THE UNIVERSITY OF CHICAGO

TOPICS ON EMPIRICAL BAYES NORMAL MEANS

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

LEI SUN

CHICAGO, ILLINOIS

MARCH 2020

In memory of

Derun Sun

孫德潤父親大人

(1951–2017)

An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.

— John Tukey

Those who ignore Statistics are condemned to reinvent it.

— Bradley Efron

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

This dissertation and my graduate career in general will not come to fruition without the heartwarming and painstaking guidance of my dissertation advisor, Matthew Stephens, at every turn. To me Matthew is a quintessential British gentleman: extremely considerate and nuanced, with a calm sense of humor and sharp, sometimes even biting, wisdom. Almost every time I talk with him about my work, he helps crystalize my thinking, improve my argument, and point out a way forward. For academic work, Matthew has a simple standard – one has to understand the problem through and through – and he truly means it. The training he has put me through is immensely exhilarating and rewarding, and conditioned on that now I have successfully defended my dissertation, worth every bit of it. The warmth and hospitality of Matthew's wife, Lisa, is also worth mentioning. Matthew has turned my life around and I will forever be grateful.

I also want to express my deep gratitude to other members in my dissertation committee: Rina Foygel Barber, James O. Berger, and Nicholas G. Polson. All have provided extensive support at every step. Their ideas and consultations have tremendously enriched my work.

The Department of Statistics at the University of Chicago provides a welcoming and exciting scholarly environment I am proud to call home in the past five years, and I thank the Department Chairs, Yali Amit and Dan Nicholae, and the Director of Graduate Studies, Steve Lalley, for that. I am also deeply appreciative of Peter McCullagh for being a caring and strict mentor. Wonderful faculty, staff, classmates, colleagues, and alumni including Chao Gao, Wei-Biao Wu, Mihai Anitescu, John Reinitz, Mei Wang, Linda Brant Collins, Michael Stein, Jonathan Weare, Stephen Stigler, Steve Lalley, John Lafferty, Debashis Mondal, Mary Sara McPeek, Lek-Heng Lim, Risi Kondor, Jian Ding, Laura Rigazzi, Kirsten Wellman, John Zekos, Jonathan Rodriguez, Keisha Prowoznik, Mitzi Nakatsuka, Christopher McKennan, Yuefeng Han, Yuancheng Zhu, Mahtiyar Bonakdarpour, Ran Dai, Fan Yang, Yongseok Kim, Nathan Gill, Yongrui Chen, Tae Hyun Kim, Joelle Mbatchou,

Jonathan Eskreis-Winkler, Yi Liu, Sen Na, Micol Tresoldi, Keshav Vemuri, Petr Panov, Jing Yu, Bumeng Zhuo, Kan Xu, Siao Lu, Likai Chen, Soudeep Deb, Wooseok Ha, Sayar Karmakar, Vivak Patel, Yunfan Tang, Sze Wai Wong, Ang Li, Li Li, Wanting Xu, Danna Zhang, Marc Goessling, Andrew Poppick, Walter Dempsey, Somak Dutta, Eric Janofsky, Lian Huan Ng, Sheng Zhong, Matthew Reimherr, Harry Crane are all appreciated in my heart.

The research teams under the exemplary leadership of Matthew Stephens, John Novembre, and Xin He have collectively created a uniquely collegial and stimulating atmosphere on the fourth floor of the Cummings Life Science Center, where I have spent most of my days and had coordinated the joint weekly seminar for four years. In addition to the three team leaders, I am thankful to principal investigators Matthias Steinruecken, Hae Kyung Im, Lixing Yang, Mark Abney, support staff Anita Williams-Logan, David Marti, Susan Levison, Danielle Smith, Juan Camacho, my best friends Hussein Al-Asadi, Kushal Dey, David Gerard, computer geniuses Gao Wang, Nan Xiao, John Blischak, Peter Carbonetto, and colleagues and alumni Evan Koch, Joel Smith, Joe Marcus, Daniel Rice, Xiang Zhu, Siming Zhao, Mengyin Lu, Wei Wang, Jacob Degner, Joyce Hsiao, Sarah Urbut, Abhishek Sarkar, Michael Turchin, Yuxin Zou, Jean Morrison, Arjun Biddanda, Jason Willwerscheid, Nicholas Knoblauch, Min Qiao, Yanyu Liang, Shengtong Han, Yuwen Liu, Dongyue Xie, Ben Peter, Mark Reppell, Andrew Goldstein, Yifan Zhou, Desislava Petkova, William Wen, Xiang Zhou, Audrey Fu, Heejung Shim, Ida Moltke, Timothée Flutre, Zhengrong Xing for making this journey a lot more fun and less taxing.

The University of Chicago community offers vibrant intellectual climate and rich career opportunities I feel fortunate to be able to take advantage of. In particular, I thank Julie Marie Lemon, Naomi Blumberg, Lauren M. Jackson, Richard Jean So, Michael Dawson, Dave Thieme, and the Arts, Science & Culture Initiative Grant for a fulfilling collaboration experience. I also thank Michael Tessel and UChicagoGRAD for terrific career and profes-

sional services. I enjoy going to all kinds of talks, workshops, lectures, seminars, symposiums for, besides plenty of free food, fascinating discussions on topics often completely unrelated to my dissertation, such as economics, politics, history, business, public policy, law, finance, and China. I particularly thank John J. Mearsheimer for long conversations in his class and his office which are curiously relaxing and exhausting at the same time.

In addition, I owe great debt to many teachers and mentors in many other places during all these years. They include Lizhong Peng, Marion R. Reynolds, Kenneth Long, Mary Lanier, William H. Woodall, Jeffrey B. Birch, Eric P. Smith, Pang Du, Scotland C. Leman, Feng Guo.

My mother, Baoling Zhang, a tough, intelligent, and accomplished woman, is my biggest and unwavering believer. She serves as a permanent inspiration for my continuous improvement. Enormous love also goes to my wife, Hui Zhang, also a tough, intelligent, and accomplished woman who inspires me every day. My parents-in-law, Yueping Zhang and Weihong Yin, have lent precious support in our hardest time. I also remember my maternal grandmother, Shuhua Sun, and my paternal grandfather, Shiheng Sun, whose love and care I will never have a chance to fully pay back. My daughter, Zoe, came into this world a year ago, and made the final part of my Ph.D. simultaneously more chaotic and more enjoyable. Last but not least, this dissertation is dedicated to my father, Derun Sun, whose unexpectedly fast deterioration of health and ultimate passing away two years ago has completely changed my life. To this day I still feel it is my fault. I wish I had made him proud.

