$ is typically larger than the posterior probability $\mathbb{P}(\theta_i = 0 \mid Z_i)$.*

In Figure 5.10, we plot the estimates of the local fdr based on Grenander's method and Lindsey's method (Lindsey (1974a), Lindsey (1974b), Efron (2012)). Lindsey's method estimates $\hat{\pi}_0 = 0.998$ using maximum likelihood, as well as the null distribution,

$$\hat{f}_0 = N(0.0026, \hat{\sigma}_0 = 1.09).$$

Although this differs a little from the theoretical $N(0, 1)$ null, in order to make a comparison with the Grenander procedure, which takes as input the two-sided p-values $p_i :=$

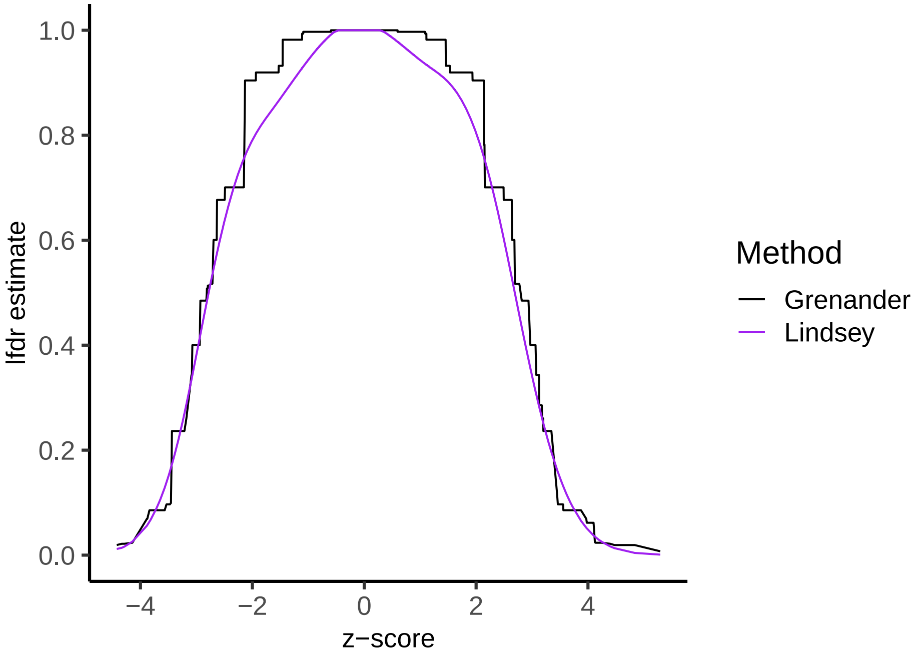


Figure 5.10: Estimated lfdr curves are plotted for the z-scores in the prostate cancer dataset. The Grenander estimate is plotted in black, with Lindsey's method overlaid in purple. Both curves are truncated below 1 since the estimand (5.4) is a probability.

$2(1 - \Phi(|Z_i|))$, we use the theoretical $N(0, 1)$ null when fitting Lindsey's method. To obtain the black curve in Figure 5.10, we transformed the p-values back to z-scores and plotted them against the reciprocal of the Grenander estimator evaluated at the corresponding p-value.

Running the BH procedure at levels $\alpha = 0.01, 0.02, 0.05, 0.10$ on the two-sided p-values yields 2, 13, 21, and 60 rejections. The average lfdr estimate based on Grenander or Lindsey's method among these rejections are displayed in Table 5.3.

The smallest fitted lfdr values based on the Grenander procedure tend to be larger than the smallest of Lindsey's estimates. Among the 106 rejections in the BH(0.2) set, the fitted lfdr values based on the Grenander procedure range from 0.007 to 0.485 with an average of 0.203, and for Lindsey's method from 0.001 to 0.421 with an average of 0.199. We also run the SL procedure at the same levels, and record the maximum estimated lfdr values within this set in the Table 5.4.

One strategy to mitigate the high rate of false discoveries at the edge of the BH set is to run the procedure at a range of levels, lowering the tuning parameter α until the estimated rate near the boundary falls below the targeted level. Based on the tables above, another rough approach could be to simply run an FDR procedure (e.g. BH) at level $\alpha/2$ to target bFDR at level α, but this approximation only appears reasonable in a small range of values, e.g. between $\alpha = 0.01$ and 0.05 for the prostate data. If the goal is to control the rate of false discoveries at level α near the decision point, then we recommend directly running the SL procedure, which has the same computational cost as a single run of the BH procedure as well as a provable guarantee.

Average lfdr estimate within the BH(α) set

α	#BH rejections	Grenander	Lindsey
0.01	2	0.013	0.003
0.02	13	0.024	0.020
0.05	21	0.043	0.035
0.10	60	0.099	0.109

Table 5.3: Comparison between the average estimate of lfdr using either Strimmer (2008) (column 3) or Efron et al. (2011) (column 4). The average is computed among rejections made by the BH(α) procedure run at levels $\alpha \in \{0.01, 0.02, 0.05, 0.10\}$.

Maximum lfdr estimate within the SL(α) set

α	#SL rejections	Grenander	Lindsey
0.01	1	0.007	0.001
0.02	4	0.019	0.013
0.05	12	0.024	0.035
0.10	49	0.100	0.167
0.20	50	0.118	0.167

Table 5.4: Comparison between the maximum estimate of lfdr using either Strimmer (2008) (column 3) or Efron et al. (2011) (column 4). The maximum lfdr estimate is reported among rejections made by the BH(α) procedure run at levels $\alpha \in \{0.01, 0.02, 0.05, 0.10\}$.

5.6 Discussion

5.6.1 Calibration of p-values

We return to the question posed at the start of this chapter: on average, of all the studies in this literature with p-values close to 0.001, for what fraction was the null hypothesis actually true? We have argued in Section 5.2.3 to answer this question by evaluating the local false discovery rate at $t = 0.001$. When $m = 1$, i.e. there is one fixed hypothesis to test, the formula (5.5) for the lfdr at any $t \in [0, 1]$ reduces to either 0 or 1, reflecting the difficulty in making posterior inferences based on a single observation. Not much can be done in this case without specifying a subjective prior probability on the truth status of the hypothesis. As Theorem 5.2.2 and Proposition 5.2.1 demonstrate, when there are many hypotheses to test, lfdr is the local frequency of nulls in a small neighborhood of the sample space,

$$\alpha \approx \frac{\#\{i : H_i = 0, \mathrm{lfdr}(p_i) \approx \alpha\}}{\#\{i : \mathrm{lfdr}(p_i) \approx \alpha\}}.$$

The value $\text{lfdr}(p_i)$ thus forecasts the proportion of hypotheses that are null with p-values near p_i in the large sample limit. In the next section, we compare two approaches to calibration on the nudge data from Section 5.5.1; the first is based on the SL procedure, and the second is a conservative approach proposed by Sellke et al. (2001).

Comparison with Sellke et al. (2001) on nudge data

We stated in Section 5.5 for the nudge data that by reporting statistical significance on the scale of the lfdr (as opposed to the p-value), researchers claiming a certain false positive probability will collectively achieve the rate they claim. For instance, a researcher who publishes a result with one sided p-value near 0.002 implicitly reports a value of $\text{lfdr}(0.002) \approx 20\%$, according to our estimates based on the meta-analysis data collected by Mertens et al. (2022a). In other words, about 20% of researchers who observed nearby p-values make false discoveries. For studies with p-values ten times as small as this, we estimate fewer than 5% of them to be false discoveries (see Table 5.2).

The p-value is often misinterpreted as a posterior probability that the tested hypothesis is null (Goodman; 1999); however, the actual numeric value of the observed p_i can differ substantially from the proportion of p-values near p_i that correspond to true null hypotheses (Sellke et al.; 2001). We have claimed in Theorem 5.2.2 that the latter is close to the local false discovery rate at p_i when there are many independent tests, and a numerical illustration of this difference can be found in Section 2 of Sellke et al. (2001). In their chapter, the authors proposed to combat this misinterpretation by introducing a method for calibration, which is to compute

$$B(p_i) := -ep_i \log(p_i),$$

for $p_i < 1/e \approx 0.368$, and interpret this number as a lower bound on the Bayes factor for $H_i = 0$ against $H_i = 1$. Multiplying this value by the prior odds gives a lower bound on the posterior odds when $H_i = 1 \Rightarrow p_i \sim \text{Beta}(\xi, 1)$ for any prior on $\xi \in (0, 1]$,

$$\frac{\text{lfdr}(p_i)}{1 - \text{lfdr}(p_i)} \geq B(p_i) \times \frac{\bar{\pi}_0}{1 - \bar{\pi}_0},$$

which can be converted to a lower bound on $\text{lfdr}(p_i)$. Plugging in our estimate $\bar{\pi}_0 \approx 0.28$ from Section 5.5.1 gives an estimated lower bound on the local false discovery rate at various p-value cut-offs. Four of these estimates are displayed in Table 5.5.

5.6.2 Multiple testing with lfdr

Exact guarantees like the ones in Theorem 5.3.1 and in Theorem 1 of Benjamini and Hochberg (1995) are useful to have in finite samples, where we don't know if we've done a good job estimating the signal distribution based on the few non-nulls distinguishable from the bulk of the data. However, the prioritization of exact FDR guarantees has led attention away from the wider implications of the BH idea, namely that by including a few extra strong bets

one sided p-value cut-off	2.44×10^{-4}	6.25×10^{-4}	2.04×10^{-3}	3.70×10^{-3}
lfdr cut-off	0.05	0.10	0.20	0.30
Sellke et al. (2001) lower bound	0.045	0.09	0.17	0.23

Table 5.5: Entries in this table compare the Grenander estimate of the lfdr with the lower bound from Sellke et al. (2001). The value in each entry of the first row is obtained by running SL on the adjusted p-values at level $\alpha/\bar{\pi}_0$ for $\alpha = 0.05, 0.10, 0.20, 0.30$, and recording the rejection threshold on the scale of the original (un-adjusted) p-values.

in our rejection set, we become more willing to reject hypotheses for which the evidence is weak. Average case guarantees let bad bets in through the cracks, and it is easy to recognize that the lowest quality bets are made near the rejection threshold, so why make them in the first place? SL also satisfies an exact bound, but its true merit comes from the fact that it is consistent with what an oracle would do[7]. An oracle would not choose to control FDR; they would minimize a combination of type 1 and 2 errors, and we have argued that lfdr is the right tool for converting p-values to the appropriate scale for making this trade-off.

There is room to extend the ideas in this chapter to multiple testing problems where more information is observed. Namely, it may be worthwhile to modify procedures that account for additional structure such as covariate information and dependence (e.g. Barber and Candès (2015), Lei and Fithian (2018), Fithian and Lei (2022)) to control the rate of false discoveries among their least promising rejections. The local fdr concept will be important in developing such extensions of frequentist FDR methods.

5.7 Proofs

Proof of Theorem 5.2.1. For the first part, the lfdr is a limit of posterior probabilities

$$\begin{aligned}
\mathrm{lfdr}(t) &= \lim_{\varepsilon \to 0} \mathbb{P}(H_J = 0 \mid |p_J - t| \le \varepsilon \text{ for some } J \in [m]) \\
&= \lim_{\varepsilon \to 0} \frac{\mathbb{P}(H_J = 0, |p_J - t| \le \varepsilon \text{ for some } J \in [m])}{\mathbb{P}(|p_J - t| \le \varepsilon \text{ for some } J \in [m])}.
\end{aligned} \tag{5.20}$$

The event in the numerator can be written as a union,

$$\{H_J = 0, |p_J - t| \le \varepsilon \text{ for some } J \in [m]\} = \bigcup_{j : H_j = 0} \left\{ |p_j - t| \le \varepsilon \right\}.$$

[7] an oracle without knowledge of $\bar{\pi}_0$

By independence,

$$\begin{aligned}\mathbb{P}\Bigg(\bigcup_{j:H_j=0}\{|p_j-t|\le\varepsilon\}\Bigg) &= 1-\mathbb{P}\Bigg(\bigcap_{j:H_j=0}\{|p_j-t|>\varepsilon\}\Bigg)\\ &= 1-\prod_{j:H_j=0}\mathbb{P}\big(|p_j-t|>\varepsilon\big)\\ &= 1-\prod_{j:H_j=0}\Big(1-\Big(F^{(j)}(t+\varepsilon)-F^{(j)}(t-\varepsilon)\Big)\Big)\\ &= 1-\prod_{j:H_j=0}\Big(1-2\varepsilon f^{(j)}(\xi_j)\Big),\end{aligned}$$

for some $\xi_1,\ldots,\xi_m\in[t-\varepsilon,t+\varepsilon]$ by the mean value theorem. Now letting $\varepsilon\to 0$, since each $f^{(j)}$ is continuous,

$$\prod_{j:H_j=0}\Big(1-2\varepsilon f^{(j)}(\xi_j)\Big)\sim\exp\left(-2m_0\varepsilon\bar{f}_0(t)\right),$$

where $\bar{f}_0(t)=\frac{1}{m_0}\sum_{j:H_j=0}f^{(j)}(t)$ is the average null density. The above implies

$$\mathbb{P}\Bigg(\bigcup_{j:H_j=0}\{|p_j-t|\le\varepsilon\}\Bigg)\sim 2m_0\varepsilon\bar{f}_0(t)\quad\text{as }\varepsilon\to 0.$$

An identical argument will show

$$\mathbb{P}\Bigg(\bigcup_{j=1}^{m}\{|p_j-t|\le\varepsilon\}\Bigg)\sim 2m\varepsilon\bar{f}(t)\quad\text{as }\varepsilon\to 0,$$

where $f(t)=\frac{1}{m}\sum_{j=1}^{m}f^{(j)}(t)$ is the average density. Dividing these two expressions gives

$$\lim_{\varepsilon\to 0}\frac{\mathbb{P}(H_J=0,|p_J-t|\le\varepsilon\text{ for some }J\in[m])}{\mathbb{P}(|p_J-t|\le\varepsilon\text{ for some }J\in[m])}=\lim_{\varepsilon\to 0}\frac{2m_0\varepsilon\bar{f}_0(t)}{2m\varepsilon f(t)}=\frac{\bar{\pi}_0\bar{f}_0(t)}{\bar{f}(t)}.$$

Now suppose that the test statistics are generated from the independently from the two-groups model (5.1). The lfdr is a limit of posterior probabilities,

$$\text{lfdr}(t)=\lim_{\varepsilon\to 0}\frac{\mathbb{P}\left(\bigcup_{j=1}^{m}\{H_j=0,|p_j-t|\le\varepsilon\}\right)}{\mathbb{P}\left(\bigcup_{j=1}^{m}\{|p_j-t|\le\varepsilon\}\right)}.\tag{5.21}$$

Since the pairs (H_j, p_j) are independent across $j = 1, \ldots, m$, the numerator is equal to

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{j=1}^{m}\{H_j = 0, |p_j - t| \le \varepsilon\}\right) &= 1 - \prod_{j=1}^{m}\big(1 - \mathbb{P}(H_j = 0, |p_j - t| \le \varepsilon)\big) \\
&= 1 - \prod_{i=1}^{m}(1 - \pi_0(F_0(t+\varepsilon) - F_0(t-\varepsilon))) \\
&= 1 - (1 - \pi_0(F_0(t+\varepsilon) - F_0(t-\varepsilon)))^m \\
&= 1 - (1 - \pi_0 \cdot 2\varepsilon f_0(\xi)))^m
\end{aligned}$$

for some $\xi \in (t - \varepsilon, t + \varepsilon)$ by the mean value theorem. As $\varepsilon \to 0$,

$$(1 - \pi_0 \cdot 2\varepsilon f_0(\xi)))^m \sim \exp\left(-m\pi_0 \cdot 2\varepsilon f_0(\xi))\right) \sim 1 - 2m\pi_0\varepsilon f_0(t)$$

since f_0 is continuous and $\xi \to t$ as $\varepsilon \to 0$. It follows that

$$\mathbb{P}\left(\bigcup_{j=1}^{m}\{H_j = 0, |p_j - t| \le \varepsilon\}\right) \sim 2m\pi_0\varepsilon f_0(t) \quad \text{ as } \varepsilon \to 0.$$

An identical argument shows

$$\mathbb{P}\left(\bigcup_{j=1}^{m}\{|p_j - t| \le \varepsilon\}\right) \sim 2m\varepsilon f(t),$$

which implies that the ratio (5.21) tends to $\frac{\pi_0 f_0(t)}{f(t)}$, as desired. □

Proof of Proposition 5.2.1. Let $S_m := \{t \in [0,1] : |\mathrm{lfdr}(t) - \alpha| \le \varepsilon_m\}$, and define the count variables

$$N_0 := \#\{i \in \mathcal{H}_0 : p_i \in S_m\}, \quad N_1 := \#\{i \in \mathcal{H}_1 : p_i \in S_m\}.$$

To show that $\frac{N_0}{1 \vee (N_0 + N_1)} - \alpha \to 0$ in probability, it is equivalent to show that

$$\frac{N_0}{\mathbb{E}N_0} - 1 \xrightarrow{\mathbb{P}} 0, \quad \text{and} \quad \frac{N_0 + N_1}{\mathbb{E}(N_0 + N_1)} - 1 \xrightarrow{\mathbb{P}} 0 \quad \text{ as } m \to \infty, \tag{5.22}$$

since the ratio of expectations $\frac{\mathbb{E}(N_0)}{\mathbb{E}(N_0+N_1)} \to \mathrm{lfdr}(\mathrm{lfdr}^{-1}(\alpha)) = \alpha$. By Chebyshev's inequality, we have for any fixed $\delta > 0$,

$$\mathbb{P}\left(\left|\frac{N_0}{\mathbb{E}N_0} - 1\right| > \delta\right) \le \frac{\mathrm{Var}(N_0)}{(\mathbb{E}N_0)^2\delta^2} \asymp \frac{m_0\varepsilon_m}{(m_0\varepsilon_m)^2}, \tag{5.23}$$

because $H_i = 0$ implies $\mathbb{P}(\mathrm{lfdr}(p_i) \in \alpha \pm \varepsilon_m) \asymp \varepsilon_m$ as $m \to \infty$, since

$$\alpha - \varepsilon_m \leq \frac{\bar{\pi}_0}{\bar{f}(p_i)} \leq \alpha + \varepsilon_m \iff \bar{f}^{-1}\left(\frac{1}{\alpha/\bar{\pi}_0 - \varepsilon_m}\right) \leq p_i \leq \bar{f}^{-1}\left(\frac{1}{\alpha/\bar{\pi}_0 + \varepsilon_m}\right),$$

where we have assumed $\bar{f}$ is decreasing and $(\bar{f}^{-1})'(\bar{\pi}_0/\alpha) > 0$. The right hand side of (5.23) tends to zero since $m_0\varepsilon_m \to \infty$. Similarly,

$$\mathbb{P}\left(\left|\frac{N_0 + N_1}{\mathbb{E}(N_0 + N_1)} - 1\right| > \delta\right) \leq \frac{\mathrm{Var}(N_0 + N_1)}{(\mathbb{E}(N_0 + N_1))^2\delta^2} \asymp \frac{m\varepsilon_m}{(m\varepsilon_m)^2} \leq \frac{1}{m_0\varepsilon_m} \to 0,$$

from which (5.22) follows. □

Proposition 5.7.1. *Suppose $p_i \sim f^{(i)}$ are independently generated for $i = 1, \ldots, m$ where $H_i = 0 \Rightarrow f^{(i)}$ is the Uniform$(0,1)$ distribution, $\bar{f}$ is decreasing and differentiable, and $\varepsilon_m \to 0$ is a sequence for which $\varepsilon_m \gtrsim m_0^{-1+\delta}$ for some constant $\delta > 0$. If $(\bar{f}^{-1})'(\bar{\pi}_0/\alpha) > 0$ is bounded away from zero, then for any α in the range of* lfdr : $[0,1] \to [0,1]$,

$$\mathrm{FDP}\left(\{i : |\mathrm{lfdr}(p_i) - \alpha| \leq \varepsilon_m\}\right) \xrightarrow{a.s.} \alpha \text{ as } m \to \infty.$$

Proof of Proposition 5.7.1. By the reasoning in the proof of Theorem 5.2.1, it suffices to show

$$\frac{N_0}{\mathbb{E}N_0} - 1 \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \frac{N_0 + N_1}{\mathbb{E}(N_0 + N_1)} - 1 \xrightarrow{\text{a.s.}} 0 \quad \text{as } m \to \infty. \tag{5.24}$$

Let $S_m := \{t \in [0,1] : |\mathrm{lfdr}(t) - \alpha| \leq \varepsilon_m\}$. The first convergence is equivalent to

$$\frac{1}{m_0\varepsilon_m} \sum_{i \in \mathcal{H}_0} \left(1_{\{p_i \in S_m\}} - \mathbb{P}(p_i \in S_m)\right) \to 0, \quad \text{almost surely} \tag{5.25}$$

since $\mathbb{E}N_0 \asymp m_0\varepsilon_m$. Since $i \in \mathcal{H}_0 \Rightarrow p_i \sim$ Uniform$(0,1)$, the variance of each summand is

$$\mathrm{Var}\left(\frac{1_{\{p_i \in S_m\}}}{m_0\varepsilon_m}\right) \asymp \frac{\mathbb{P}(p_i \in S_m)}{(m_0\varepsilon_m)^2} \asymp \frac{1}{m_0^2\varepsilon_m}.$$

The sum of variances is

$$\sum_{m_0=1}^{\infty} \frac{1}{m_0^2\varepsilon_m} \lesssim \sum_{m_0=1}^{\infty} \frac{1}{m_0^{1+\delta}} < \infty.$$

Since $1_{\{p_i \in S_m\}} - \mathbb{P}(p_i \in S_m)$ has mean zero, Kolmogorov's strong law implies the convergence (5.25). The second convergence in (5.24) is proved similarly. □

Proof of Proposition 5.2.2. Let $\mathrm{lfdr}^{-1}(\alpha) := \{t : \mathrm{lfdr}(t) = \alpha\}$. The argument in the first part of the proof for Theorem 5.2.1 implies the left hand side of (5.14) is equivalent to

$$\mathbb{P}(H_J = 0 \mid p_J \in \mathrm{lfdr}^{-1}(\alpha) \text{ for some } J \in [m]) \sim \frac{\sum_{i:H_i=0} \mathbb{P}\left(\min_{t\in\mathrm{lfdr}^{-1}(\alpha)} |p_i - t| \leq \varepsilon\right)}{\sum_{i=1}^{m} \mathbb{P}\left(\min_{t\in\mathrm{lfdr}^{-1}(\alpha)} |p_i - t| \leq \varepsilon\right)},$$

as $\varepsilon \to 0$. If $\mathrm{lfdr}^{-1}(\alpha)$ is contiguous subset, then

$$\begin{aligned}\frac{\sum_{i:H_i=0} \mathbb{P}\left(\min_{t\in\mathrm{lfdr}^{-1}(\alpha)} |p_i - t| \leq \varepsilon\right)}{\sum_{i=1}^{m} \mathbb{P}\left(\min_{t\in\mathrm{lfdr}^{-1}(\alpha)} |p_i - t| \leq \varepsilon\right)} &= \frac{\sum_{i:H_i=0} \int_{\{t\in\mathrm{lfdr}^{-1}(\alpha)\pm\varepsilon\}} f^{(i)}(t)\mathrm{d}t}{\sum_{i=1}^{m} \int_{\{t\in\mathrm{lfdr}^{-1}(\alpha)\pm\varepsilon\}} f^{(i)}(t)\mathrm{d}t} \\ &= \frac{\bar{\pi}_0 \int_{\{t\in\mathrm{lfdr}^{-1}(\alpha)\pm\varepsilon\}} \bar{f}_0(t)\mathrm{d}t}{\int_{\{t\in\mathrm{lfdr}^{-1}(\alpha)\pm\varepsilon\}} \bar{f}(t)\mathrm{d}t}.\end{aligned}$$

Since $t \in \mathrm{lfdr}^{-1}(\alpha)$ implies $\bar{\pi}_0\bar{f}_0(t) = \alpha\bar{f}(t)$, the conditional probability tends to α as $\varepsilon \to 0$. □

Proof of Corollary 5.2.3. Let $g^{-1}(\alpha) \coloneqq \{t : g(t) = \alpha\}$. If g is calibrated, then

$$\begin{aligned}\alpha &= \mathbb{P}(H_J = 0 \mid g(p_J) = \alpha \text{ for some } J \in [m]) \\ &\sim \frac{\sum_{i:H_i=0} \mathbb{P}(\min_{t\in g^{-1}(\alpha)} |p_i - t| \leq \varepsilon)}{\sum_{i=1}^{m} \mathbb{P}(\min_{t\in g^{-1}(\alpha)} |p_i - t| \leq \varepsilon)},\end{aligned}$$

as $\varepsilon \to 0$, since the p-values are independent and continuously distributed. It follows that for a uniformly distributed index $I \sim \text{Uniform}\{1, \ldots, m\}$,

$$\mathbb{P}(H_I = 0 \mid g(p_I) = \alpha) = \lim_{\varepsilon\to 0} \frac{\sum_{i:H_i=0} \mathbb{P}(\min_{t\in g^{-1}(\alpha)} |p_i - t| \leq \varepsilon)}{\sum_{i=1}^{m} \mathbb{P}(\min_{t\in g^{-1}(\alpha)} |p_i - t| \leq \varepsilon)} = \alpha.$$

By Proposition 1 of Gupta et al. (2020), there exists some function h for which

$$g(t) = \mathbb{P}\left(H_I = 0 \mid h(p_I) = h(t)\right), \quad t \in [0, 1].$$

By the tower property,

$$\mathbb{P}(H_I = 0 \mid h(p_I) = h(t)) = \mathbb{P}(H_I = 0 \mid p_I \in h^{-1}(h(t)))$$

so the result follows with $A_t \coloneqq \{s \in [0, 1] : h(s) = h(t)\}$. □

Proof of Theorem 5.2.2. The conditional expectation is equal to

$$\mathrm{pFDR}([t-\varepsilon, t]) = \frac{\mathbb{E}(\mathrm{FDP}([t-\varepsilon, t]) \cdot 1_{\{p_i\in[t-\varepsilon,t] \text{ for some } i\}})}{\mathbb{P}(p_i \in [t-\varepsilon, t] \text{ for some } i)}$$

Observe that

$$\begin{aligned}\mathrm{FDP}([t-\varepsilon,t])\cdot 1_{\{p_i\in[t-\varepsilon,t] \text{ for some } i\}} &\leq 1_{\{p_i\in[t-\varepsilon,t] \text{ and } H_i=0 \text{ for some } i\}}\\ \mathrm{FDP}([t-\varepsilon,t])\cdot 1_{\{p_i\in[t-\varepsilon,t] \text{ for some } i\}} &\geq 1_{\{p_i\in[t-\varepsilon,t] \text{ and } H_i=0 \text{ for exactly one } i\}}.\end{aligned}$$

It follows from continuity of the densities that the numerator and denominator are

$$\begin{aligned}\mathbb{E}(\mathrm{FDP}([t-\varepsilon,t])\cdot 1_{\{p_i\in[t-\varepsilon,t] \text{ for some } i\}}) &\sim \varepsilon m_0 \bar{f}_0(t)\\ \mathbb{P}(p_i\in[t-\varepsilon,t] \text{ for some } i) &\sim \varepsilon m \bar{f}(t),\end{aligned}$$

so their ratio tends to $\mathrm{lfdr}(t)$ as $\varepsilon\to 0$. The same argument shows that the mFDR tends to the lfdr as $\varepsilon\to 0$ when all p-values are continuously distributed. When there is an atom at t, the mFDR tends to

$$\mathrm{mFDR}([t-\varepsilon,t]) = \frac{\mathbb{E}\left(\#\{i: H_i=0, p_i\in[t-\varepsilon,t]\}\right)}{\mathbb{E}\left(\#\{i: p_i\in[t-\varepsilon,t]\}\right)} \to \frac{m_0\bar{P}_0(\{t\})}{m\bar{P}(\{t\})} \quad \text{as } \varepsilon\to 0. \qquad \square$$

Proof of Theorem 5.3.1. Suppose without loss of generality that $H_m=0$. Then since the nulls are exchangeable, this probability is

$$\mathbb{P}(H_{(R_\alpha)}=0) = m\bar{\pi}_0\mathbb{P}(p_{(R_\alpha)}=p_m).$$

Let $q_{(1)}\leq\cdots\leq q_{(m-1)}$ denote the order statistics of $p_1,\ldots,p_{m-1}$, and note that p_m achieves the maximum in (5.16) as the $(k+1)^{\text{th}}$ order statistic if $q_{(k)}<p_m<q_{(k+1)}$ and

$$\frac{\alpha(k+1)}{m}-p_m > \left[\max_{j=k+1,\ldots,m-1}\left\{\frac{\alpha(j+1)}{m}-q_{(j)}\right\}\right]\vee\left[\max_{j=0,\ldots,k}\left\{\frac{\alpha j}{m}-q_{(j)}\right\}\right],$$

for $k\in\{0,\ldots,m-1\}$, where $q_{(0)}\coloneqq 0$. Rearranging the above inequalities gives the range in which p_m achieves the maximum as the $(k+1)^{\text{th}}$ order statistic,

$$q_{(k)}<p_m<\frac{\alpha k}{m}-\left[\max_{j=k+1,\ldots,m-1}\left\{\frac{\alpha j}{m}-q_{(j)}\right\}\right]\vee\left[\max_{j=0,\ldots,k}\left\{\frac{\alpha j}{m}-q_{(j)}\right\}-\frac{\alpha}{m}\right].$$

This range is non-empty when $\Delta_k\coloneqq\frac{\alpha k}{m}-q_{(k)}$ exceeds each of $\Delta_{k+1},\ldots,\Delta_{m-1}$ as well as $\max_{j=0,\ldots,m-1}\Delta_j-\frac{\alpha}{m}$, and has length $\Delta_k-(\max_{j=k+1,\ldots,m-1}\Delta_j)\vee(\max_{j=0,\ldots,k}\Delta_j-\frac{\alpha}{m})$. The sum of lengths of the non-empty ranges is telescoping and equal to $\frac{\alpha}{m}$, as illustrated in Figure 5.11. $\square$

Proof of BH guarantee. Letting $I\sim\mathrm{Uniform}\{1,\ldots,R\}$ where R is the number of $\mathrm{BH}(\alpha)$

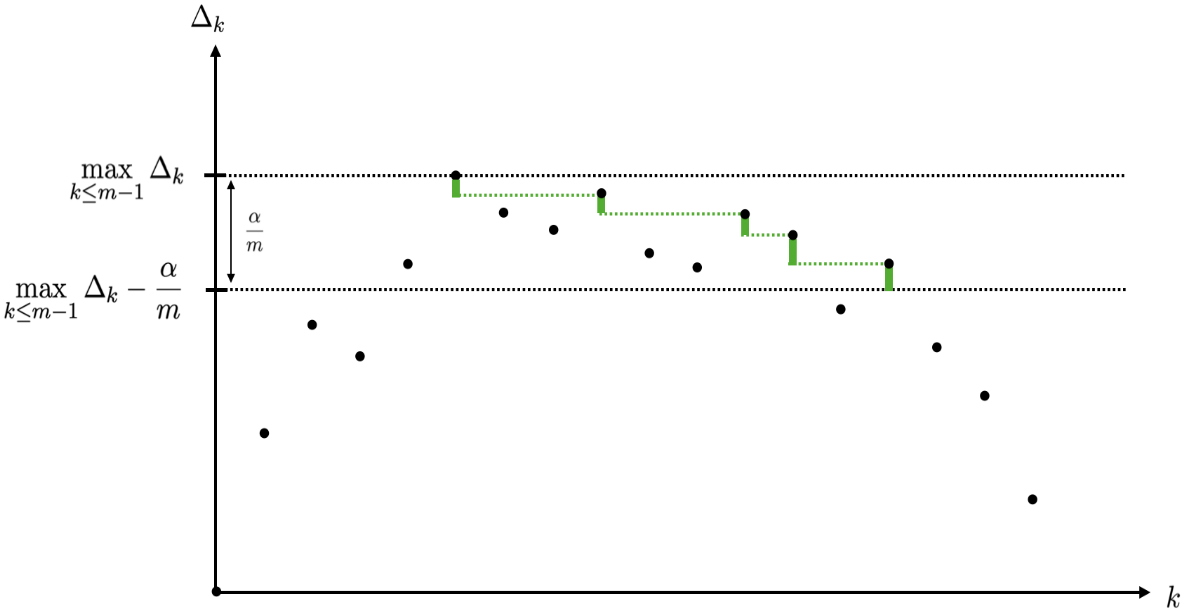


Figure 5.11: Each length of the interval range in which p_m achieves the maximum in (5.16) is indicated by a vertical green bar, and the sum of these lengths is $\frac{\alpha}{m}$.