ABSTRACT

The normal means problem plays a fundamental role in many areas of modern statistics, both in theory and practice. The Empirical Bayes (EB) approach to solving this problem has been shown to be highly effective, again both in theory and practice. Indeed, the parallel nature of EB appears to be particularly suited to solve the large-scale data problems prevalent in modern scientific investigations.

Here we present new extensions and applications of this important framework. We design visualization tools for existing EB methods to diagnose their model adequacy in estimating the prior distribution. We devise an EB-based approach to model a certain type of stochastic ordering and to detect the difference in strength between two groups of signals from noisy observations. We also develop new EB methods for solving the normal means problem that take account of unknown correlations among observations. We provide practical software implementations of these methodologies, and illustrate them using realistic numerical experiments and real data problems.

CHAPTER 1

INTRODUCTION

An important inspiration for statistical theories and modeling device in data applications, the Normal Means problem (Robbins, 1951; Johnstone, 2019) can be written as follows:

$$X_j|\theta_j, s_j \overset{\text{ind}}{\sim} N(\theta_j, s_j^2) \ , \qquad j = 1, \ldots, p \ . \tag{1.1}$$

Here $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 and $N(\cdot|\mu, \sigma^2)$ its probability density function (PDF); $\{X_j\} := \{X_1, \ldots, X_p\}$ are observations; $\{s_j\} := \{s_1, \ldots, s_p\}$ are standard deviations that are assumed known; and $\{\theta_j\} := \{\theta_1, \ldots, \theta_p\}$ are unknown means to be estimated. The notation $j = 1, \ldots, p$ is often omitted in the rest of the dissertation for simplicity unless otherwise noted. The goal of the statistical analysis is to estimate or identify non-zero elements of $\{\theta_j\}$ from the observations.

We consider the Empirical Bayes (EB) approach to the Normal Means problem (Efron and Morris, 1973; Johnstone and Silverman, 2004), which assumes that $\{\theta_j\}$ are independent and identically distributed (iid) from some "prior" distribution,

$$\theta_j \overset{\text{iid}}{\sim} g(\cdot) \ , \tag{1.2}$$

and performs inference for θ_j in two steps: first obtain an estimate of g, $\hat{g}$ say, and second compute the posterior distributions $p(\theta_j|X_j, s_j, \hat{g})$. We refer to the two-step process as "solving the Empirical Bayes Normal Means (EBNM) problem." The first step, estimating g, is sometimes of direct interest in itself, and is an example of a "deconvolution" problem (e.g. Kiefer and Wolfowitz, 1956; Laird, 1978; Stefanski and Carroll, 1990; Fan, 1991; Cordy and Thomas, 1997; Bovy et al., 2011; Efron, 2016).

First named by Robbins (1956), EB methods have seen extensive theoretical study (e.g. Robbins, 1964; Morris, 1983; Efron, 1996; Jiang and Zhang, 2009; Brown and Greenshtein,

2009; Scott and Berger, 2010; Petrone et al., 2014; Rousseau and Szabo, 2017; Efron, 2018, 2019), and are becoming widely used in practice. Indeed, according to Efron and Hastie (2016), "large parallel data sets are a hallmark of twenty-first-century scientific investigation, promoting the popularity of empirical Bayes methods."

The EB approach provides a particularly attractive solution to the Normal Means problem. For example, the posterior means of $\{\theta_j\}$ provide shrinkage point estimates, with all the accompanying risk-reduction benefits (Efron and Morris, 1972; Berger, 1985). In addition, the posterior distributions for $\{\theta_j\}$ provide corresponding "shrinkage" interval estimates, which can have good coverage properties even "post-selection" (Dawid, 1994; Stephens, 2017). Further, by estimating g, EB methods "borrow strength" across observations, and automatically determine an appropriate amount of shrinkage from the data (Johnstone and Silverman, 2004). Because of these benefits, methods for solving the EBNM problem – and related extensions – are increasingly used in data applications (e.g. Clyde and George, 2000; Johnstone and Silverman, 2005b; Brown, 2008; Koenker and Mizera, 2014; Wang and Stephens, 2018; Dey and Stephens, 2018; Xing et al., 2019; Urbut et al., 2019).

This dissertation extends the classic EBNM framework and develops new methodologies to solve a wide range of problems arising in large-scale data analysis. We also provide efficient implementations for these methods and illustrate their applications using real data examples and realistic simulations. The rest of the dissertation consists of the following chapters.

- Chapter 2: Designing the diagnostic plots for existing EB methods to assess their model adequacy in estimating g, the crucial first step of solving the EBNM problem.

- Chapter 3: Developing `biashr`, an EBNM-based methodology, to model a certain type of stochastic ordering and compare two groups of signals from noisy observations.

- Chapter 4: Developing `cashr` to solve the EBNM problem with correlated noise, with applications in large-scale multiple testing under dependency.

- Chapter 5: Summarizing the main contributions of this dissertation and discussing possible future research directions on EB.

CHAPTER 2

DIAGNOSTIC PLOTS FOR EMPIRICAL BAYES NORMAL MEANS PROBLEMS

2.1 Introduction

Many EB methodologies assume the following model:

$$\theta_j \overset{\text{iid}}{\sim} g(\cdot) \ , \tag{2.1}$$

$$X_j \overset{\text{ind}}{\sim} f_j(\cdot|\theta_j) \ . \tag{2.2}$$

Here $\{\theta_j\} := \{\theta_1, \ldots, \theta_p\}$ are unobserved signals to be estimated, sampled from an unknown distribution g, and $\{X_j\} := \{X_1, \ldots, X_p\}$ are observations, each generated through a known probability family $f_j(\cdot|\theta_j)$. A large number of EB methods first estimate g from $\{X_j\}$ by methods such as marginal maximum likelihood estimation, and then perform inference on $\{\theta_j\}$ based on their posterior distributions, using the estimated $\hat{g}$ as the prior (e.g. Efron and Morris, 1973; Johnstone and Silverman, 2004; Jiang and Zhang, 2009; Efron, 2014; Koenker and Mizera, 2014; Stephens, 2017; Efron, 2019). In this chapter in particular, we consider a special case of applying EB to the normal means problem:

$$X_j \overset{\text{ind}}{\sim} N(\theta_j, s_j^2) \ , \tag{2.3}$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. We also focus on the `ashr` methods developed in Stephens (2017) to solve the problem.

A crucial task of EB is to estimate g with sufficient accuracy. Recent development in methodologies have made substantial progress on this front (Johnstone and Silverman, 2005a; Efron, 2016; Narasimhan and Efron, 2016; Stephens, 2017; Koenker and Gu, 2017; Sun and

Stephens, 2018). Essentially all of them necessarily operate under specific assumptions. For example, almost all existing EB methods assume independence among observations $\{X_j\}$. Many also impose regularizing assumptions on g. Specifically, `ashr` assumes g is unimodal and thus can be approximated, to arbitrary accuracy, by a mixture of uniform or Gaussian distributions. However, to the best of our knowledge, we have not yet had a simple and direct way to assess the performance of these methods. Here we propose intuitive visualization tools to diagnose the model adequacy of EB methods.

Under the framework of (2.1)-(2.2), $\{X_j\}$ can be seen as independent random samples from their respective marginal distributions

$$h_j(\cdot) := \int_{\mathbb{R}} f_j(\cdot|\theta_j) g(\theta_j) \mathrm{d}\theta_j \ . \tag{2.4}$$

Thus, for continuous $\{X_j\}$, the marginal CDF at $\{X_j\}$

$$H_j := \Pr_{X_j}(\cdot \leq X_j) = \int_{-\infty}^{X_j} h_j(x) \mathrm{d}x \tag{2.5}$$

should be iid samples from $U[0,1]$, where $U[a,b]$ denotes a uniform distribution on $[a,b]$. If the estimated $\hat{g}$ is sufficiently close to the true g, the plug-in estimated CDF

$$\hat{H}_j := \int_{-\infty}^{X_j} \int_{\mathbb{R}} f_j(x|\theta_j) \hat{g}(\theta_j) \mathrm{d}\theta_j \mathrm{d}x \tag{2.6}$$

should be close to $U[0,1]$-distributed. Therefore, the deviation between the distribution of $\{\hat{H}_j\} := \{\hat{H}_1, \ldots, \hat{H}_p\}$ and $U[0,1]$ can be used to evaluate the model adequacy in the estimation of $\hat{g}$.

The rest of the chapter is organized as follows. Section 2.2 computes $\{\hat{H}_j\}$ for `ashr` and designs diagnostic plots to evaluate its uniformity on $[0,1]$. Section 2.3 shows examples of diagnostic plots for `ashr` in various situations when `ashr` fits the data well or poorly. Section

2.4 concludes the chapter.

2.2 Method

2.2.1 Computing the marginals

`ashr` (Stephens, 2017) assumes g is unimodal and can be approximated semi-parametrically by a mixture of uniform or Gaussian distributions as below,

$$g = \sum_k \pi_k g_k \ , \tag{2.7}$$

where g_k can be $N(\mu_k, \sigma_k^2)$ or $U[a_k, b_k]$. All μ_k, σ_k or a_k, b_k, $k = 1, 2, \ldots,$ are pre-specified grids of values. The number of the mixture components and the density of the grids are chosen so that g can be satisfactorily approximated to a reasonable extent. The paper develops and implements the `ashr` methods to estimate $\hat{g}$ by estimating $\hat{\pi}_k$ via convex optimization. Combining (2.3)-(2.7),

$$\hat{H}_j = \sum_k \pi_k \int_{-\infty}^{X_j} \int_{\mathbb{R}} N(\cdot|\theta_j, s_j^2)\hat{g}_k(\theta_j)\mathrm{d}\theta_j := \sum_k \pi_k \hat{H}_{jk} \ , \tag{2.8}$$

where $\hat{H}_{jk}$ can be written out analytically. Let $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and CDF of $N(0, 1)$ respectively.

- When g_k is $N(\mu_k, \sigma_k^2)$,

$$\hat{H}_{jk} = \Phi\left(\frac{X_j - \mu_k}{\sqrt{\sigma_k^2 + s_j^2}}\right) \ . \tag{2.9}$$

- When g_k is $U[a_k, b_k]$,

$$\hat{H}_{jk} = \frac{s_j}{b_k - a_k}\left(\left(\frac{X_j - a_k}{s_j}\Phi\left(\frac{X_j - a_k}{s_j}\right) + \varphi\left(\frac{X_j - a_k}{s_j}\right)\right)\right.$$
$$\left. - \left(\frac{X_j - b_k}{s_j}\Phi\left(\frac{X_j - b_k}{s_j}\right) + \varphi\left(\frac{X_j - b_k}{s_j}\right)\right)\right) . \tag{2.10}$$

- When g_k is δ_{μ_k}, a point mass at μ_k which is usually set to be the case for $k = 0$,

$$\hat{H}_{jk} = \Phi\left(\frac{X_j - \mu_k}{s_j}\right) \tag{2.11}$$

These $\{\hat{H}_j\}$ are then used to construct the diagnostic plots.

2.2.2 Q-Q plot

One of the most widely-used statistical graphs to compare the distribution of $\{\hat{H}_j\}$ and $U[0, 1]$ is the Q-Q (quantile-quantile) plot (Wilk and Gnanadesikan, 1968), plotting the sample quantiles of $\{\hat{H}_j\}$ against the theoretical quantiles of $U[0, 1]$. The sample quantiles are $\{\hat{H}_{(j)}\}$, the order statistics, while the jth theoretical quantiles have no default choice. Choosing these quantiles in a Q-Q plot is not a trivial task. This problem is also known as the "plotting positions" problem in the literature (Kimball, 1960; Harter, 1984; Makkonen, 2008). Available options include $(j - 0.5)/p$ (Hazen, 1914), $j/(p + 1)$ (Weibull, 1939), $(j - 0.3)/(p + 0.4)$ (Lebedev, 1952), $(j - \alpha)/(p - 2\alpha + 1), 0 \leq \alpha \leq 1$ (Blom, 1958), and so on. Here we use $j/(p + 1)$ following Kimball (1946).