rejections, and letting $H_m = 0$ without loss of generality, the FDR of the BH procedure is

$$\begin{aligned}
\mathbb{P}(H_{(I)} = 0) &= m_0 \mathbb{P}(p_{(I)} = p_m) \\
&= m_0 \mathbb{E}\big[\mathbb{P}(p_{(I)} = p_m \mid p_{-m})\big] \\
&= m_0 \mathbb{E}\left[\sum_{k=1}^{R} \mathbb{P}(p_{(k)} = p_m, I = k \mid p_{-m}) 1_{\{p_m \le p_{(R)}\}}\right] \\
&= m_0 \mathbb{E}\left[\frac{1}{R}\sum_{k=1}^{R} \mathbb{P}(p_{(k)} = p_m \mid p_{-m}) 1_{\{p_m \le p_{(R)}\}}\right],
\end{aligned}$$

where $p_{-m} = (p_1, \ldots, p_{m-1})$. We claim that

$$\mathbb{P}(p_m = p_{(k)} \mid p_{-m}) = q_{(k)} - q_{(k-1)} \quad \text{for } k < R,$$

where $q_{(1)} \le \cdots \le q_{(m-1)}$ are the order statistics of the entries in p_{-m}. This holds because the event $\{p_m = p_{(k)}\}$ is equivalent to $\{p_m \in (q_{(k-1)}, q_{(k)})\}$, which has probability $q_{(k)} - q_{(k-1)}$ under the assumption $p_m \sim \text{Uniform}(0,1)$ independently from p_{-m}. We also claim that

$$\mathbb{P}(p_{(R)} = p_m \mid p_{-m}) 1_{\{p_m \le p_{(R)}\}} = \left(\frac{\alpha R}{m} - q_{(R-1)}\right) 1_{\{p_m \le p_{(R)}\}},$$

since on the event $1_{\{p_m \leq p_{(R)}\}}$, the integer R is a function of p_{-m}, and

$$p_{(R)} = p_m \iff p_{(R-1)} < p_m \leq \frac{\alpha R}{m} \iff q_{(R-1)} < p_m \leq \frac{\alpha R}{m}.$$

Summing the gaps gives a telescoping sum,

$$\mathbb{P}(H_{(I)} = 0) = m_0 \mathbb{E}\left[\frac{1}{R} \cdot \left(q_{(1)} + (q_{(2)} - q_{(1)}) + \cdots + \frac{\alpha R}{m} - q_{(R-1)}\right)\right] = \frac{m_0}{m}\alpha.$$

□

Proof of Lemma 5.3.1. By the law of total probability,

$$\begin{aligned}\mathbb{P}(\text{rank}(p_m) = R) &= \sum_{\ell=1}^{L} \frac{1}{L} \cdot \mathbb{P}(\text{rank}(p_m) = R \mid p_m = \ell/L) \\ &= \sum_{\ell=1}^{L} \frac{1}{L} \cdot 1_{\{\hat{\tau}_\alpha(\ell/L) = \ell/L\}} \cdot \frac{1}{n_\ell + 1},\end{aligned}$$

where $n_\ell \coloneqq \#\{i < m : p_i = \ell/L\}$, and we have used explicit notation $\hat{\tau}_\alpha(p_m)$ to denote the threshold $\hat{\tau}_\alpha$ as a function of p_m,

$$\begin{aligned}\hat{\tau}_\alpha(p_m) &\coloneqq \underset{p_{(k)}}{\text{argmax}} \left\{\frac{\alpha k}{m} - p_{(k)}\right\} \\ &= \frac{1}{L} \cdot \underset{\ell=0,\ldots,L}{\text{argmax}} \left\{\frac{\alpha L}{m} \cdot \#\{i \leq m : p_i \leq \ell/L\} - \ell\right\},\end{aligned}$$

treating $p_1, \ldots, p_{m-1}$ as non-random elements of the grid $\{1/L, \ldots, L/L\}$. Define

$$\Delta_\ell \coloneqq \frac{\alpha L}{m} N_\ell - \ell, \quad \ell = 1, \ldots, L$$

where $N_\ell \coloneqq \#\{i < m : p_i \leq \ell/L\}$, and let $\ell^* \coloneqq \underset{\ell}{\text{argmax}}\ \Delta_\ell$. We claim that

$$\hat{\tau}_\alpha(\ell/L) = \ell/L \iff \Delta_\ell > \left[\Delta_{\ell^*} - \frac{\alpha L}{m}\right] \vee \max_{k > \ell} \Delta_k, \tag{5.26}$$

which follows from the same argument as in the alternative proof of Theorem 5.3.1 in Appendix 5.7. The claim implies

$$\begin{aligned}\mathbb{P}(\text{rank}(p_m) = R) &= \sum_{\ell=1}^{L} \frac{1}{L} \cdot 1_{\{\hat{\tau}_\alpha(\ell/L)=\ell/L\}} \cdot \frac{1}{n_\ell + 1} \\ &\leq \frac{1}{L} \sum_{\ell=1}^{L} 1_{\left\{\Delta_\ell > \left[\Delta_{\ell^*} - \frac{\alpha L}{m}\right] \vee \max_{k>\ell} \Delta_k\right\}}.\end{aligned}$$

Without loss of generality, suppose that $p_1 \leq p_2 \leq \cdots \leq p_{m-1}$. If $p_i < p_j$ are among these p-values and both $\Delta_{p_i L}, \Delta_{p_j L}$ satisfy the rhs of (5.26), then as $L \to \infty$, it follows from the definition of Δ_ℓ that as L grows,

$$\Delta_\ell - \Delta_{\ell+1} = 1 \quad \text{for all but a fixed number of } \ell \in (Lp_i, Lp_j).$$

Thus for the distinct values $\ell^* =: \ell_1 < \ell_2 < \cdots < \ell_k$ for which the rhs of (5.26) holds, we must have $\Delta_{\ell_i} - \Delta_{\ell_{i+1}} = 1$ for all but a fixed number of winners ℓ_i, the number of which is less than m. Therefore

$$\begin{aligned}\mathbb{P}(\text{rank}(p_m) = R) &\leq \frac{1}{L} \sum_{\ell=1}^{L} 1_{\left\{\Delta_\ell > \left[\Delta_{\ell^*} - \frac{\alpha L}{m}\right] \vee \max_{k>\ell} \Delta_k\right\}} \\ &\leq \frac{1}{L} \left(m + \sum_{i=1}^{k-1} (\Delta_i - \Delta_{i+1}) + \Delta_k - \Delta_{\ell^*} + \frac{\alpha L}{m} \right) \\ &= \frac{1}{L} \left(m + \frac{\alpha L}{m} \right) = \frac{\alpha}{m} + O\left(\frac{m}{L}\right),\end{aligned}$$

as $L \to \infty$. □

Proof of Theorem 5.3.3. By symmetry of the nulls,

$$\mathbb{P}(H_{(R_\alpha)} = 0) = m_0 \mathbb{P}(\text{rank}(p_m) = R).$$

By the argument in Theorem 5.3.2, the probability that p_m is the R^{th} smallest p-value (breaking ties uniformly at random) is

$$\mathbb{P}(\text{rank}(p_m) = R) = \mathbb{E}\left(\frac{1}{L} \sum_{\ell=1}^{L} 1_{\left\{\Delta_\ell > \left[\Delta_{\ell^*_m} - \alpha L/m\right] \vee \max_{j>\ell} \Delta_j\right\}} \cdot \frac{1}{n_\ell + 1} \right),$$

where $n_\ell := \#\{i < m : p_i = \ell/L\}$ and Δ_ℓ are defined

$$\Delta_\ell := \frac{\alpha L}{m} N_\ell - \ell, \quad \ell = 0, \ldots, L,$$

with $N_\ell \coloneqq \sum_{k=1}^{\ell} n_k$ and $\ell^*_m \coloneqq \operatorname*{argmax}_{\ell=0,\ldots,L} \Delta_\ell$. As $m \to \infty$, we have the following convergence in probability,

$$\frac{n_\ell}{mf^*(\ell/L)} \xrightarrow{p} 1, \quad \Delta_\ell \xrightarrow{p} \alpha L \sum_{k=0}^{\ell} f^*(k/L) - \ell.$$

Since the maximizer of $\alpha L \sum_{k=0}^{\ell} f^*(k/L) - \ell$ is unique, we have $\ell^*_m = \ell^*$ with high probability as $m \to \infty$, and

$$mf^*(\ell^*/L) \cdot \sum_{\ell=1}^{L} 1_{\left\{\Delta_\ell > \left[\Delta_{\ell^*_m} - \alpha L/m\right] \vee \max_{j>\ell} \Delta_j\right\}} \cdot \frac{1}{n_\ell + 1} \xrightarrow{\mathbb{P}} 1_{\{\ell^* > 0\}},$$

from which it follows that

$$\mathbb{E}\left[\frac{m_0}{mf^*(\ell^*/L)} \cdot \frac{L}{L} \cdot mf^*(\ell^*/L) \cdot \sum_{\ell=1}^{L} 1_{\left\{\Delta_\ell > \left[\Delta_{\ell^*_m} - \alpha L/m\right] \vee \max_{j>\ell} \Delta_j\right\}} \cdot \frac{1}{n_\ell + 1}\right] \to \frac{\pi_0^*}{Lf^*(\ell^*/L)},$$

when $\ell^* > 0$, and zero otherwise. □

Proof of Theorem 5.4.1. Lemma 5.7.5 implies that for m large enough, we have with probability $\geq 1 - \delta$ that

$$\hat{\tau}_\alpha - \tau^*_\alpha \leq \varepsilon \coloneqq C' m^{-1/3} \log(2m/\delta), \tag{5.27}$$

for some constant $C' > 0$ depending on α, J and δ. Since f is decreasing on $(0, \tau^*_\alpha + \alpha)$ and has derivative greater than $-J^{-1}$ over the interval $(\tau^*_\alpha, \tau^*_\alpha + \varepsilon)$, the above inequality implies

$$f(\hat{\tau}_\alpha) \geq f(\tau^*_\alpha) - J^{-1}\varepsilon = \alpha^{-1} - J^{-1}\varepsilon.$$

It follows that

$$\operatorname{lfdr}(\hat{\tau}_\alpha) = \frac{\pi_0}{f(\hat{\tau}_\alpha)} \leq \pi_0 \alpha + C m^{-1/3} \log(m/\delta),$$

for another constant $C > 0$ depending on α, J and δ. The other direction follows similarly from Lemma 5.7.6. □

Proof of Proposition 5.4.1. Put $q \coloneqq (p_{(1)}, \ldots, p_{(m)})$ and $H^q \coloneqq (H_{(1)}, \ldots, H_{(m)})$, and let

$$\sigma \sim \operatorname{Uniform}(\mathcal{S}_m), \quad \widetilde{H}^q \coloneqq H^q_\sigma, \quad \widetilde{q} \coloneqq q_\sigma.$$

Since $(\widetilde{H}^q, \widetilde{q})$ is equal in distribution to $(\widetilde{H}, \widetilde{p})$,

$$\mathrm{clfdr}_i(\widetilde{q}) = \mathbb{P}(\widetilde{H}_i^q = 0 \mid \widetilde{q}) = \mathbb{P}(H_{\sigma^{-1}(i)}^q = 0 \mid q). \tag{5.28}$$

By exchangeability of the pairs $(\widetilde{H}_i, \widetilde{p}_i)_{i=1}^m$, we have

$$\mathrm{clfdr}(u_\pi) = \mathrm{clfdr}_\pi(u), \quad \text{for any permutation } \pi \in \mathcal{S}_m \text{ and } u \in [0,1]^m.$$

On the event $\{p_{(K)} = p_i\}$, this property implies that

$$\begin{aligned}
\mathrm{clfdr}_i(p) &= \mathrm{clfdr}_K(q) \\
&= \mathrm{clfdr}_{\sigma(K)}(\widetilde{q}) \\
&= \mathbb{P}(\widetilde{H}_{\sigma(K)}^q = 0 \mid \widetilde{q}) \\
&= \mathbb{P}(H_K^q = 0 \mid q) && \text{(by (5.28))} \\
&= \mathbb{P}(H_{(K)} = 0 \mid p_{(1)}, \dots, p_{(m)}).
\end{aligned}$$

□

Proof of Theorem 5.4.2. Supposing without loss of generality that $H_1 = 0$ and $H_2 = 1$,

$$\mathrm{clfdr}_i(p) = \frac{\bar{\pi}_0 f_0(p_i)}{\bar{\pi}_0 f_0(p_i) + \bar{\pi}_1 f_1(p_i) \cdot X_i},$$

where X_i is a likelihood ratio,

$$X_i := \frac{\sum_{\sigma \in \mathcal{S}_m : \sigma(i)=2} \prod_{j \in [m] \setminus \{i\}}^m f^{(\sigma(j))}(p_j)}{\sum_{\sigma \in \mathcal{S}_m : \sigma(i)=1} \prod_{j \in [m] \setminus \{i\}}^m f^{(\sigma(j))}(p_j)}$$

for testing between the following two hypotheses:

$$\begin{aligned}
&\mathrm{Hyp}_0 : \text{We observe a random permutation of } p_{-i} \text{ when } H_i = 1 \\
&\mathrm{Hyp}_1 : \text{We observe a random permutation of } p_{-i} \text{ when } H_i = 0.
\end{aligned}$$

A simpler testing problem is:

$$\begin{aligned}
&\widetilde{\mathrm{Hyp}}_0 : \widetilde{p}_1, \dots, \widetilde{p}_{m_0} \overset{\text{iid}}{\sim} f_0, \text{ and } (\widetilde{p}_{m_0+1}, \dots, \widetilde{p}_{m-1}) \overset{\text{iid}}{\sim} f_1 \\
&\widetilde{\mathrm{Hyp}}_1 : \frac{1}{m_0} \sum_{\ell=1}^{m_0} \left[(\widetilde{p}_{1:m_0})_{-\ell} \overset{\text{iid}}{\sim} f_0, \text{ and } (\widetilde{p}_\ell, \widetilde{p}_{m_0+1}, \dots, \widetilde{p}_{m-1}) \overset{\text{iid}}{\sim} f_1 \right],
\end{aligned}$$

since $\mathrm{Hyp}_0, \mathrm{Hyp}_1$ can be obtained from $\widetilde{\mathrm{Hyp}}_0, \widetilde{\mathrm{Hyp}}_0$ by adding a random permutation. If $H_i = 0$ (resp. $H_i = 1$), then the distribution of X_i is as if the data were generated by Hyp_1

(resp. Hyp_0). The likelihood ratio of $\widetilde{\text{Hyp}_1}$ to $\widetilde{\text{Hyp}_0}$ has variance

$$\text{Var}_0\left(\frac{1}{m_0}\sum_{\ell=1}^{m_0}\frac{f_1}{f_0}(\widetilde{p}_\ell)\right) = \frac{1}{m_0}\text{Var}_0\left(\frac{f_1}{f_0}(p_1)\right) \to 0$$

by assumption, where Var_0 denotes the variance operation when $\widetilde{\text{Hyp}_0}$ holds. It follows from Lemma 5.7.1 that

$$\mathbb{E}_{H_i=1}(X_i-1)^2 \leq \widetilde{\mathbb{E}}_0\left(\frac{1}{m_0}\sum_{\ell=1}^{m_0}\frac{f_1}{f_0}(\widetilde{p}_\ell)-1\right)^2 \to 0.$$

A symmetric argument yields

$$\mathbb{E}_{H_i=0}(X_i-1)^2 \leq \widetilde{\mathbb{E}}_1\left(\frac{1}{m_1}\sum_{\ell=m_0}^{m-1}\frac{f_0}{f_1}(\widetilde{p}_\ell)-1\right)^2 \to 0,$$

under the condition that $\text{Var}\left(\frac{f_0}{f_1}(p_2)\right)$ when $H_2 = 1$. Here we are abusing notation by writing the index ℓ from m_0 to $m-1$, to denote summing over the $m_1 - 1$ many p-values drawn from f_1 in the scenario described by $\widetilde{\text{Hyp}_1}$. It now follows from Chebyshev's inequality that

$$\mathbb{P}(|X_i - 1| > \varepsilon) \leq \frac{1}{\varepsilon^2}\cdot\frac{\text{Var}\left(\frac{f_0}{f_1}(p_2)\right) \vee \text{Var}\left(\frac{f_1}{f_0}(p_1)\right)}{m_0 \wedge m_1} \to 0,$$

as $m \to \infty$, since $0 < \lim_{m\to\infty}\frac{m_0}{m} < 1$. □

Lemma 5.7.1 (Greenshtein and Ritov (2009)). *Consider two pairs of distributions, $\{G_0, G_1\}$ and $\{\widetilde{G}_0, \widetilde{G}_1\}$, such that the first pair represents a weaker experiment in the sense that there is a Markov kernel $\mathbb{K}$, and $G_i(\cdot) = \int \mathbb{K}(y,\cdot)\mathrm{d}\widetilde{G}_i(y), i = 1, 2$. Then*

$$\mathbb{E}_{G_0}\psi\left(\frac{\mathrm{d}G_1}{\mathrm{d}G_0}\right) \leq \mathbb{E}_{\widetilde{G}_0}\psi\left(\frac{\mathrm{d}\widetilde{G}_1}{\mathrm{d}\widetilde{G}_0}\right),$$

for any convex function ψ.