2.2.3 Error bounds

Under the null hypothesis of $\{\hat{H}_j\}$ being close to iid $U[0,1]$-distributed, the distribution of $\hat{H}_{(j)}$ is $Beta(j, p+1-j)$ with mean $j/(p+1)$, so

$$\Pr(\hat{H}_{(j)} \in [B^j_{\alpha/2}, B^j_{1-\alpha/2}]) = 1 - \alpha \ , \tag{2.12}$$

where $B^j_{\alpha/2}$ and $B^j_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of $Beta(j, p+1-j)$. We use these quantiles as error bounds, and the distribution of $\{\hat{H}_j\}$ is deemed to be non-$U[0,1]$ if a good number of them are outside the bounds. To make the potential transgression more visible, we de-mean the order statistics by subtracting $j/(p+1)$ from $\hat{H}_{(j)}$ and compare them with $B^j_{\alpha/2} - j/(p+1)$ and $B^j_{1-\alpha/2} - j/(p+1)$ in the plot. In addition, as we are comparing p order statistics at once, which are not independent, there exists an implicit multiple testing under dependency problem. The users can choose the value of α to strike a balance between specificity and power. We set $\alpha = 0.01$ as a software default. We also plot the histogram of $\{\hat{H}_j\}$ and provide the p-value from the Kolmogorov-Smirnov (K-S) test by `ks.test` in `R`.

2.2.4 Histogram of observations

Each X_j is an independent sample from $h_j(\cdot)$ in (2.4), so when p is large, the empirical distribution of $\{X_j\}$ can be approximated by

$$h(\cdot) := \frac{1}{p} \sum_j h_j(\cdot) \ . \tag{2.13}$$

We plot the histogram of $\{X_j\}$ and compare it with the estimated $\hat{h}$ computed from $\hat{g}$:

$$\hat{h}(\cdot) = \frac{1}{p} \sum_j \sum_k \hat{\pi}_k \int_{\mathbb{R}} N(\cdot|\theta_j, s_j^2) g_k(\theta_j) \mathrm{d}\theta_j \ . \tag{2.14}$$

All the above calculations and plots are implemented in the function `plot_diagnostic()` in the `R` package `ashr`.

2.3 Examples

In the following examples, we create synthetic data from underlying models similar to (2.1) and (2.3), apply `ashr` to the data, and use the diagnostic plots to assess the goodness of fit of `ashr`. In all examples, $p = 10^4$.

2.3.1 A good fit

The data-generating models are

$$\theta_j \overset{\text{iid}}{\sim} N(0,1) \ , \tag{2.15}$$

$$X_j \overset{\text{iid}}{\sim} N(\theta_j, 1) \ . \tag{2.16}$$

Here $g = N(0,1)$ is unimodal, in accordance with the assumptions of `ashr`, so `ashr` is expected to perform well. Indeed, the diagnostic plots in Figure 2.1 show that $\{\hat{H}_j\}$ computed from $\hat{g}$ given by `ashr` are roughly $U[0,1]$-distributed, indicating a good fit.

2.3.2 A bad fit

The data-generating models are

$$\theta_j \overset{\text{iid}}{\sim} 0.5N(-2,1) + 0.5N(2,1) \ , \tag{2.17}$$

$$X_j \overset{\text{iid}}{\sim} N(\theta_j, 1) \ . \tag{2.18}$$

The conspicuous bimodal $g = 0.5N(-2,1) + 0.5N(2,1)$ markedly breaks `ashr`'s unimodal assumption, so `ashr` may perform poorly. Correspondingly, the diagnostic plots in Figure

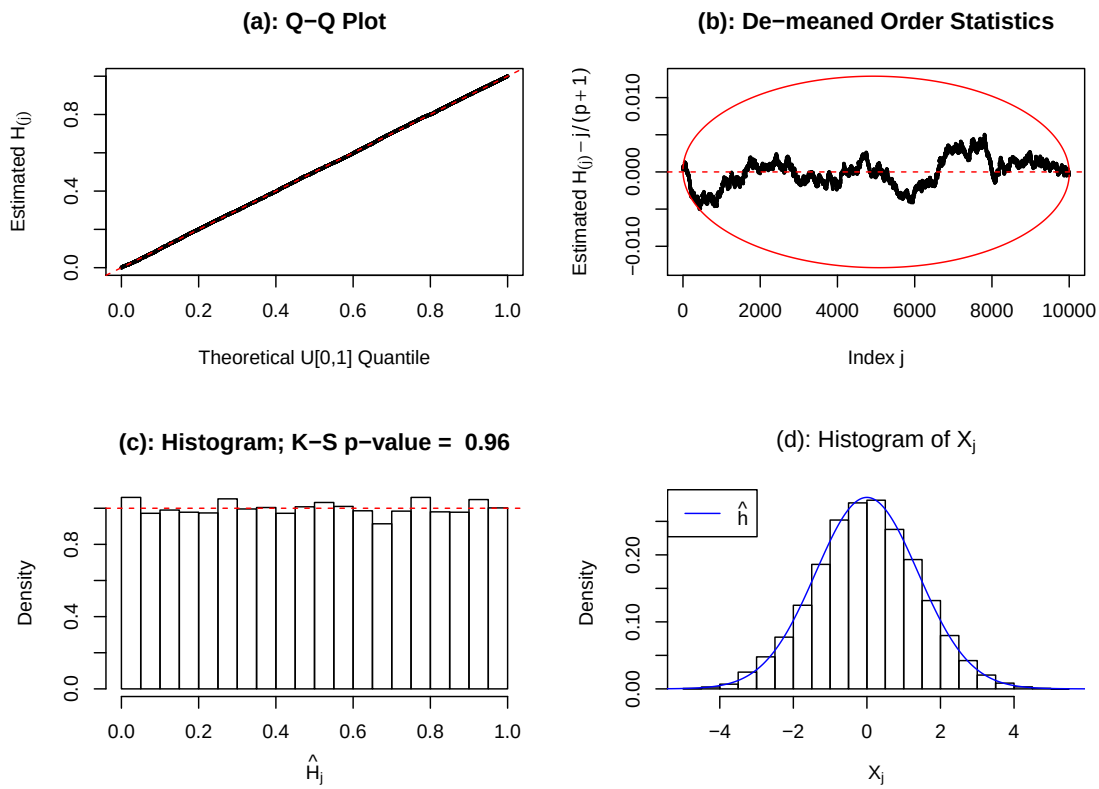

Figure 2.1: The diagnostic plots illustrate that `ashr` fits well when the data are generated in accordance with `ashr`'s assumptions. Panel (a): the Q-Q plot shows that the distribution of $\{\hat{H}_j\}$ is apparently close to $U[0,1]$. Panel (b): the deviation between $\{\hat{H}_{(j)}\}$ and their mean $j/(p+1)$ are well inside the (red) α error bounds. Panel (c): $\{\hat{H}_j\}$ seem to be evenly distributed on $[0,1]$ and the Kolmogorov-Smirnov p-value (0.96) is large. Panel (d): The distribution of $\{X_j\}$ appears to be similar to the computed $\hat{h}$.

2.2 suggest that the distribution of $\{\hat{H}_j\}$ is significantly different from $U[0,1]$, suggesting a bad fit of `ashr`.

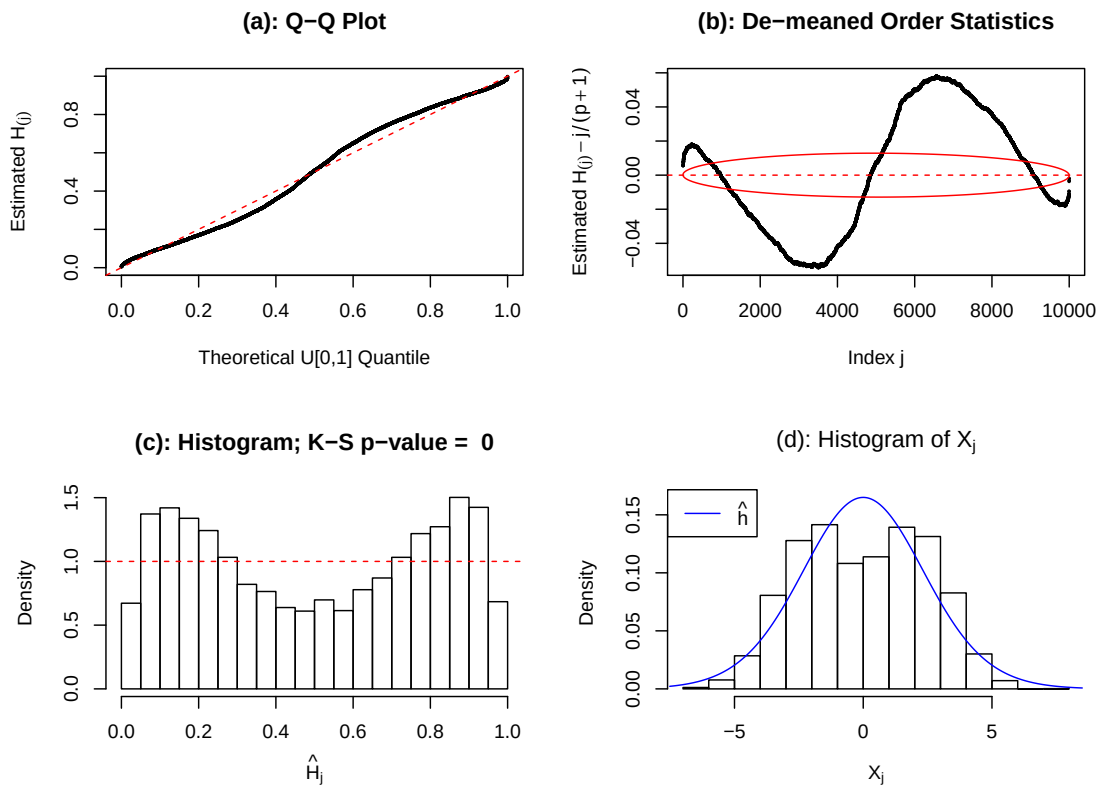

Figure 2.2: The diagnostic plots illustrate that `ashr` may fit poorly when the data are generated disagreeing with `ashr`'s assumptions. Panel (a): the Q-Q plot shows an apparent difference between the distribution of $\{\hat{H}_j\}$ and $U[0,1]$. Panel (b): the deviation between $\{\hat{H}_{(j)}\}$ and their mean $j/(p+1)$ are noticeably large, well outside the (red) α error bounds. Panel (c): $\{\hat{H}_j\}$ are unevenly distributed on $[0,1]$ and the Kolmogorov-Smirnov p-value (0) is highly significant. Panel (d): The distribution of $\{X_j\}$ is markedly different from $\hat{h}$.

2.3.3 A special case: correlated noise in the normal means problem

The data-generating models are

$$\theta_j \overset{\text{iid}}{\sim} N(0,1) \ , \tag{2.19}$$

$$Z_j \sim N(0,1) \ , \ \text{correlated} \ , \tag{2.20}$$

$$X_j = \theta_j + Z_j \ . \tag{2.21}$$

Here $\{Z_j\} := \{Z_1, \ldots, Z_p\}$ are produced using the real RNA-seq gene expression data as detailed in Chapter 4. This way, $X_j \sim N(\theta_j, 1)$ but not independently, breaking a core assumption of the normal means problem `ashr` relies on. Correlation is known to distort the empirical distribution of statistics and thus mislead EB (Efron, 2007a, also see Chapter 4 for more discussion). As expected, the diagnostic plots in Figure 2.3 indicate `ashr` does not fit the data well – the distribution of $\{\hat{H}_j\}$ sees a substantial departure from $U[0, 1]$. It is worth noting that in this situation, even if $\hat{g}$ is perfectly estimated, the empirical distribution of $\{\hat{H}_j\}$ are not supposed to be close to $U[0, 1]$, since the empirical distribution of correlated random variables may be substantially different from their theoretical marginal distribution. Chapter 4 discusses this issue and develops EB methods to tackle it.

2.4 Discussion

In this chapter we introduce ideas to check the model adequacy of EB methods and develop visualization tools for diagnosing `ashr`. We use synthetic data examples to show that `ashr` is sensitive to its central assumptions, including the unimodality of g and the independence of the observations, and that the diagnostic plots are able to visualize `ashr`'s goodness of fit in various situations. Although we focus on `ashr` here, the methods used to construct these plots can easily be adapted and generalized to essentially all EB methods that employ the framework of (2.1)-(2.2).