5.7.1 bFDR of SL under relaxed null assumptions

The result in Theorem 5.3.1 holds under the relaxed condition on each null density,

$$H_i = 0 \Rightarrow f^{(i)}(t) \leq 1 \quad \text{for all } t \in [0, \alpha].$$

This holds for p-values computed in the Gaussian one-sided location testing problem, but holds more generally for one-sided testing of a hypothesis $\theta \leq \theta_0$ within an exponential family. Without loss of generality we assume $\theta_0 = 0$ in the following result, from which Proposition 5.3.1 follows by recentering the parameter $\theta' = \theta - \theta_0$.

Lemma 5.7.2. *For an exponential family (g_θ) of continuous distributions with densities given by*

$$g_\theta(z) = \exp(\theta z - A(\theta))g_0(z),$$

with cdfs (G_θ), the condition $G_0^{-1}(1-\alpha) > \mathbb{E}_0 Z$ implies the density of the one-sided p-value $p := 1 - G_0(Z)$ is bounded by 1 over the interval $[0, \alpha]$.

Proof. When $Z \sim g_\theta$, the density of $p = 1 - F_0(Z)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}_\theta(p \leq t) = \frac{g_\theta}{g_0}(G_0^{-1}(1-t)).$$

At $\theta = 0$, the above ratio is equal to 1. When $\theta \leq 0$, the log density has a positive derivative in θ when

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\log \frac{g_\theta}{g_0}(G_0^{-1}(1-t))\right] = G_0^{-1}(1-t) - \mathbb{E}_\theta(Z) > 0$$

which is true for all $t \leq \alpha$ when $G_0^{-1}(1-\alpha) > \mathbb{E}_0(Z)$. □

Counterexample to the super-uniform assumption

The SL guarantee may break down when the nulls are super-uniformly distributed instead of Uniform$(0, 1)$ distributed, as illustrated by the following example.

Counterexample. Let $m = 2$, $H_1 = 0$, and $H_2 = 1$, with $p_1 \sim f^{(1)}$ defined

$$f^{(1)}(t) := \begin{cases} \frac{1}{2} & 0 \leq t \leq \frac{1}{4} \\ \frac{3}{2} & \frac{1}{4} < t \leq \frac{1}{2} \\ 1 & \frac{1}{2} < t \leq 1, \end{cases}$$

and $p_2 \equiv \frac{1}{4}$. It is straightforward to verify that $f^{(1)}$ is super-uniform, i.e.

$$\int_0^t f^{(1)}(s)\mathrm{d}s \leq t \quad \text{for any } t \in [0, 1].$$

The event $p_{(R_\alpha)} = p_1$ is equivalent to

$$\left(\{p_1 \leq p_2\} \cap \left\{p_1 - \frac{\alpha}{2} < (p_2 - \alpha) \wedge 0\right\}\right) \bigcup \left(\{p_1 > p_2\} \cap \left\{p_1 - \alpha < \left(p_2 - \frac{\alpha}{2}\right) \wedge 0\right\}\right).$$

Plugging in $\alpha = \frac{1}{2}$ and $p_2 = \frac{1}{4}$ gives

$$H_{(R_\alpha)} = 0 \iff p_{(R_\alpha)} = p_1 \iff p_1 \in (1/4, 1/2)\,,$$

which occurs with probability $\frac{3}{2} \times \frac{1}{4} = \frac{3}{8} > \frac{\alpha}{2} = \frac{1}{4}$. □

Counterexample to the discrete-uniform assumption

The SL guarantee breaks down in the setting where each null p-value is uniformly distributed on the grid $\left\{\frac{1}{L}, \frac{2}{L}, \ldots, \frac{L-1}{L}, 1\right\}$.

Counterexample. Let $m = 6, L = 9, \alpha = 1/2$, and the alternative p-values are

$$\begin{aligned} p_1 = p_2 &= 1/L \\ p_i &= i/L \quad i = 2, 3, 4. \end{aligned}$$

Then the probability that $p_m \sim$ Uniform $\left\{\frac{1}{L}, \frac{2}{L}, \ldots, \frac{L-1}{L}, 1\right\}$ is the last SL(α) rejection is

$$\begin{aligned} \mathbb{P}(p_{(R_\alpha)} = p_m) &= \sum_{\ell=1}^{L} \frac{1}{L} \mathbb{P}(p_{(R_\alpha)} = p_m \mid p_m = \ell/L) \\ &= \sum_{\ell=1}^{L} \frac{1}{L} \cdot 1_{\{p_{(R_\alpha)} = \ell/L\}} \cdot \frac{1}{n_\ell + 1}, \end{aligned}$$

where $n_\ell := \#\{i < m : p_i = \ell/L\}$ for $\ell = 1, \ldots, L$. It is straightforward to check that $1_{\{p_{(R_\alpha)} = \ell/L\}} = 1$ for $\ell = 1, 2, 3, 4$, so that the above evaluates to

$$\mathbb{P}(p_{(R_\alpha)} = p_m) = \frac{1}{9}\left(\frac{1}{3} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right) = 0.204 > 0.167 = \frac{2\alpha}{m}.$$

□

Numerical evidence suggests that the boundary FDR of the SL set can be violated above α even as the grid size $1/L \to 0$. A counterexample is illustrated in Figure 5.12, where $m, L \to \infty$ with $L/m = 1.8$, and the alternative p-values are fixed at the left end of the grid.

5.7.2 Technical Lemmas

Lemma 5.7.3. *Suppose $p_i \sim f^{(i)}$ independently for $i = 1, \ldots, m$, and each density $f^{(i)}$ is non-increasing. Let R_i denote the rank of p_i among $p_1, \ldots, p_m$, where $R_i = 1$ when $p_i = \min_{j=1,\ldots,m} p_j$. Then for any $p_i \sim$ Uniform$(0, 1)$, the conditional pmf of R_i is non-*

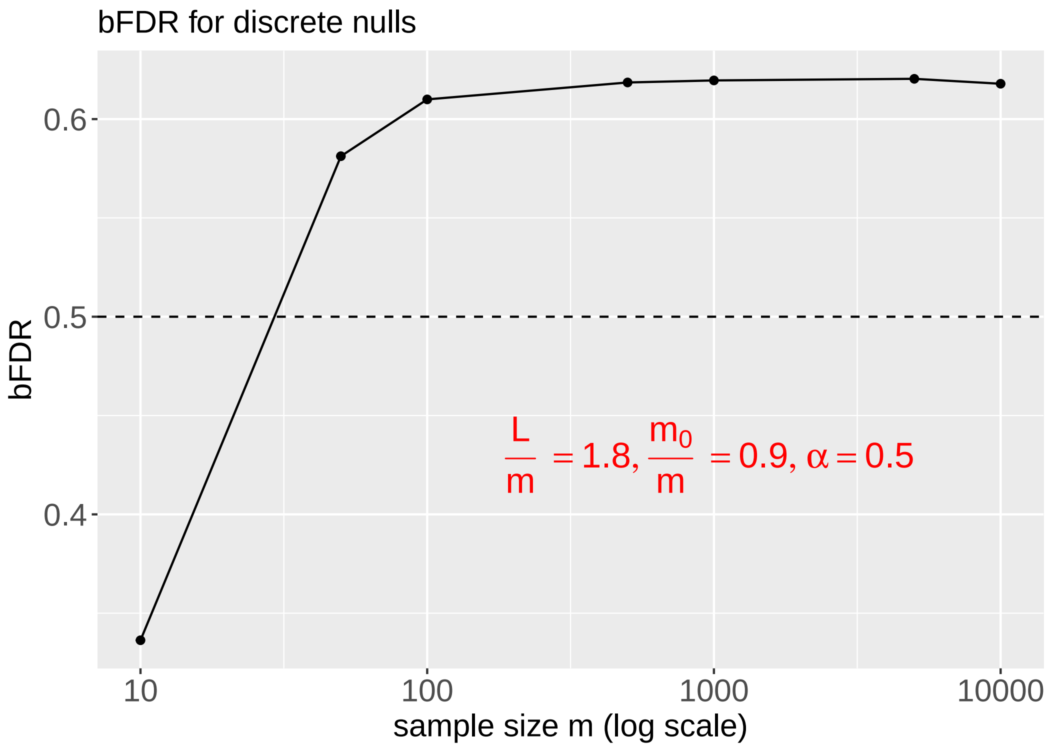


Figure 5.12: The simulation setting is $\bar{\pi}_0 = 0.9$, $\alpha = 0.5$, and $L = 1.8m$, and the non-nulls are fixed along the grid $\left\{\frac{1}{L}, \frac{2}{L}, \ldots, \frac{m_1}{L}\right\}$. The sample size ranges from $m \in \{10, 50, 100, 500, 1000, 5000, 10^4\}$. The bFDR of the SL procedure is estimated using $N = 10^5$ Monte Carlo samples.

decreasing,

$$\mathbb{P}(R_i = j \mid P_m) \leq \mathbb{P}(R_i = j + 1 \mid P_m), \quad j = 1, \ldots, m - 1$$

where P_m is the empirical distribution of $p_1, \ldots, p_m$.

Proof. For $j \leq m$, let $[j] \in \{1, \ldots, m\}$ denote the anti-rank of j, i.e. the index for which $R_{[j]} = j$, or equivalently, $p_{[j]} = p_{(j)}$. By the tower property, it suffices to show that for any $j \leq m - 1$,

$$\mathbb{P}(R_i = j \mid \mathcal{G}_j) \leq \mathbb{P}(R_i = j + 1 \mid \mathcal{G}_j), \tag{5.29}$$

where $\mathcal{G}_j$ denotes the sigma field generated by the order statistics and all anti-ranks besides $[j], [j + 1]$,

$$\mathcal{G}_j := \sigma(F_m, \{[1], \ldots, [m]\} \backslash \{[j], [j + 1]\}).$$

If $i \notin \{[j], [j + 1]\}$, then both sides of (5.29) are equal to zero and the inequality is satisfied.

Now suppose $\{[j], [j+1]\} = \{i, k\}$ for some $1 \leq k \leq m$. Then

$$\mathbb{P}(R_i = j \mid \mathcal{G}_j) \propto f^{(i)}(p_{(j)}) f^{(k)}(p_{(j+1)}) \prod_{\ell \notin \{i,k\}} f^{(\ell)}(p_\ell)$$
$$\mathbb{P}(R_i = j+1 \mid \mathcal{G}_j) \propto f^{(i)}(p_{(j+1)}) f^{(k)}(p_{(j)}) \prod_{\ell \notin \{i,k\}} f^{(\ell)}(p_\ell),$$

where the constant of proportionality is the same in both cases,

$$C \coloneqq \frac{1}{\Big(f^{(i)}(p_{(j)}) f^{(k)}(p_{(j+1)}) + f^{(i)}(p_{(j+1)}) f^{(k)}(p_{(j)})\Big) \prod_{\ell \notin \{i,k\}} f^{(\ell)}(p_\ell)}.$$

Since $f^{(i)} = 1_{[0,1]}$, the ratio of conditional probabilities is

$$\frac{\mathbb{P}(R_i = j \mid \mathcal{G}_j)}{\mathbb{P}(R_i = j+1 \mid \mathcal{G}_j)} = \frac{f^{(k)}(p_{(j+1)})}{f^{(k)}(p_{(j)})} \leq 1,$$

since $f^{(k)}$ is non-increasing. Multiplying both sides of the above inequality by $\mathbb{P}(R_i = j+1 \mid \mathcal{G}_j)$ and taking expectation with respect to the conditional distribution given F_m completes the proof. □

Lemma 5.7.4. *Let τ_α^* be a solution to $f(\tau_\alpha^*) = \alpha^{-1}$ and $\hat{\tau}_\alpha$ is the rejection threshold of the SL(α) procedure (5.16). If $\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon$, then there exists an index $k \geq 1$ for which*

$$p_{(i^*+k)} \leq \tau_\alpha^* + \frac{\alpha k}{m} \quad \text{and} \quad k > \frac{m\varepsilon}{\alpha},$$

where $i^* \coloneqq \max\{i : p_{(i)} \leq \tau_\alpha^*\}$ and $i^* = 0$ if no such i exists.

Proof. Let $\hat{k}$ be the index for which $\hat{\tau}_\alpha = p_{(i^*+\hat{k})}$. The first inequality can be written

$$\frac{i^* + \hat{k}}{m} - \frac{i^*}{m} - \alpha^{-1}(p_{(i^*+\hat{k})} - \tau_\alpha^*) \geq 0,$$

which holds because $F_m(t) - \alpha^{-1}F_0(t)$ is maximized at $t = p_{(i^*+\hat{k})}$. Since $\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon$, the above inequality implies $\hat{k} > \frac{m\varepsilon}{\alpha}$. □

Lemma 5.7.5. *Let τ_α^* and $\hat{\tau}_\alpha$ be defined as in Lemma 5.7.4, let $\delta > 0$ and suppose $\bar{f}$ is decreasing on $[\tau_\alpha^*, \tau_\alpha^* + \alpha]$ and that there exists some $J > 0$ for which $J \leq |\bar{f}'(t)| \leq J^{-1}$ for all t with $|t - \tau_\alpha^*| \leq \varepsilon$, where $\varepsilon \coloneqq \left(\frac{24}{\alpha L^2}\right)^{1/3} m^{-1/3} \log(2m/\delta)$. Then*[8]

$$\mathbb{P}(\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon) \leq \delta,$$

[8] In Appendix 5.7.3, we show a simulation in which the average density is decreasing $(0, \tau_\alpha + \alpha)$ (but not

for any $m \geq C(\alpha, J, \delta)$, a constant depending only on α, J and δ.

Proof of Lemma 5.7.5. Applying Lemma 5.7.4 with ε defined as above, we have

$$\begin{aligned}\mathbb{P}(\hat{\tau}_\alpha > \tau^*_\alpha + \varepsilon) &\leq \sum_{k > \frac{m\varepsilon}{\alpha}} \mathbb{P}\left(p_{(i^*+k)} \leq \tau^*_\alpha + \frac{\alpha k}{m}\right) \\ &= \sum_{\frac{m\varepsilon}{\alpha} < k \leq \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}\left(N_k \geq k\right) + \sum_{k > \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}\left(N_k \geq k\right), \qquad (5.30)\end{aligned}$$

where i^* is defined in Lemma 5.7.4, and N_k is the number of p-values between τ^*_α and $\tau^*_\alpha + \frac{\alpha k}{m}$, distributed Generalized-Binomial with sample size m and average success probability $\bar{F}(\tau^*_\alpha + \alpha k/m) - \bar{F}(\tau^*_\alpha)$,

$$N_k = \sum_{j=1}^m 1_{\{p_j \in (\tau^*_\alpha, \tau^*_\alpha + \alpha k/m)\}} \Rightarrow \mathbb{E}N_k = m(\bar{F}(\tau^*_\alpha + \alpha k/m) - \bar{F}(\tau^*_\alpha)),$$

where $\bar{F} := \frac{1}{m}\sum_{i=1}^m F^{(i)}$ is the average cdf of the p-values. Note that since $\bar{F}' = \bar{f}$, we have by the mean value theorem that

$$\mathbb{E}N_k = m(\bar{F}(\tau^*_\alpha + \alpha k/m) - \bar{F}(\tau^*_\alpha)) = m\bar{f}(\xi) \cdot \frac{\alpha k}{m},$$

for some $\xi \in (\tau^*_\alpha, \tau^*_\alpha + \alpha k/m)$. By the monotonicity assumption, $\bar{f}(\xi) \leq \bar{f}(\tau^*_\alpha) = \alpha^{-1}$ implies we have $\mathbb{E}N_k \leq k$. Consider the corresponding Binomial random variable, $\widetilde{N}_k \sim \text{Binomial}(m, \bar{F}(\tau^*_\alpha + \alpha k/m) - \bar{F}(\tau^*_\alpha))$. Since $\mathbb{E}\widetilde{N}_k = \mathbb{E}N_k \leq k$, it follows from Theorem 5 in Hoeffding (1956) that

$$\mathbb{P}\left(N_k \geq k\right) \leq \mathbb{P}\left(\widetilde{N}_k \geq k\right) = \mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right).$$

To bound the probability on the right hand side, we use the following bounds on the expectation $\mathbb{E}\widetilde{N}_k = m(\bar{F}(\tau^*_\alpha + \alpha k/m) - \bar{F}(\tau^*_\alpha))$,

$$\mathbb{E}\widetilde{N}_k \leq k + \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon \qquad (5.31)$$

$$\mathbb{E}\widetilde{N}_k \geq \frac{m\varepsilon}{2\alpha}. \qquad (5.32)$$

Before proving inequalities (5.31) and (5.32), we show how they can be used to complete the proof. When $\frac{m\varepsilon}{\alpha} < k \leq \frac{m\varepsilon \log m}{\alpha}$, the upper bound (5.31) gives $\mathbb{E}\widetilde{N}_k \leq k + \frac{Jm\varepsilon^2}{2} - J\alpha\varepsilon \cdot \frac{m\varepsilon}{\alpha}$,

on the entire unit interval) where the estimate $\hat{\tau}_\alpha$ tends to τ^*_α at the $m^{-1/3}$ rate. Namely, the assumption that $\bar{f}$ be decreasing over the entire unit interval is not necessary for the threshold based on the Grenander estimator to be accurate in large samples.