The diagnostic tools developed in this chapter are based on the fact that the marginal CDF of all observations, $\{H_j\}$, should be independently and uniformly distributed on $[0, 1]$, and so if $\hat{g}$ estimated by EB is sufficiently close to g, the empirical distribution of $\{\hat{H}_j\}$ computed from $\hat{g}$ should be asymptotically close to $U[0, 1]$. All the diagnostic plots, as well as the K-S test, focus on the empirical distribution of the test statistics; therefore, they are applicable to our task. We have not, however, provided a thorough theoretical study of the distribution of $\{\hat{H}_j\}$, as well as their correlation structure induced by the estimation process;

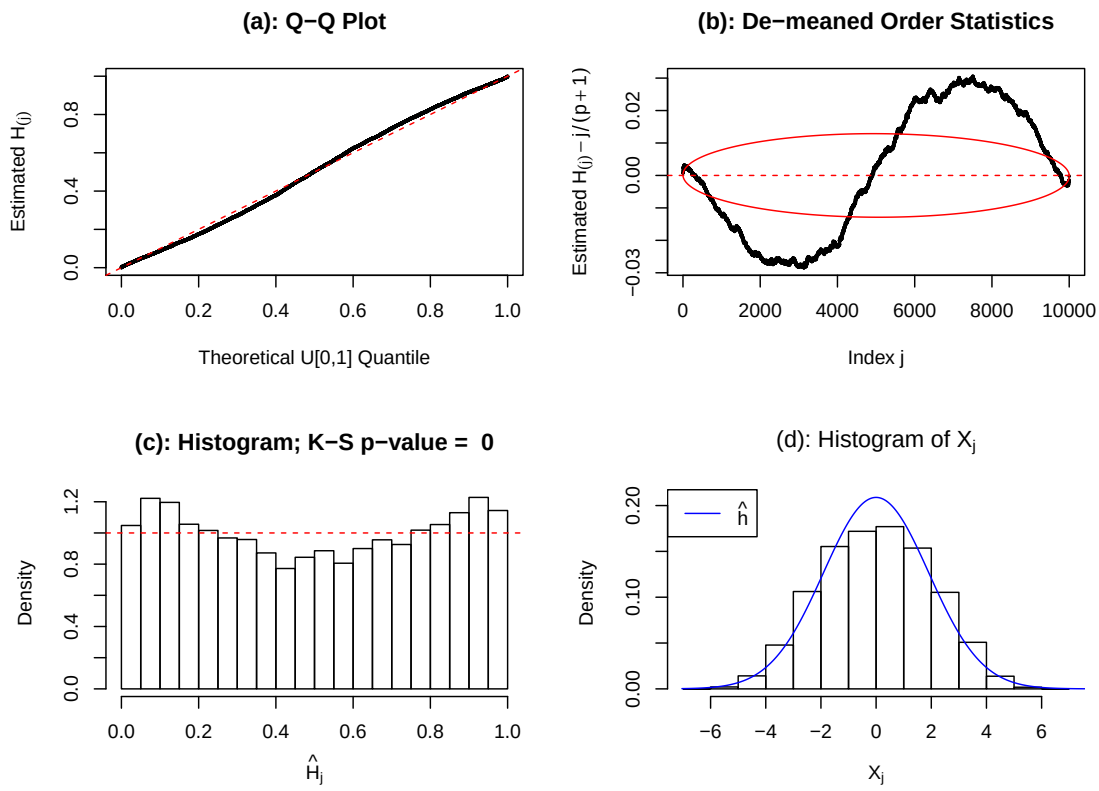

Figure 2.3: The diagnostic plots illustrate that `ashr` may fit poorly when the data contain correlated noise. Panel (a): the Q-Q plot shows the distribution of $\{\hat{H}_j\}$ is different from $U[0,1]$, although the difference is not necessarily conspicuous. Panel (b): $\{\hat{H}_{(j)}\}$ deviate substantially from their mean $j/(p+1)$, well beyond the (red) α error bounds. Panel (c): the distribution of $\{\hat{H}_j\}$ does not fit $U[0,1]$; the Kolmogorov-Smirnov p-value (0) is highly significant. Panel (d): The distribution of $\{X_j\}$ does not fit $\hat{h}$ well, most visibly at the center.

thus, the p-value produced by the K-S test using such plug-in estimates should not be taken too literally, since the underlying assumption of the independence of the marginal CDF of all observations may not hold. Such treatment can be an interesting topic for further research.

CHAPTER 3

EMPIRICAL BAYES METHODS TO COMPARE THE STOCHASTIC ORDERING OF TWO GROUPS OF SIGNALS

3.1 Introduction and Motivation

We consider the problem of comparing the distributions of two independent groups of signals. The signals in each group are themselves independent and identically distributed (iid) from respective underlying distributions as follows.

$$\text{Group 1:} \qquad \{\theta_{1i}\} := \{\theta_{11}, \ldots, \theta_{1p_1}\} \ , \qquad \theta_{1i} \overset{\text{iid}}{\sim} h \ , \quad i = 1, \ldots, p_1 \ ; \tag{3.1}$$

$$\text{Group 2:} \qquad \{\alpha_{2j}\} := \{\alpha_{21}, \ldots, \alpha_{2p_2}\} \ , \qquad \alpha_{2j} \overset{\text{iid}}{\sim} f \ , \quad j = 1, \ldots, p_2 \ . \tag{3.2}$$

Furthermore, the signals $\{\theta_{1i}\}$, $\{\alpha_{2j}\}$ are not directly observed. Instead, only noisy point estimates are obtained such that

$$\text{Group 1:} \qquad X_{1i} \overset{\text{ind}}{\sim} N(\theta_{1i}, s_{1i}^2) \ , \quad i = 1, \ldots, p_1 \ ; \tag{3.3}$$

$$\text{Group 2:} \qquad X_{2j} \overset{\text{ind}}{\sim} N(\alpha_{2j}, s_{2j}^2) \ , \quad j = 1, \ldots, p_2 \ . \tag{3.4}$$

Here $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 and $N(\cdot; \mu, \sigma^2)$ denotes its PDF; $\{X_{1i}\} := \{X_{11}, \ldots, X_{1p_1}\}$, $\{X_{2j}\} := \{X_{21}, \ldots, X_{2p_2}\}$ are observations; $\{s_{1i}\} := \{s_{11}, \ldots, s_{1p_1}\}$, $\{s_{2j}\} := \{s_{21}, \ldots, s_{2p_2}\}$ are respective standard deviations assumed known. For simplicity, the notations $i = 1, \ldots, p_1$, $j = 1, \ldots, p_2$ are often omitted in the rest of the chapter unless otherwise noted.