which implies

$$\mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right) \leq \mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{k - \frac{Jm\varepsilon^2}{2}}\right).$$

Now since $\frac{1}{1-x} \geq 1 + x$, the rhs of the above is

$$\leq \mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \left(1 + \frac{Jm\varepsilon^2}{2k}\right)\right) \leq \exp\left(-\frac{1}{3} \cdot \mathbb{E}\widetilde{N}_k \cdot \left(\frac{Jm\varepsilon^2}{2k}\right)^2\right),$$

where the last inequality follows from a Binomial tail bound, recorded in Lemma 5.7.7. Now using $k \leq \frac{m\varepsilon \log m}{\alpha}$ and applying the lower bound (5.32), we obtain

$$\leq \exp\left(-\frac{1}{3} \cdot \frac{m\varepsilon}{2\alpha} \cdot \left(\frac{J\alpha\varepsilon}{2\log m}\right)^2\right).$$

Simplifying, we have shown that when $\frac{m\varepsilon}{\alpha} < k \leq \frac{m\varepsilon \log m}{\alpha}$,

$$\mathbb{P}\left(\widetilde{N}_k \geq k\right) \leq \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{24\log^2 m}\right).$$

Plugging in the formula for ε, the above inequality implies that the first piece of (5.30) is bounded,

$$\sum_{\frac{m\varepsilon}{\alpha} < k \leq \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}(N_k \geq k) \leq m \exp\left(-\log(2m/\delta)\right) = \delta/2. \tag{5.33}$$

When $k > \frac{m\varepsilon \log m}{\alpha}$, the upper bound (5.31) gives

$$\mathbb{E}\widetilde{N}_k \leq k + \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon = k\left(1 + \frac{Jm\varepsilon^2}{2k} - J\alpha\varepsilon\right) \leq k\left(1 - \frac{J\alpha\varepsilon}{2}\right),$$

for m large enough, since $\frac{Jm\varepsilon^2}{2k} \leq O\left(\frac{\varepsilon}{\log m}\right)$. Again using $\frac{1}{1-x} \geq 1 + x$, this upper bound on $\mathbb{E}\widetilde{N}_k$ implies

$$\begin{aligned}
\mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right) &\leq \mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{k\left(1-\frac{J\alpha\varepsilon}{2}\right)}\right) \\
&\leq \mathbb{P}\left(\widetilde{N}_k \geq \mathbb{E}\widetilde{N}_k \cdot \left(1+\frac{J\alpha\varepsilon}{2}\right)\right) \\
&\leq \exp\left(-\frac{1}{3}\cdot \mathbb{E}\widetilde{N}_k \cdot \left(\frac{J\alpha\varepsilon}{2}\right)^2\right) && \text{(by Lemma 5.7.7)} \\
&\leq \exp\left(-\frac{1}{3}\cdot \frac{m\varepsilon}{2\alpha} \cdot \left(\frac{J\alpha\varepsilon}{2}\right)^2\right) && \text{(by (5.32))} \\
&= \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{24}\right).
\end{aligned}$$

Since $\delta \leq 1$, the above implies that the second piece of (5.30) is bounded,

$$\sum_{k > \frac{m\varepsilon \log m}{\alpha}} \mathbb{P}(N_k \geq k) \leq m \exp\left(-\log^3(2m/\delta)\right) \leq \delta/2.$$

Together with (5.33), we have shown

$$\mathbb{P}(\hat{\tau}_\alpha > \tau_\alpha^* + \varepsilon) \leq \sum_{k > \frac{m\varepsilon}{\alpha}} \mathbb{P}(N_k \geq k) \leq \delta.$$

It remains to verify (5.31) and (5.32). To show (5.31), note that for any $t \in [\tau_\alpha^*, \tau_\alpha^* + \varepsilon]$, the mean value theorem gives

$$\bar{f}(t) - \bar{f}(\tau_\alpha^*) \leq -J(t - \tau_\alpha^*)$$

since $\bar{f}' \leq -J$ on $[\tau_\alpha^*, \tau_\alpha^* + \varepsilon]$. Since $\bar{f}$ is decreasing on $[\tau_\alpha^*, \tau_\alpha^* + \alpha]$, this implies

$$\bar{f}(t) \leq \begin{cases} \bar{f}(\tau_\alpha^*) - J(t - \tau_\alpha^*) & \tau_\alpha^* \leq t \leq \tau_\alpha^* + \varepsilon \\ \bar{f}(\tau_\alpha^*) - J\varepsilon & \tau_\alpha^* + \varepsilon < t \leq \tau_\alpha^* + \alpha. \end{cases}$$

Thus the expectation can be bounded,

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &= m\int_{\tau_\alpha^*}^{\tau_\alpha^*+\frac{\alpha k}{m}} \bar{f}(t)\mathrm{d}t \\
&= m\int_{\tau_\alpha^*}^{\tau_\alpha^*+\varepsilon} \bar{f}(t)\mathrm{d}t + m\int_{\tau_\alpha^*+\varepsilon}^{\tau_\alpha^*+\frac{\alpha k}{m}} \bar{f}(t)\mathrm{d}t \\
&\le m\int_{\tau_\alpha^*}^{\tau_\alpha^*+\varepsilon} \bar{f}(\tau_\alpha^*) - J(t-\tau_\alpha^*)\mathrm{d}t + m\int_{\tau_\alpha^*+\varepsilon}^{\tau_\alpha^*+\frac{\alpha k}{m}} \bar{f}(\tau_\alpha^*) - J\varepsilon\mathrm{d}t \\
&= m\left[\bar{f}(\tau_\alpha^*)\cdot\frac{\alpha k}{m} - \frac{J(t-\tau_\alpha^*)^2}{2}\Big|_{\tau_\alpha^*}^{\tau_\alpha^*+\varepsilon} - J\varepsilon\left(\frac{\alpha k}{m}-\varepsilon\right)\right] \\
&= k - \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon + Jm\varepsilon^2 = k + \frac{Jm\varepsilon^2}{2} - J\alpha k\varepsilon,
\end{aligned}$$

which shows (5.31). For (5.32), note that the mean value theorem and the condition $\bar{f}' \ge -J^{-1}$ on $[\tau_\alpha^*, \tau_\alpha^*+\varepsilon]$ imply that $\bar{f}(t) \ge \bar{f}(\tau_\alpha^*) - J^{-1}(t-\tau_\alpha^*)$ for any $t \in [\tau_\alpha^*, \tau_\alpha^*+\varepsilon]$. Thus we have

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &= m\int_{\tau_\alpha^*}^{\tau_\alpha^*+\frac{\alpha k}{m}} \bar{f}(t)\mathrm{d}t \\
&\ge m\int_{\tau_\alpha^*}^{\tau_\alpha^*+\varepsilon} \left(\bar{f}(\tau_\alpha^*) - J^{-1}(t-\tau_\alpha^*)\right)\mathrm{d}t \\
&= m\varepsilon\bar{f}(\tau_\alpha^*) - \frac{m\varepsilon^2}{2J} \\
&= \frac{m\varepsilon}{\alpha} - \frac{m\varepsilon^2}{2J} \ge \frac{m\varepsilon}{2\alpha},
\end{aligned}$$

since for m larger than some constant $C(\alpha, J, \delta) > 0$, we have $\frac{m}{\log^3(2m/\delta)} \ge 24\alpha^2/J^5$, which is equivalent to the last inequality above. □

A high probability lower bound can be shown under an extended monotonicity constraint of f over the interval $(0, \tau_\alpha^*)$, as described in the next lemma.

Lemma 5.7.6. *Let $\delta > 0$. Suppose f is decreasing on the interval $(0, \tau_\alpha^*)$ and that there exists some $J > 0$ for which $|f'(t)| \ge J$ for all t with $|t-\tau_\alpha^*| \le \varepsilon$, where $\varepsilon := \left(\frac{48}{\alpha J^2}\right)^{1/3} m^{-1/3}\log(2m/\delta)$. Then*

$$\mathbb{P}(\hat{\tau}_\alpha < \tau_\alpha^* - \varepsilon) \le \delta,$$

for any $m \ge C(\alpha, J, \delta)$, a constant depending only on α, J and δ.

Proof. Define i^* as in Lemma 5.7.4. If $\hat{\tau}_\alpha < \tau_\alpha^* - \varepsilon$, then there exists some $0 \le k \le i^*$ for which $\hat{\tau}_\alpha = p_{(i^*-k)}$ and thus

$$p_{(i^*-k)} - \frac{\alpha(i^*-k)}{m} \le p_{(i^*)} - \frac{\alpha i^*}{m} \quad \text{and} \quad p_{(i^*-k)} < \tau_\alpha^* - \varepsilon.$$

Since $p_{(i^*)} \le \tau_\alpha^*$, it follows that the probability can be bounded,

$$\begin{aligned}
\mathbb{P}(\hat{\tau}_\alpha < \tau_\alpha^* - \varepsilon) &\le \mathbb{P}\left(\bigcup_{k=0}^{m} \left\{p_{(i^*-k)} \le \left(\tau_\alpha^* - \frac{\alpha k}{m}\right) \wedge (\tau_\alpha^* - \varepsilon)\right\} \cap \{i^* \ge k\}\right) \\
&\le \mathbb{P}\left(\bigcup_{0 \le k \le \frac{m\varepsilon}{\alpha}} \left\{p_{(i^*-k)} \le \tau_\alpha^* - \varepsilon\right\} \cap \{i^* \ge k\}\right) \qquad (5.34) \\
&+ \mathbb{P}\left(\bigcup_{k > \frac{m\varepsilon}{\alpha}} \left\{p_{(i^*-k)} \le \tau_\alpha^* - \frac{\alpha k}{m}\right\} \cap \{i^* \ge k\}\right). \qquad (5.35)
\end{aligned}$$

For (5.34), note that

$$p_{(i^*-k)} \le \tau_\alpha^* - \varepsilon \Rightarrow N_\varepsilon := \sum_{j=1}^{m} 1_{\{p_j \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]\}} \le k,$$

since if at least $i^* - k$ of the p-values fall below $\tau_\alpha^* - \varepsilon$, and exactly i^* of the p-values are below τ_α^*, then at most k of the p-values fall in the interval $[\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$. Since the p-values are independent, we again have $N_\varepsilon \sim$ Generalized-Binomial with sample size m and average success probability $\bar{F}(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \varepsilon)$. By the mean value theorem, for some $\xi \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$, we have

$$\mathbb{E}N_\varepsilon = m(\bar{F}(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \varepsilon)) = m\bar{f}(\xi)\varepsilon \ge m\bar{f}(\tau_\alpha^*)\varepsilon \ge k,$$

since $\bar{f}$ is decreasing on $(0, \tau_\alpha^*)$, $\bar{f}(\tau_\alpha^*) = \alpha^{-1}$, and $k \le \frac{m\varepsilon}{\alpha}$. It follows from Theorem 5 in Hoeffding (1956) that

$$\mathbb{P}(p_{(i^*-k)} \le \tau_\alpha^* - \varepsilon, i^* \ge k) \le \mathbb{P}(N_\varepsilon \le k) \le \mathbb{P}(\widetilde{N}_\varepsilon \le k), \qquad (5.36)$$

where $\widetilde{N}_\varepsilon \sim \text{Binomial}(m, \bar{F}(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \varepsilon))$. Further note that for any $t \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$, the mean value theorem and the condition $\bar{f}' \le -J$ on $[\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$ imply

$$\bar{f}(\tau_\alpha^*) - \bar{f}(t) = \bar{f}'(\xi)(\tau_\alpha^* - t) \le -J(\tau_\alpha^* - t),$$

which further implies the following lower bound on the mean,

$$\begin{aligned}\mathbb{E}\widetilde{N}_\varepsilon &= m\int_{\tau_\alpha^*-\varepsilon}^{\tau_\alpha^*} \bar{f}(t)\mathrm{d}t \\ &\geq m\int_{\tau_\alpha^*-\varepsilon}^{\tau_\alpha^*} \bar{f}(\tau_\alpha^*) + J(\tau_\alpha^* - t)\mathrm{d}t \\ &= m\bar{f}(\tau_\alpha^*)\varepsilon - \frac{mJ}{2}(\tau_\alpha^* - t)^2\bigg|_{\tau_\alpha^*-\varepsilon}^{\tau_\alpha^*} = \frac{m\varepsilon}{\alpha} + \frac{mJ\varepsilon^2}{2}. \end{aligned} \tag{5.37}$$

It follows that (5.36) is bounded,

$$\begin{aligned}\mathbb{P}(\widetilde{N}_\varepsilon \leq k) &= \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \cdot \frac{k}{\mathbb{E}\widetilde{N}_\varepsilon}\right) \\ &\leq \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \cdot \frac{k}{\frac{m\varepsilon}{\alpha}\left(1+\frac{J\alpha\varepsilon}{2}\right)}\right) \\ &\leq \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon \cdot \frac{1}{1+\frac{J\alpha\varepsilon}{2}}\right). && (k \leq \tfrac{m\varepsilon}{\alpha})\end{aligned}$$

Now since $\frac{1}{1+x} \leq 1 - x/2$ for $x \in [0,1]$, and since $\frac{J\alpha\varepsilon}{2} \leq 1$ for m larger than a constant, the above is bounded

$$\begin{aligned}&\leq \mathbb{P}\left(\widetilde{N}_\varepsilon \leq \mathbb{E}\widetilde{N}_\varepsilon\left(1 - \frac{J\alpha\varepsilon}{4}\right)\right) \\ &\leq \exp\left(-\frac{1}{3}\cdot\mathbb{E}\widetilde{N}_\varepsilon\cdot\left(\frac{J\alpha\varepsilon}{4}\right)^2\right) && \text{(Lemma 5.7.7)} \\ &\leq \exp\left(-\frac{1}{3}\cdot\frac{m\varepsilon}{\alpha}\cdot\left(\frac{J\alpha\varepsilon}{4}\right)^2\right),\end{aligned}$$

since (5.37) implies $\mathbb{E}\widetilde{N}_\varepsilon \geq \frac{m\varepsilon}{\alpha}$. Plugging the definition of ε, we have shown

$$\mathbb{P}(\widetilde{N}_\varepsilon \leq k) \leq \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{48}\right) = \exp\left(-\log^3(2m/\delta)\right) \leq \frac{\delta}{2m},$$

so by the union bound, (5.34) is no larger than $\delta/2$.