In this chapter, we focus on situations where one of the groups, Group 2 say, and its distribution f represent signals in the "background," and it is desired to determine whether Group 1 and its distribution h differ from this "background." Specifically we focus on

assessing one aspect of the difference between h and f, namely, whether h is more likely to generate *stronger* signals than f.

To formalize this, we begin by defining a specific stochastic ordering as follows:

Definition 1. *A random variable θ (and its distribution h) is "stochastically stronger (on both sides)" than a random variable α (and its distribution f) if*

$$\forall x < 0 \ , \qquad Pr(\theta \leq x) \geq Pr(\alpha \leq x) \quad \textit{and} \quad Pr(\theta \geq -x) \geq Pr(\alpha \geq -x) \ . \tag{3.5}$$

Under this framework, the question is now to determine whether h is stochastically stronger than f. Classic statistical methods have provided ways to test the stochastic ordering of two distributions (e.g. Lee and Wolfe, 1976; Robertson and Wright, 1981; Franck, 1984; Mau, 1988; Wang, 1996). However, since in our problem $\{s_{1i}\}$, $\{s_{2j}\}$ are potentially heteroskedastic, $\{X_{1i}\}$, $\{X_{2j}\}$ are thus not iid. Therefore, many existing methods are not readily applicable to our data. Conventional approaches circumvent this issue by using z-scores (computed as $z_{\cdot\cdot} = X_{\cdot\cdot}/s_{\cdot\cdot}$) or their corresponding p-values (computed as $p_{\cdot\cdot} = 2\Phi(-|z_{\cdot\cdot}|)$). But this may be misleading: the distribution of z-scores in Group 1 may appear to be stochastically stronger than that in Group 2 simply because $\{s_{1i}\}$ are generally smaller than $\{s_{2j}\}$, even if h and f are exactly the same. In other words, the disparity in measurement precision (or equivalently, statistical power) of the two groups, *not* in their true signal distributions, may pollute the comparison results when using only z-scores or p-values. Figure 3.1 illustrates this point with a simple simulation example. Young et al. (2010); Mandelboum et al. (2019) discussed similar issues in the context of RNA-seq data analysis, where gene lengths, which may affect the measurement error of gene expression, can bias the comparison of the gene expression distributions. In contrast, our method attempts to tackle this issue by employing EB and taking both point estimates and their standard deviations into consideration.

Despite its somehow unfamiliar form, the question being framed as to determine whether

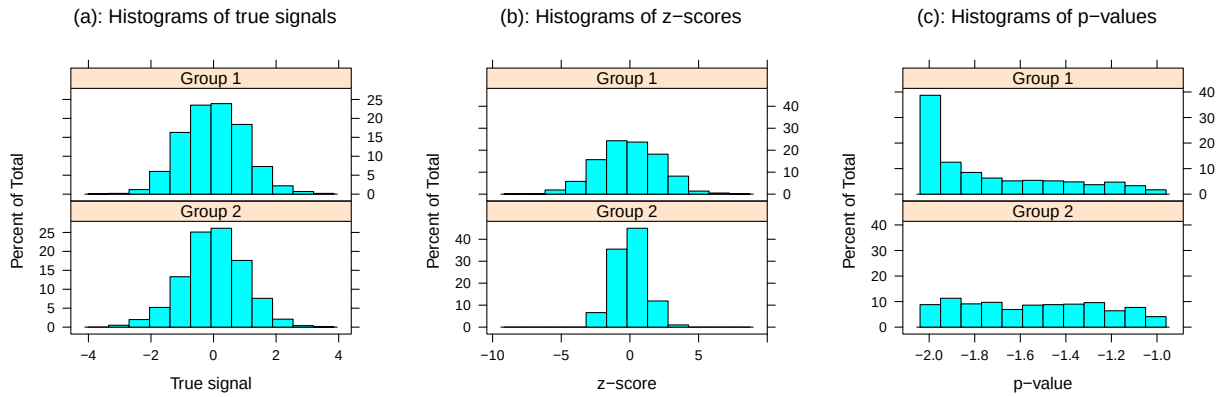

Figure 3.1: Illustration that disparities in measurement precision may mislead the comparison on signal distributions. Synthetic data are simulated according to (3.1)-(3.4), where $h = f = N(0, 1)$, $\{s_{1i}\} \equiv 0.5$, $\{s_{2j}\} \equiv 2$, $p_1 = p_2 = 1000$. That is, the signals in Group 1 and Group 2 are identically distributed, as shown in panel (a), but the measurement error in Group 1 is much smaller than that in Group 2. As a result, panel (b) shows that z-scores in Group 1 are visibly more dispersed than those in Group 2, appearing to provide evidence that the signal distribution in Group 1 is stochastically stronger than that in Group 2 if analysis is solely based on z-scores. Likewise, panel (c) shows a greater concentration on smaller p-values in Group 1 than in Group 2, which in itself may be misleading for comparing signals distributions.

h is stochastically stronger than f from noisy observations can be applied to a variety of problems in which a group of signals of interest are compared with those in the background. In the following, we will introduce two statistical genetics problems as motivating examples to illustrate this.

3.1.1 Gene set enrichment analysis

Many genetic analyses focus not on individual genes, but on sets of biologically related genes, also known as *gene sets* or *pathways*. The research goal, in the context of gene expression studies, is to determine whether an *a priori* defined gene set is "*enriched*," that is, whether the genes in the gene set are more differentially expressed under different biological conditions than other genes on the genome. This is colloquially referred to as *gene set enrichment analysis* (Subramanian et al., 2005). By highlighting the interaction and networking patterns

of a group of genes in a complex biological context, this analysis complements traditional genetic studies which are mainly focused on the behavior of individual genes. Since its introduction in Subramanian et al. (2005), gene set enrichment analysis has been applied to a wide range of fields (e.g. Holden et al., 2008; Ballard et al., 2010; Wang et al., 2011; Jia et al., 2012; Carbonetto and Stephens, 2013; Pinto et al., 2014; Creixell et al., 2015; Zhu and Stephens, 2018; Reimand et al., 2019). The methodology of conducting such analysis is also an active research area, and other approaches have been proposed (e.g. Efron and Tibshirani, 2007; Irizarry et al., 2009; Merico et al., 2010; Carbonetto and Stephens, 2013; Mooney et al., 2014; De Leeuw et al., 2016; Lamparter et al., 2016; Zhu and Stephens, 2018).