For (5.35), similar to the first step in the analysis of (5.34), we have the implication

$$p_{(i^*-k)} \leq \tau_\alpha^* - \frac{\alpha k}{m} \Rightarrow N_k := \sum_{j=1}^m 1_{\{p_j \in [\tau_\alpha^* - \frac{\alpha k}{m}, \tau_\alpha^*]\}} \leq k.$$

We have $N_k \sim$ Generalized-Binomial with sample size m and average success probability $\bar{F}\left(\tau_\alpha^*) - \bar{F}(\tau_\alpha^* - \frac{\alpha k}{m}\right)$ because the p-values are independent. By the mean value theorem, for some $\xi \in [\tau_\alpha^* - \frac{\alpha k}{m}, \tau_\alpha^*]$, we have

$$\mathbb{E}N_k = m\bar{f}(\xi) \cdot \frac{\alpha k}{m} \geq k,$$

since $\bar{f}$ is decreasing on $(0, \tau_\alpha^*)$ and $\bar{f}(\tau_\alpha^*) = \alpha^{-1}$. It thus follows from Theorem 5 in Hoeffding (1956) that

$$\mathbb{P}\left(p_{(i^*-k)} \leq \tau_\alpha^* - \frac{\alpha k}{m}, i^* \geq k\right) \leq \mathbb{P}(N_k \leq k) \leq \mathbb{P}(\widetilde{N}_k \leq k),$$

where $\widetilde{N}_k \sim \text{Binomial}\left(m, \bar{F}(\tau_\alpha^*) - \bar{F}\left(\tau_\alpha^* - \frac{\alpha k}{m}\right)\right)$. For any $t \in [\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$, the mean value theorem gives

$$\bar{f}(\tau_\alpha^*) - \bar{f}(t) \leq -J(\tau_\alpha^* - t)$$

since $\bar{f}' \leq -J$ on $[\tau_\alpha^* - \varepsilon, \tau_\alpha^*]$. Since $\bar{f}$ is decreasing on $(0, \tau_\alpha^*)$, this implies

$$\bar{f}(t) \geq \begin{cases} \bar{f}(\tau_\alpha^*) + J(\tau_\alpha^* - t) & \tau_\alpha^* - \varepsilon \leq t \leq \tau_\alpha^* \\ \bar{f}(\tau_\alpha^*) + J\varepsilon & t < \tau_\alpha^* - \varepsilon. \end{cases}$$

Thus $\mathbb{E}\widetilde{N}_k$ is bounded below,

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &= m\int_{\tau_\alpha^* - \frac{\alpha k}{m}}^{\tau_\alpha^*} \bar{f}(t)\mathrm{d}t \\
&= m\int_{\tau_\alpha^* - \frac{\alpha k}{m}}^{\tau_\alpha^* - \varepsilon} \bar{f}(t)\mathrm{d}t + m\int_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} \bar{f}(t)\mathrm{d}t && (k > \tfrac{m\varepsilon}{\alpha}) \\
&\geq m\int_{\tau_\alpha^* - \frac{\alpha k}{m}}^{\tau_\alpha^* - \varepsilon} (\bar{f}(\tau_\alpha^*) + J\varepsilon)\mathrm{d}t + m\int_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} f(\tau_\alpha^*) + J(\tau_\alpha^* - t)\mathrm{d}t \\
&= m\bar{f}(\tau_\alpha^*) \cdot \frac{\alpha k}{m} + Jm\varepsilon\left(\frac{\alpha k}{m} - \varepsilon\right) - \frac{mJ}{2}(\tau_\alpha^* - t)^2\Big|_{\tau_\alpha^* - \varepsilon}^{\tau_\alpha^*} \\
&= k + J\alpha k\varepsilon - mJ\varepsilon^2 + \frac{mJ\varepsilon^2}{2}.
\end{aligned}$$

Simplifying, we have shown

$$\begin{aligned}
\mathbb{E}\widetilde{N}_k &\geq k + J\alpha k\varepsilon - \frac{mJ\varepsilon^2}{2} \\
&> k + J\alpha k\varepsilon - \frac{J\alpha k\varepsilon}{2} && (m\varepsilon < \alpha k) \\
&= k\left(1 + \frac{J\alpha\varepsilon}{2}\right). && (5.38)
\end{aligned}$$

Now since $\frac{1}{1+x} \leq 1 - x/2$ for $x \in [0,1]$, and since $\frac{J\alpha\varepsilon}{2} \leq 1$ for m larger than a constant, we have

$$\begin{aligned}
\mathbb{P}(\widetilde{N}_k \leq k) &= \mathbb{P}\left(\widetilde{N}_k \leq \mathbb{E}\widetilde{N}_k \cdot \frac{k}{\mathbb{E}\widetilde{N}_k}\right) \\
&\leq \mathbb{P}\left(\widetilde{N}_k \leq \mathbb{E}\widetilde{N}_k \cdot \left(1 - \frac{J\alpha\varepsilon}{4}\right)\right) \\
&\leq \exp\left(-\frac{1}{3} \cdot \mathbb{E}\widetilde{N}_k \cdot \left(\frac{J\alpha\varepsilon}{4}\right)^2\right) \\
&\leq \exp\left(-\frac{1}{3} \cdot \frac{m\varepsilon}{\alpha} \cdot \frac{J^2\alpha^2\varepsilon^2}{16}\right),
\end{aligned}$$

since (5.38) together with $k > \frac{m\varepsilon}{\alpha}$ imply $\mathbb{E}\widetilde{N}_k \geq \frac{m\varepsilon}{\alpha}$. Plugging in the definition of ε, we have shown

$$\mathbb{P}(\widetilde{N}_\varepsilon \leq k) \leq \exp\left(-\frac{\alpha J^2 m\varepsilon^3}{48}\right) = \exp\left(-\log^3(2m/\delta)\right) \leq \frac{\delta}{2m},$$

so by the union bound, (5.35) is no larger than $\delta/2$. Since we've now shown that both terms (5.34) and (5.35) are below $\delta/2$, the proof is complete. □

Lemma 5.7.7. *Let $X \sim$ Binomial(n,p). Then for any $0 < \delta < 1/2$, we have*

$$\mathbb{P}(X \geq np(1+\delta)) \leq \exp\left(-\frac{1}{3}np\delta^2\right).$$

Proof. By Markov's inequality, for any $t \geq 0$ we have

$$\mathbb{P}(X \geq np(1+\delta)) \leq \frac{\mathbb{E}e^{tX}}{e^{tnp(1+\delta)}} = \frac{(1-p+pe^t)^n}{e^{tnp(1+\delta)}} \leq \exp\left(np(e^t-1) - tnp(1+\delta)\right).$$

Letting $t = \log(1+\delta)$, we have

$$\mathbb{P}(X \geq np(1+\delta)) \leq e^{np(\delta - (1+\delta)\log(1+\delta))}.$$

Now since $(1+\delta)\log(1+\delta) \geq \delta + \frac{1}{3}\delta^2$ for any $\delta \in (0, 1/2)$, we obtain the result. □

5.7.3 Simulation

Theorem 5.4.1 only requires the assumption that the average density $\bar{f}$ is decreasing over $(0, \tau_\alpha + \alpha)$, and the simulation described in Figure 5.13 checks empirically that $\bar{f}$ need not be decreasing over the entire unit interval in order for $\hat{\tau}_\alpha$ to closely approximate the population threshold τ_α^*.

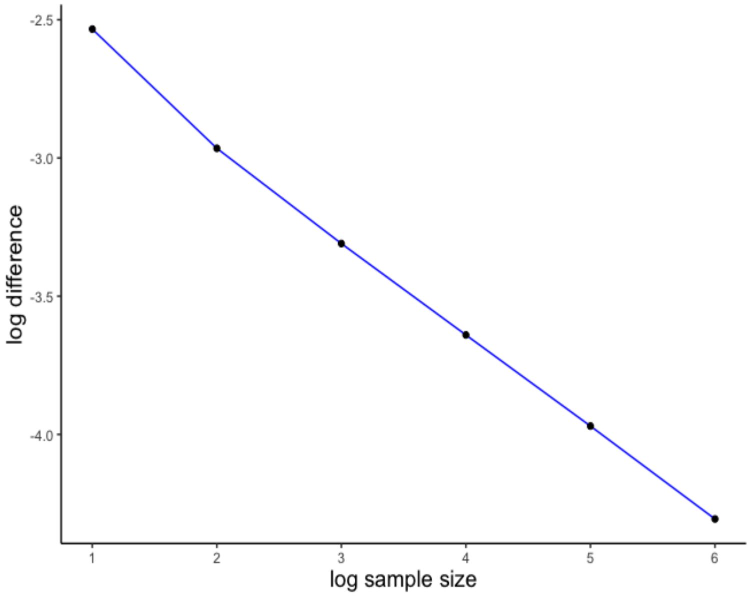


Figure 5.13: For each sample size $m \in \{10^k : k = 1, \ldots, 6\}$, we draw m p-values from the Beta(0.5,0.5) distribution, whose density tends to infinity at 0 and 1. We compute $\hat{\tau}_\alpha$ at level $\alpha = 0.1$ for $N = 10^4$ Monte Carlo trials to estimate the expected difference $\mathbb{E}|\hat{\tau}_\alpha - \tau_\alpha^*|$ for each m. The logged estimates of the error against the log of the sample size. The slope is roughly $-1/3$, which matches the theoretical prediction of Theorem 5.4.1, as lfdr in this setting is a continuous function. The logged absolute differences (vertical axis) are plotted against the log sample size (horizontal axis). The least squares estimate for the slope is $\hat{\beta}_1 \in -0.348 \pm 2(0.0084) = (-0.365, -0.331)$

5.8 Point process formulation

By viewing the set of observations as a realization from a Poisson point process, the lfdr can be meaningfully evaluated at an order statistic. By discarding the labels in our conditioning, we essentially fix the rank of the p-value at which lfdr is evaluated, rather than the index.

Theorem 5.8.1. *Let $n_i \overset{\text{iid}}{\sim}$ Poisson(1) and $p_{ij} \sim f^{(i)}$ independently for $j = 1, \ldots, n_i$ and*

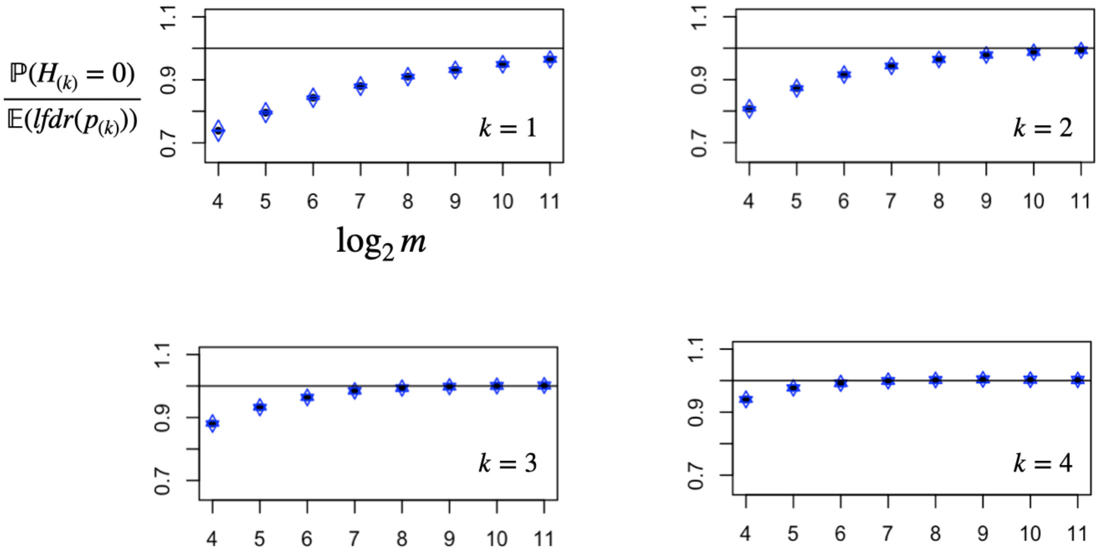


Figure 5.14: For $m \in \{16, 32, 64, \dots, 2048\}$, and $m_1 = 6$ non-nulls, the ratio $\frac{\mathbb{P}(H_{(k)}=0)}{\mathbb{E}(\text{lfdr}(p_{(k)}))}$ is estimated for $k = 1, 2, 3, 4$ based on $N = 10^6$ monte carlo samples and plotted against the log sample size. The nulls are drawn $U(0,1)$ and alternatives from the Beta$(1/4, 1)$ distribution. Blue triangles represent 90% confidence intervals for the true ratio.

$i = 1, \dots, m$. *If each* $f^{(i)}$ *is a continuous density on* $[0,1]$*, then*

$$\text{lfdr}(p_{(k)}) = \mathbb{P}(H_{(k)} = 0 \mid P_n), \quad k = 1, \dots, n \tag{5.39}$$

where $p_{(1)} \leq \cdots \leq p_{(n)}$ *are the order statistics of the total* $n = \sum_{i=1}^m n_i$ *samples, and* $H_{(1)}, \dots, H_{(n)}$ *are the corresponding hypotheses*[9].

Proof. For each i, the set of p-values drawn from $f^{(i)}$ is a Poisson process with intensity $f^{(i)}$, and the union over i of these sets is a Poisson process with intensity $m\bar{f}$. An equivalent way to draw the set of observations is to first draw $n \sim \text{Poisson}(m)$, and conditional on n, draw the p-values iid from the oracle two-groups model,

$$\widetilde{H}_j \mid n \overset{\text{iid}}{\sim} \text{Bernoulli}(1 - \bar{\pi}_0), \quad j = 1, \dots, n$$
$$\widetilde{p}_j \mid \widetilde{H}_j, n \sim \begin{cases} \bar{f}_0 & \widetilde{H}_j = 0 \\ \bar{f}_1 & \widetilde{H}_j = 1, \end{cases} \quad \text{independently.}$$

[9] We assume each $H_i \in \{0, 1\}$ is fixed and for each $n_i > 0$, the H_{ij} corresponding to p_{ij} is set equal to H_i.

By the Superposition Theorem (Kingman; 1992) the resulting set has

$$\left\{(\widetilde{H}_{(1)},\widetilde{p}_{(1)}),\ldots,(\widetilde{H}_{(n)},\widetilde{p}_{(n)})\right\} \overset{(d)}{=} \left\{(H_{(1)},p_{(1)}),\ldots,(H_{(n)},p_{(n)})\right\}. \tag{5.40}$$

The result holds conditional on n, which can be deduced from P_n,

$$\operatorname{lfdr}(p_{(k)}) = \mathbb{P}(H_{(k)} = 0 \mid P_n, n) = \mathbb{P}(H_{(k)} = 0 \mid P_n),$$

where the first equality follows from (5.40) and the Marking Theorem (Kingman; 1992). □

In our original model (5.3) with a single realization from each $P^{(i)}$, a substantial difference between $\operatorname{lfdr}(p_{(k)})$ and $\mathbb{P}(H_{(k)} = 0 \mid P_m)$ is possible without regularity assumptions. This occurs for instance when $\bar{\pi}_0$ is small and each alternative distribution is a dirac measure on a distinct point. In such a scenario, the set of observations would appear dissimilar to a realization from the Poissonized model, where on average $e^{-1}m_1$ of the atoms are not represented. However, the two become close in mean if each distribution has a well-behaved density and the sample size gets large, as illustrated in Figure 5.14 for a Beta alternative. If we instead assume the fixed effects model (5.3), where $H_i = 0 \Rightarrow p_i \sim \text{Uniform}(0,1)$ and each alternative has a continuous density, then (5.39) no longer holds exactly, but is still approximately true in large samples under sparsity assumptions as demonstrated by Theorem 5.8.2, stated and proved below.

Theorem 5.8.2. *Suppose each p_i is independently drawn from the fixed effects model (5.3), with $H_i = 0 \Rightarrow p_i \sim$ Uniform$(0,1)$ and $H_i = 1 \Rightarrow p_i \sim f_1$ satisfying $\sup_{t\in(0,1)} f_1(t) \le B$ and supported on the unit interval, with $\bar{\pi}_0 > 2/3$. Then for any $K \in [m]$ determined by P_m,*

$$\mathbb{P}\left(\left|\operatorname{lfdr}(p_{(K)}) - \mathbb{P}(H_{(K)} = 0 \mid P_m)\right| > CB^2\left(\frac{m_1}{m_0} + \sqrt{\frac{5m_1\log(2m_0/\delta)}{m}}\right)\right) \le \delta,$$

for some universal constant $C > 0$.

Theorem 5.8.2 implies that the null probability for the hypothesis corresponding to the K^{th} smallest observation is essentially determined by $p_{(K)}$ as the fraction of non-nulls decreases, so long as K is measurable with respect to the order statistics.

Proof of Theorem 5.8.2. On the event $\{K = k\}$, we have by exchangeability of the nulls,

$$\mathbb{P}(H_{(k)} = 0 \mid P_m) = m_0\mathbb{P}(p_1 = p_{(k)} \mid P_m),$$

where we have assumed wlog that $H_1 = 0$. By the definition of conditional probability,

$$\mathbb{P}(p_1 = p_{(k)} \mid P_m) = \frac{\sum_{\sigma\in\mathcal{S}_m:\sigma(1)=k}\prod_{i\in\mathcal{H}_1} f_1(p_{(\sigma(i))})}{\sum_{\sigma\in\mathcal{S}_m}\prod_{i\in\mathcal{H}_1} f_1(p_{(\sigma(i))})} \tag{5.41}$$

Let A denote the numerator in the above expression. The denominator can be split into two pieces,

$$\begin{aligned}\sum_{\sigma \in \mathcal{S}_m} \prod_{i \in \mathcal{H}_1} f_1(p_{(\sigma(i))}) &= \sum_{j \in \mathcal{H}_0} \sum_{\sigma : \sigma(j) = k} \prod_{i \in \mathcal{H}_1} f_1(p_{(\sigma(i))}) + \sum_{j \in \mathcal{H}_1} \sum_{\sigma : \sigma(j) = k} \prod_{i \in \mathcal{H}_1} f_1(p_{(\sigma(i))}) \\ &= m_0 A + m_1 f_1(p_{(k)}) \sum_{\sigma : \sigma(2) = k} \prod_{i \in \mathcal{H}_1 \setminus \{2\}} f_1(p_{(\sigma(i))}),\end{aligned}$$

where we have assumed wlog that $H_2 = 1$. Letting $B \coloneqq \sum_{\sigma:\sigma(2)=k} \prod_{i \in \mathcal{H}_1 \setminus \{2\}} f_1(p_{(\sigma(i))})$, plugging the above expression into the denominator of (5.41) yields

$$\mathbb{P}(H_{(k)} = 0 \mid P_m) = \frac{m_0 A}{m_0 A + m_1 f_1(p_{(k)}) B} = \frac{\frac{m_0}{m}}{\frac{m_0}{m} + \frac{m_1}{m} f_1(p_{(k)}) \cdot \frac{B}{A}}.$$

To analyze the ratio $\frac{B}{A}$, note that

$$\begin{aligned} B &= \sum_{\sigma : \sigma(2) = k} \prod_{i \in \mathcal{H}_1 \setminus \{2\}} f_1(p_{(\sigma(i))}) = m_0!(m_1 - 1)! \sum_{\substack{S \subset [m] \setminus \{k\} \\ |S| = m_1 - 1}} \prod_{i \in S} f_1(p_{(i)}) \\ A &= \sum_{\sigma : \sigma(1) = k} \prod_{i \in \mathcal{H}_1} f_1(p_{(\sigma(i))}) = (m_0 - 1)! m_1! \sum_{\substack{S \subset [m] \setminus \{k\} \\ |S| = m_1}} \prod_{i \in S} f_1(p_{(i)}),\end{aligned}$$

since for each summand in B, the null indices can be re-ordered $m_0!$ ways without affecting the product, and the non-null indices can be re-ordered in $(m_1 - 1)!$ ways without affecting the product (since $\sigma(2) = k$ is fixed). Similar reasoning gives the combinatorial factor for A. Dividing the two formulas yields the following expression for the ratio

$$\frac{B}{A} = \frac{m_0}{m_1} \times \frac{\sum_{\substack{S \subset [m] \setminus \{k\} \\ |S| = m_1 - 1}} \prod_{i \in S} f_1(p_{(i)})}{\sum_{\substack{S \subset [m] \setminus \{k\} \\ |S| = m_1}} \prod_{i \in S} f_1(p_{(i)})}.$$

The summation in the denominator can be written

$$\sum_{\substack{S \subset [m] \setminus \{k\} \\ |S| = m_1}} \prod_{i \in S} f_1(p_{(i)}) = \frac{1}{m_1} \sum_{\substack{S \subset [m] \setminus \{k\} \\ |S| = m_1 - 1}} \sum_{j \in S^c} \left[(f_1(p_{(j)}) - 1) \prod_{i \in S} f_1(p_{(i)}) + \prod_{i \in S} f_1(p_{(i)}) \right].$$

where S^c means the complement of S in $[m] \setminus \{k\}$, and division by m_1 discounts the repeated summands. Since on the right hand side $|S^c| = m_0$, it simplifies to

$$= \frac{m_0}{m_1} \sum_{\substack{S \subset [m]\setminus\{k\} \\ |S|=m_1-1}} \prod_{i \in S} f_1(p_{(i)}) \left[1 + \frac{1}{m_0} \sum_{j \in S^c} (f_1(p_{(j)}) - 1) \right].$$

Then the ratio simplifies to

$$\frac{B}{A} \sim 1 + \frac{\sum_{\substack{S \subset [m]\setminus\{k\} \\ |S|=m_1-1}} \prod_{i \in S} f_1(p_{(i)}) \frac{1}{m_0} \sum_{j \in S^c} (f_1(p_{(j)}) - 1)}{\sum_{\substack{S \subset [m]\setminus\{k\} \\ |S|=m_1-1}} \prod_{i \in S} f_1(p_{(i)})}, \tag{5.42}$$

when the rhs is close to one.

Note that the average within the numerator is upper bounded

$$\frac{1}{m_0} \sum_{j \in S^c} f_1(p_j) \leq \frac{1}{m_0} \sum_{j=1}^{m} f_1(p_j) \leq \frac{m_1}{m_0} B + \frac{1}{m_0} \sum_{j \in \mathcal{H}_0} f_1(p_j) \tag{5.43}$$

and lower bounded,

$$\begin{aligned} \frac{1}{m_0} \sum_{j \in S^c} f_1(p_j) &\geq \frac{|S^c \cap \mathcal{H}_0|}{m_0} \cdot \frac{1}{|S^c \cap \mathcal{H}_0|} \sum_{j \in S^c \cap \mathcal{H}_0} f_1(p_j) \\ &\geq \frac{m_0 - m_1}{m_0} \cdot \frac{1}{|S^c \cap \mathcal{H}_0|} \sum_{j \in S^c \cap \mathcal{H}_0} f_1(p_j). \end{aligned} \tag{5.44}$$

For any subset $T \subset \mathcal{H}_0$, Hoeffding's inequality implies

$$\mathbb{P}\left(\left| \sum_{j \in T} f_1(p_j) - |T| \cdot \mathbb{E} f_1(U) \right| \geq x \right) \leq 2 \exp\left(-\frac{2x^2}{|T| B^2} \right),$$

where $U \sim \text{Uniform}(0,1)$. Since $\mathbb{E} f_1(U) = \int f_1(u) \mathrm{d}u = 1$, the above inequality implies

$$\begin{aligned} \mathbb{P}\left(\bigcup_{T \subset \mathcal{H}_0 : |T| \geq m_0 - m_1} \left| \frac{1}{|T|} \sum_{j \in T} f_1(p_j) - 1 \right| \geq \frac{x}{|T|} \right) &\leq 2 \sum_{\ell = m_0 - m_1}^{m_0} \binom{m_0}{\ell} \exp\left(-\frac{2x^2}{\ell B^2} \right) \\ &\leq 2 m_1 \binom{m_0}{m_0 - m_1} \exp\left(-\frac{2x^2}{m_0 B^2} \right), \end{aligned}$$

where we have used $m_0 - m_1 > m_0/2$, which follows from the assumption $\bar{\pi}_0 > 2/3$. Equivalently, we have shown that with probability $\geq 1 - \delta$, it holds simultaneously over all subsets

$T \subset \mathcal{H}_0$ satisfying $|T| \geq m_0 - m_1$ that,

$$\begin{aligned}
\left| \frac{1}{|T|} \sum_{j \in T} f_1(p_j) - 1 \right| &\leq B \sqrt{\frac{m_0 \log\left(2m_1 \binom{m_0}{m_0 - m_1}/\delta\right)}{2(m_0 - m_1)^2}} \\
&= B \sqrt{\frac{m_0}{2(m_0 - m_1)^2} \log\left(\frac{2m_1}{\delta} \cdot \frac{m_0!}{(m_0 - m_1)! m_1!}\right)} \\
&= B \sqrt{\frac{m_0}{2(m_0 - m_1)^2} \log\left(\frac{2}{\delta} \cdot \frac{m_0!}{(m_0 - m_1)!(m_1 - 1)!}\right)} \\
&= B \sqrt{\frac{m_0}{2(m_0 - m_1)^2} \sum_{i=0}^{m_1 - 2} \log\left(\frac{2}{\delta} \cdot \frac{m_0 - i}{m_1 - 1 - i}\right)} \\
&\leq B \sqrt{\frac{m_0 m_1 \log(2m_0/\delta)}{2(m_0 - m_1)^2}}.
\end{aligned}$$

Now because $m_0 - m_1 = m_0 - (m - m_0) = 2m_0 - m = m(2\bar{\pi}_0 - 1) \geq m(4/3 - 1) = m/3$, the above is bounded by

$$\leq B \sqrt{\frac{9m_1 \log(2m_0/\delta)}{2m}} \leq B \sqrt{5 \cdot \frac{m_1 \log(2m_0/\delta)}{m}}.$$

Together with (5.44) and (5.43), this implies that with probability $\geq 1 - \delta$, the following bound holds for every $S \subset [m]$ with $|S^c| \geq m_0 - m_1$

$$\left| \frac{1}{m_0} \sum_{j \in S^c} f_1(p_j) - 1 \right| \leq \frac{m_1}{m_0} B + B \sqrt{\frac{5m_1 \log(2m_0/\delta)}{m}} = B \left(\frac{m_1}{m_0} + \sqrt{\frac{5m_1 \log(2m_0/\delta)}{m}} \right).$$

Combining this with expression (5.42), it holds on this high probability event that

$$\left| \frac{B}{A} - 1 \right| \leq B \left(\frac{m_1}{m_0} + \sqrt{\frac{5m_1 \log(2m_0/\delta)}{m}} \right),$$

which implies that

$$\begin{aligned}
\mathbb{P}(H_{(k)} = 0 \mid P_m) &= \frac{\frac{m_0}{m}}{\frac{m_0}{m} + \frac{m_1}{m} f_1(p_{(k)}) \cdot \left(1 + \frac{B}{A} - 1\right)} \\
&= \operatorname{lfdr}(p_{(k)}) \pm CB^2 \left(\frac{m_1}{m_0} + \sqrt{\frac{5m_1 \log(2m_0/\delta)}{m}} \right)
\end{aligned}$$

for some constant $C > 0$. □

5.9 π_0 adjustment

5.9.1 Estimating m_0

Consider the Storey estimator for the proportion of true null hypotheses (Storey; 2002),

$$\hat{\pi}_0^\lambda := \frac{1 + \#\{i : p_i > \lambda\}}{(1-\lambda)m}, \quad \lambda \in (0,1).$$

The Storey-modified SL procedure was defined in Soloff et al. (2022) as follows:

$$R_\alpha^\lambda := \underset{k \geq 0:\, p_{(k)} \leq \lambda}{\operatorname{argmax}} \left\{ \frac{\alpha k}{\hat{\pi}_0^\lambda m} - p_{(k)} \right\}, \tag{5.45}$$

scanning only over p-values falling below a deterministic threshold λ. The boundary FDR of the Storey-modified procedure is controlled below α, a result we state and prove below.

Theorem 5.9.1. *Suppose that $p_1, \ldots, p_m$ are generated independently from* (5.3) *and that $H_i = 0$ implies $p_i \sim$ Uniform$(0,1)$. Then the Storey-modified procedure* (5.45) *satisfies*

$$\mathrm{bFDR}(\mathcal{R}_\alpha^\lambda) \leq \alpha,$$

where $\mathcal{R}_\alpha^\lambda := \{i : p_i \leq p_{(R_\alpha^\lambda)}\}$.

Proof of Theorem 5.9.1. By exchangeability of the nulls,

$$\mathrm{bFDR}(\mathcal{R}_\alpha^\lambda) = \mathbb{P}(H_{(R_\alpha^\lambda)} = 0) = m_0 \mathbb{P}(p_{(R_\alpha^\lambda)} = p_m),$$

where we have assumed $H_m = 0$ without loss of generality. Letting $p_{-m} := (p_1, \ldots, p_{m-1})$, we claim that

$$\mathbb{P}(p_{(R_\alpha^\lambda)} = p_m \mid \mathcal{F}) \leq \frac{\alpha}{\hat{\pi}_0^\lambda \lambda m} \cdot 1_{\{p_m \leq \lambda\}}, \tag{5.46}$$

where $\mathcal{F} := \sigma(1_{\{p_m \leq \lambda\}}, p_{-m})$ is the sigma field generated by $1_{\{p_m \leq \lambda\}}$ and p_{-m}. To see this, define $p_i^\lambda := p_i/\lambda$ for $i = 1, \ldots, m$, and note that R_α^λ is equivalent to

$$R_\alpha^\lambda = \underset{k=0,\ldots,m^\lambda}{\operatorname{argmax}} \left\{ \frac{\alpha' k}{m^\lambda} - p_{(k)}^\lambda \right\}, \quad m^\lambda := \#\{i : p_i \leq \lambda\}.$$

where $\alpha' := \frac{\alpha m^\lambda}{\hat{\pi}_0^\lambda \lambda m}$. Since $m^\lambda, \hat{\pi}_0^\lambda$ are measurable with respect to $\mathcal{F}$, Lemma 2 of Soloff et al.

(2022) implies (5.46). Marginalizing over $\mathcal{F}$ on both sides of (5.46), we obtain

$$\begin{aligned}\mathbb{P}(p_{(R_\alpha^\lambda)} = p_m) &\le \frac{\alpha}{m} \cdot \mathbb{E}\left(\frac{1}{\hat{\pi}_0^\lambda \lambda} \mid p_m \le \lambda\right) \mathbb{P}(p_m \le \lambda) \\ &= \frac{\alpha(1-\lambda)}{m(1-\bar{F}(\lambda))} \cdot \mathbb{E}\left(\frac{m(1-\bar{F}(\lambda))}{1 + \#\{i < m : p_i > \lambda\}}\right),\end{aligned}$$

where $\bar{F} := \frac{1}{m}\sum_{i=1}^m F^{(i)}$ is the average cdf of the p-values. Since $X \mapsto \frac{m(1-\bar{F}(\lambda))}{1+X}$ is a convex function, Theorem 3 in Hoeffding (1956) implies

$$\mathbb{E}\left(\frac{m(1-\bar{F}(\lambda))}{1 + \#\{i < m : p_i > \lambda\}}\right) \le \mathbb{E}\left(\frac{m(1-\bar{F}(\lambda))}{1+Y}\right) = 1 - \bar{F}(\lambda)^m,$$

where $Y \sim \text{Binomial}(m-1, 1-\bar{F}(\lambda))$, and the last equality follows by direct calculation. Applying this bound to what we have shown above,

$$\text{bFDR}(\mathcal{R}_\alpha^\lambda) \le \frac{\alpha\bar{\pi}_0(1-\lambda)}{1-\bar{F}(\lambda)} \cdot (1 - \bar{F}(\lambda)^m) \le \alpha,$$

since $\bar{\pi}_0(1-\lambda) \le 1 - \bar{F}(\lambda)$ and $1 - \bar{F}(\lambda)^m \le 1$.

□

5.9.2 An adaptive π_0 adjustment

A different method for estimating m_0 was proposed by Benjamini and Hochberg (2000). It is adaptive in the sense that it doesn't require specification of a tuning parameter λ, and proceeds as follows:

1. Calculate $S_i := (1 - p_{(i)})/(m + 1 - i)$.

2. Starting with $i = 1$, proceed to larger i as long as $S_i \ge S_{i-1}$.

3. Stop the first time $S_j < S_{j-1}$ and use

$$\hat{m}_0 := \min\left\{\left\lceil \frac{1}{S_j} \right\rceil, m\right\}. \tag{5.47}$$

The intuition for this estimate is that

$$S_i \approx \frac{m(1-p_{(i)})}{1 - F_m(p_{(i)})}, \quad F_m(t) := \frac{1}{m}\sum_{i=1}^m 1_{\{p_i \le t\}},$$

which is inversely proportional to the Storey estimate using $\lambda = p_{(i)}$. As i increases, we would expect a less conservative estimate of m_0, and $S_j < S_{j-1}$ provides a convenient stopping

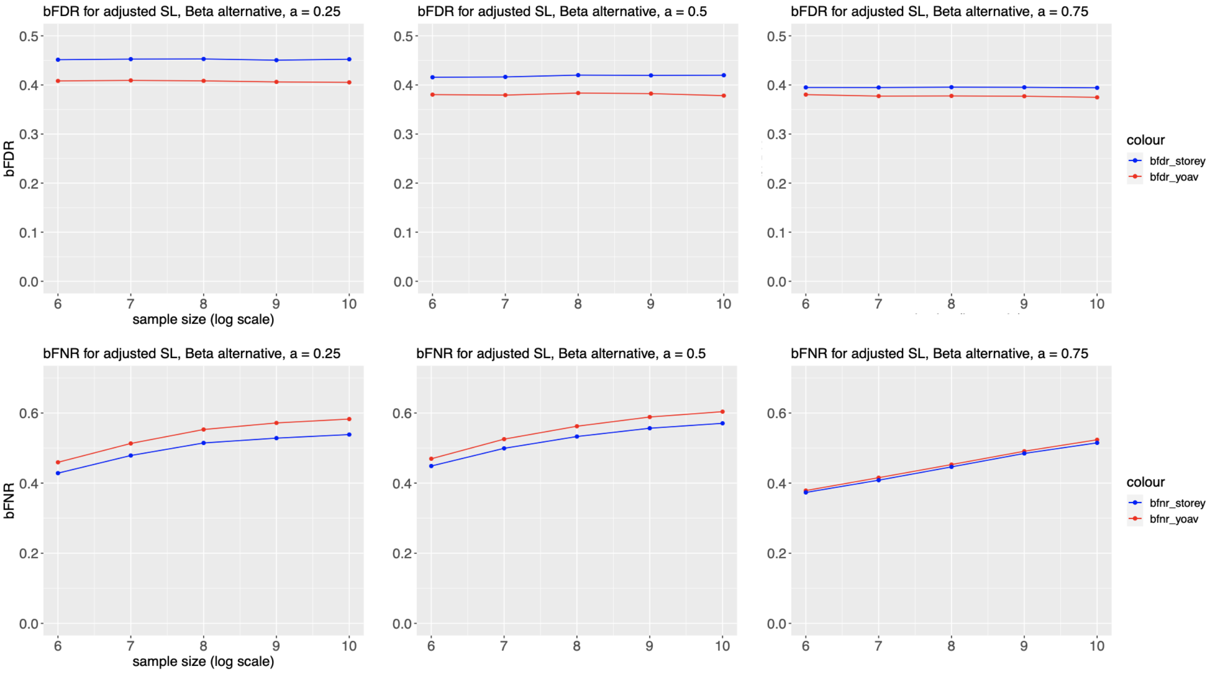


Figure 5.15: For $m \in \{2^6, \ldots, 2^{10}\}$, and $\bar{\pi}_0 = 0.75$, the boundary FDR $= \mathbb{P}(H_{(R)} = 0)$ for the π_0-adjusted procedures (5.45) (blue) and (5.48) (red) are estimated based on $N = 10^5$ Monte Carlo samples and plotted against the $\log_2$ sample size (first row). The boundary FNR $:= \mathbb{P}(H_{(R+1)} = 1)$ is estimated similarly and plotted in the second row for the same two procedures. The nulls are drawn Uniform$(0, 1)$ and alternatives from the Beta$(a, 1)$ distribution, where $a \in \{0.25, 0.5, 0.75\}$.

criterion that is consistent with this monotonicity assumption. To incorporate (5.47) into the SL procedure, we may compute

$$R_{\alpha,S} := \operatorname*{argmax}_{k=0,\ldots,m} \left\{ \frac{\alpha k}{\hat{m}_0} - p_{(k)} \right\}, \tag{5.48}$$

and reject the smallest $R_{\alpha,S}$ most significant hypotheses. Numerical evidence (Figure 5.15) suggests that the procedure $\mathcal{R}_{\alpha,S} := \{i : p_i \leq p_{(R_{\alpha,S})}\}$ controls its bFDR below α, and is slightly more conservative than the Storey-modified procedure with $\lambda = \frac{1}{2}$.

The following figure shows a numerical experiment comparing the boundary FDR and FNR for the $\bar{\pi}_0$-adjusted versions of SL$(\alpha = 0.5)$ based on expressions (5.45) and (5.48). For each Monte Carlo run of the multiple testing experiment, we check whether the last rejection $H_{(R)}$ is null, and whether the next hypothesis $H_{(R+1)}$ is non-null, and plot the empirical proportions among $N = 10^5$ Monte Carlo runs. The bFDR for the procedure (5.48) is slightly smaller than the Storey-adjusted procedure, and both are controlled below the nominal level $\alpha = 1/2$.

5.10 Grouped hypotheses

In the case where p-values are grouped by a covariate $x_i \in \{1, \dots, K\}$, where for any value of x_i,

$$H_i = 0 \Rightarrow p_i \sim \text{Uniform}(0,1),$$

we could separately run the SL procedure on each group,

$$\begin{aligned} R_{\alpha,1} &\coloneqq \underset{k=0,\dots,n_1}{\text{argmax}} \left\{ \frac{\alpha k}{m} - p_{1,(k)} \right\} \\ &\vdots \\ R_{\alpha,K} &\coloneqq \underset{k=0,\dots,n_K}{\text{argmax}} \left\{ \frac{\alpha k}{m} - p_{K,(k)} \right\}, \end{aligned}$$

where $p_{k,(i)}$ for $i = 1, \dots, n_k$ are the ordered p-values with covariate equal to k, and $n_k \coloneqq \#\{i : x_i = k\}$. In other words, we calibrate K thresholds

$$\hat{\tau}_{k,\alpha} \coloneqq p_{k,(R_{\alpha,k})}, \quad k = 1, \dots, K$$

and reject hypotheses whose p-values fall below their group threshold. Then the boundary FDR is controlled within each group according to Theorem 5.3.1. The probability that a randomly selected boundary rejection from among the K groups is a weighted combination of the boundary FDR within each group,

$$\text{bFDR}(\mathcal{R}_\alpha) \coloneqq \sum_{k=1}^{K} \frac{n_k}{m} \cdot \mathbb{P}(H_{(R_{\alpha,k})} = 0),$$

where $\mathcal{R}_\alpha$ is the union of rejection sets within each group. According to Theorem 5.3.1, the group-wise boundary FDR is controlled at $\bar{\pi}_0 \alpha$ when the p-values are independent and Uniform$(0,1)$ distributed under the null,

$$\sum_{k=1}^{K} \frac{n_k}{m} \cdot \mathbb{P}(H_{(R_{\alpha,k})} = 0) = \alpha \sum_{k=1}^{K} \frac{n_k}{m} \cdot \bar{\pi}_{0,k} = \bar{\pi}_0 \alpha,$$

where $\bar{\pi}_{0,k} \coloneqq \frac{\#\{i : H_i = 0, x_i = k\}}{n_k}$ is the null proportion within group K.

REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics, *École d'Été de Probabilités de Saint-Flour XIII—1983*, Springer, pp. 1–198.

Arias-Castro, E. and Ying, A. (2019). Detection of sparse mixtures: Higher criticism and scan statistic, *Electronic Journal of Statistics* **13**(1): 208–230.

Aston, J. A. and Kirch, C. (2012). Detecting and estimating changes in dependent functional data, *Journal of Multivariate Analysis* **109**: 204–220.

Aston, J. A., Kirch, C. et al. (2012). Evaluating stationarity via change-point alternatives with applications to fmri data, *The Annals of Applied Statistics* **6**(4): 1906–1948.

Bai, J. (2010). Common breaks in means and variances for panel data, *Journal of Econometrics* **157**(1): 78–92.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs, *The Annals of Statistics* **43**(5): 2055–2085.

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

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

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

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the kolmogorov statistics, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**(1): 47–59.

Bienvenu, L., Shafer, G. and Shen, A. (2009). On the history of martingales in the study of randomness, *Electronic Journal for History of Probability and Statistics* **5**(1): 1–40.

Bleakley, K. and Vert, J.-P. (2011). The group fused lasso for multiple change-point detection, *arXiv preprint arXiv:1106.4199* .

Butucea, C. and Ingster, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix, *Bernoulli* **19**(5B): 2652–2688.

Cai, T. T., Jeng, J. X. and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5): 629–662.

Cai, T. T. and Wu, Y. (2014). Optimal detection of sparse mixtures against a given null distribution, *IEEE Transactions on Information Theory* **60**(4): 2217–2232.

Chan, H. P., Walther, G. et al. (2015). Optimal detection of multi-sample aligned sparse signals, *The Annals of Statistics* **43**(5): 1865–1895.

Chernoff, H. (1964). Estimation of the mode, *Annals of the Institute of Statistical Mathematics* **16**(1): 31–41.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation, *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 475–507.

Church, A. (1940). On the concept of a random sequence, *Bulletin of the American Mathematical Society* **46**(2): 130–135.

Cole, D. M., Smith, S. M. and Beckmann, C. F. (2010). Advances and pitfalls in the analysis and interpretation of resting-state fmri data, *Frontiers in systems neuroscience* **4**: 8.

Collier, O., Comminges, L. and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes, *The Annals of Statistics* **45**(3): 923–958.

Crane, H. (2016). The ubiquitous ewens sampling formula, *Statistical science* **31**(1): 1–19.

Crane, H. (2018). *Probabilistic foundations of statistical network analysis*, CRC Press.

Dawid, A. P. (1982). The well-calibrated Bayesian, *Journal of the American Statistical Association* **77**(379): 605–610.

Dawid, A. P. (1985a). Calibration-based empirical probability, *The Annals of Statistics* **13**(4): 1251–1274.

Dawid, A. P. (1985b). Probability, symmetry and frequency, *The British journal for the philosophy of science* **36**(2): 107–128.

Dawid, A. P. and Vovk, V. G. (1999). Prequential probability: principles and properties, *Bernoulli* pp. 125–162.

de Finetti, B. (1938). Foresight: Its logical laws, its subjective sources, *Breakthroughs in statistics*, Springer, pp. 134–174.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures, *The Annals of Statistics* **32**(3): 962–994.

Duchi, J. (2016). Lecture notes for statistics 311/electrical engineering 377, *URL: https://stanford. edu/class/stats311/Lectures/full_notes. pdf. Last visited on* **2**: 23.

Edgington, E. and Onghena, P. (2007). *Randomization tests*, CRC press.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press.

Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes, *Statistical Science* **34**(2): 177–201.

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

Efron, B., Turnbull, B. B. and Narasimhan, B. (2011). locfdr: Computes local false discovery rates, *R package version* **1**: 1–7.

Efron, B. and Zhang, N. R. (2011). False discovery rates and copy number variation, *Biometrika* **98**(2): 251–271.

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

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference, *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 495–580.

Gao, C., Han, F., Zhang, C.-H. et al. (2020). On estimation of isotonic piecewise constant signals, *Annals of Statistics* **48**(2): 629–654.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(2): 243–268.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy, *Annals of internal medicine* **130**(12): 995–1004.

Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem, *Lecture Notes-Monograph Series* pp. 266–275.

Greenwald, A., Gonzalez, R., Harris, R. J. and Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated?, *Psychophysiology* **33**(2): 175–183.

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

Gupta, C., Podkopaev, A. and Ramdas, A. (2020). Distribution-free binary classification: prediction sets, confidence intervals and calibration, *Advances in Neural Information Processing Systems* **33**: 3711–3723.

Heller, R. and Rosset, S. (2021). Optimal control of false discovery criteria in the two-group model, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83**(1): 133–155.

Hoeffding, W. (1956). On the distribution of the number of successes in independent trials, *The Annals of Mathematical Statistics* pp. 713–721.

Horváth, L., Kokoszka, P. and Steinebach, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes, *Journal of Multivariate Analysis* **68**(1): 96–119.

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis, *TEST, 23, 219–255.* .

Hu, S., Huang, J., Chen, H. and Chan, H. P. (2021). Likelihood scores for sparse signal and change-point detection, *arXiv e-prints* pp. arXiv–2105.

Hung, K. and Fithian, W. (2020). Statistical methods for replicability assessment, *The Annals of Applied Statistics* **14**(3): 1063–1087.

Ingster, Y. I. (1999). Minimax detection of a signal for l^n-balls., *Math. Methods Statist. 7 (1998), no. 4, 401–428* .

Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O. A. and Heller, R. (2022). Quantifying replicability and consistency in systematic reviews, *Statistics in Biopharmaceutical Research* pp. 1–14.

Jeng, X. J., Cai, T. T. and Li, H. (2013). Simultaneous discovery of rare and common segment variants, *Biometrika* **100**(1): 157–172.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*, Vol. 9, Springer.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association* **107**(500): 1590–1598.

Kingman, J. F. C. (1992). *Poisson processes*, Vol. 3, Clarendon Press.

Kovács, S., Li, H., Bühlmann, P. and Munk, A. (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection, *arXiv preprint arXiv:2002.06633* .

Kovács, S., Li, H., Haubner, L., Munk, A. and Bühlmann, P. (2020). Optimistic search strategy: Change point detection for large-scale data via adaptive logarithmic queries, *arXiv preprint arXiv:2010.10194* .

Lei, L. and Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information, *Journal of the Royal statistical society: series B (Statistical Methodology)* **80**: 649–679.

Li, X. and Fithian, W. (2020). Optimality of the max test for detecting sparse signals with gaussian or heavier tail, *arXiv preprint arXiv:2006.12489* .

Lindley, D. (1966). Mathematical theory of probability and statistics.

Lindsey, J. (1974a). Comparison of probability distributions, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(1): 38–47.

Lindsey, J. (1974b). Construction and comparison of statistical models, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(3): 418–425.

Liu, H., Gao, C. and Samworth, R. J. (2019). Minimax rates in sparse, high-dimensional changepoint detection, *arXiv preprint arXiv:1907.10012* .

Loeb, K. L., Radnitz, C., Keller, K., Schwartz, M. B., Marcus, S., Pierson, R. N., Shannon, M. and DeLaurentis, D. (2017). The application of defaults to optimize parents' health-based choices for children, *Appetite* **113**: 368–375.

Macnamara, B. N. and Burgoyne, A. P. (2022). Do growth mindset interventions impact students' academic achievement? a systematic review and meta-analysis with recommendations for best practices., *Psychological Bulletin* .

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J. and Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias, *Proceedings of the National Academy of Sciences* **119**(31): e2200300119.

McCullagh, P. (2014). An asymptotic approximation for the permanent of a doubly stochastic matrix, *Journal of Statistical Computation and Simulation* **84**(2): 404–414.

McCullagh, P. and Polson, N. G. (2018). Statistical sparsity, *Biometrika* **105**(4): 797–814.

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

Mertens, S., Herberz, M., Hahnel, U. J. and Brosch, T. (2022b). Reply to Maier et al., Szaszi et al., and Bakdash and Marusich: The present and future of choice architecture research, *Proceedings of the National Academy of Sciences* **119**(31): e2202928119.

Moscovich, A., Nadler, B., Spiegelman, C. et al. (2016). On the exact berk-jones statistics and their p-value calculation, *Electronic Journal of Statistics* **10**(2): 2329–2354.

Nair, Y. and Janson, L. (2023). Randomization tests for adaptively collected data, *arXiv preprint arXiv:2301.05365* .

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science, *Revue de l'Institut International de Statistique* pp. 7–22.

Neyman, J. and Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**(694-706): 289–337.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point, *Biometrika* **42**(3/4): 523–527.

Page, E. S. (1954). Continuous inspection schemes, *Biometrika* **41**(1/2): 100–115.

Patra, R. K. and Sen, B. (2016). Estimation of a two-component mixture model with applications to multiple testing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**(4): 869–893.

Peto, R. (1996). Five years of tamoxifen–or more?, *Journal of the National Cancer Institute* **88**(24): 1791–1793.

Pilliat, E., Carpentier, A. and Verzelen, N. (2020). Optimal multiple change-point detection for high-dimensional data, *arXiv preprint arXiv:2011.07818* .

Pitman, J. (2006). *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*, Springer.

Popper, K. (2005). *The logic of scientific discovery*, Routledge.

Reichenbach, H. (1949). *Philosophical foundations of probability*, University of California Press.

Reichenbach, H. (1971). *The theory of probability*, Univ of California Press.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, The Regents of the University of California.

Sallis, A., Harper, H. and Sanders, M. (2018). Effect of persuasive messages on national health service organ donor registrations: a pragmatic quasi-randomised controlled trial with one million uk road taxpayers, *Trials* **19**: 1–10.

Savage, L. J. (1972). *The foundations of statistics*, Courier Corporation.

Sellke, T., Bayarri, M. and Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses, *The American Statistician* **55**(1): 62–71.

Shah, S. P., Lam, W. L., Ng, R. T. and Murphy, K. P. (2007). Modeling recurrent dna copy number alterations in array cgh data, *Bioinformatics* **23**(13): i450–i458.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices, *Pacific Journal of Mathematics* **21**(2): 343–348.

Soloff, J. A., Xiang, D. and Fithian, W. (2022). The edge of discovery: Controlling the local false discovery rate at the margin.

Spielman, S. (1976). Exchangeability and the certainty of objective randomness, *Journal of Philosophical Logic* pp. 399–406.

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

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(1): 187–205.

Strimmer, K. (2008). A unified approach to false discovery rate estimation, *BMC bioinformatics* **9**(1): 1–14.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control, *Journal of the American Statistical Association* **102**(479): 901–912.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S. and Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions, *Proceedings of the National Academy of Sciences* **119**(31): e2200732119.

Tavaré, S. (2021). The magical ewens sampling formula, *Bulletin of the London Mathematical Society* **53**(6): 1563–1582.

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

Tipton, E., Bryan, C., Murray, J., McDaniel, M., Schneider, B. and Yeager, D. (2022). Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity.

Tukey, J. (1989). Higher criticism for individual significances in several tables or parts of tables, *Princeton University, Princeton (Internal working paper)* .

Verzelen, N., Fromont, M., Lerasle, M. and Reynaud-Bouret, P. (2020). Optimal change-point detection and localization, *arXiv preprint arXiv:2010.11470* .

Ville, J. (1939). A counterexample to Richard von Mises's theory of collectives, translation and introduction by Glenn Shafer.

Von Mises, R. (1919). Grundlagen der wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift* **5**(191): 52–99.

Von Mises, R. (1964). *Mathematical theory of probability and statistics*, Academic press.

Wald, A. (1939). Review of probability, statistics and truth.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection series b statistical methodology.

Weinstein, A. (2021). On permutation invariant problems in large-scale inference, *arXiv preprint arXiv:2110.06250* .

Xie, J., Cai, T. T., Maris, J. and Li, H. (2011). Optimal false discovery rate control for dependent data, *Statistics and its interface* **4**(4): 417.

Yekutieli, D. and Weinstein, A. (2019). Hierarchical Bayes modeling for large-scale inference, *arXiv preprint arXiv:1908.08444* .

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods, *The Annals of Statistics* **31**(2): 379–390.

Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010). Detecting simultaneous change-points in multiple sequences, *Biometrika* **97**(3): 631–645.