To apply the framework (3.1)-(3.4) to this problem, suppose there are in total $p = p_1 + p_2$ genes in a genome-wide gene differential expression study, p_1 of them are in a pre-defined gene set (Group 1) with true signals (the $\log_2$-fold change in gene expression) $\{\theta_{1i}\}$, and the remaining p_2 genes are "in the background" (Group 2) with true signals $\{\alpha_{2j}\}$. Let h, f denote their respective signal distributions. In addition, standard genetic analysis protocols (Smyth, 2004; Robinson et al., 2010; Law et al., 2014; Ritchie et al., 2015) can produce noisy estimates X_{1i}, X_{2j} for θ_{1i}, α_{2j} with standard deviations s_{1i}, s_{2j}. Under this framework, when $h = f$, for instance, the signals in the gene set will be indistinguishable from those in the background, and there should be no enrichment. On the contrary, we consider the gene set to be enriched if it is more likely to contain strong signals, that is, if h is stochastically stronger than f. Therefore, the gene set enrichment analysis can be statistically formulated as to determine whether h is stochastically stronger than f from noisy observations $\{X_{1i}, X_{2j}\}$.

3.1.2 Confounding correction using control genes

One often-encountered issue in gene differential expression studies is that gene expression can be heavily affected by unobserved confounding factors, which can cause "unwanted variations" (Gagnon-Bartsch and Speed, 2012) in the data and hence pollute the estimation

of differential expression effects associated with the covariates of interest such as treatment conditions. Many methods have been proposed to remove these unwanted variations (Leek and Storey, 2007; Sun et al., 2012; Gagnon-Bartsch and Speed, 2012; Wang et al., 2017; Gerard and Stephens, 2019, 2020). Here we focus on the idea of using control genes (Lucas et al., 2006; Gagnon-Bartsch and Speed, 2012), which are genes known *a priori* to be unassociated with the covariates of interest. That is, the genuine differential expression of these genes are known to be zero by definition. Examples of control genes used in practice include spike-in controls (Jiang et al., 2011) and housekeeping genes (Eisenberg and Levanon, 2013). As discussed in more detail in Section 3.3.2, these control genes compose the background Group 2 in our framework, whose "signals" contain only unwanted variations but no genuine differential expression, while the remaining non-control genes as Group 1 are the subjects of interest, whose signals potentially contain both. Existing methods can produce noisy point estimates of these signals, from which the genuine differential expression effects of the non-control genes in Group 1 are to be estimated.

The rest of the chapter is organized as follows. In Section 3.2, we introduce a framework to model the specific stochastic ordering of h and f, and develop EB methods `biashr` to solve the problem. Section 3.3 illustrates the application of `biashr` with real data examples and realistic simulations. Section 3.4 concludes and discusses future research directions.

3.2 Methods

3.2.1 Modeling the stochastic ordering

To model the difference between h and f, we separate θ_{1i} in (3.1) into two independent parts,

$$\theta_{1i} = \beta_{1i} + \alpha_{1i} \ , \tag{3.6}$$

where $\{\beta_{1i}\} := \{\beta_{11}, \ldots, \beta_{1p_1}\}$ and $\{\alpha_{1i}\} := \{\alpha_{11}, \ldots, \alpha_{1p_1}\}$ are also iid from their respective distributions,

$$\beta_{1i}|g \overset{\text{iid}}{\sim} g \ ; \tag{3.7}$$

$$\alpha_{1i}|f \overset{\text{iid}}{\sim} f \ . \tag{3.8}$$

Note that α_{1i} follows the same distribution f as α_{2j} in (3.2). As a result,

$$h = g * f \ , \tag{3.9}$$

where $*$ denotes convolution. When $g = \delta_0$, a Dirac-δ distribution at zero, $h = f$. We further assume both g and f are unimodal and symmetric at zero. Then the following theorem shows that h, thus constructed, is stochastically stronger than f.

Theorem 1. *Let α and β be independent random variables. Suppose both of their respective distributions are continuous on $(-\infty, 0) \cup (0, \infty)$ with possible point mass only at zero, and both are unimodal and symmetric at zero. Then the random variable $\theta = \alpha + \beta$ is stochastically stronger than either α or β.*

Proof. Let f and F be the respective PDF and CDF of α, g and G the respective PDF and CDF of β. As both f and g are symmetric, we only need to show

$$\forall x < 0 \ , \quad \Pr(\theta \leq x) \geq \Pr(\alpha \leq x) \ , \tag{3.10}$$

and then $\Pr(\theta \geq -x) \geq \Pr(\alpha \geq -x)$ holds by symmetry. First suppose both f and g are

continuous on $\mathbb{R}$,

$$\Pr(\theta \leq x) = \Pr(\alpha + \beta \leq x) = \int_{-\infty}^{x} \int_{\mathbb{R}} f(z-y)g(y)\mathrm{d}y\mathrm{d}z = \int_{\mathbb{R}} F(x-y)g(y)\mathrm{d}y \ , \quad (3.11)$$

$$\Pr(\alpha \leq x) = F(x) \ . \quad (3.12)$$

Any distribution symmetric and unimodal at zero can be approximated to arbitrary accuracy by a mixture of symmetric uniform distributions (Khintchine, 1938; Shepp, 1962; Stephens, 2017), so we only need to show that (3.10) holds for all $g = U[-c, c]$, $c > 0$. In that case,

$$\Pr(\theta \leq x) = \frac{1}{2c} \int_{-c}^{c} F(x-y)\mathrm{d}y = \frac{1}{2c} \int_{x-c}^{x+c} F(y)\mathrm{d}y \ . \quad (3.13)$$

Now we define

$$I(c) := \int_{x-c}^{x+c} F(y)\mathrm{d}y - 2cF(x) \ , \quad c \geq 0 \ . \quad (3.14)$$

and we have

$$I'(c) = F(x+c) + F(x-c) - 2F(x) \ ; \quad (3.15)$$

$$I''(c) = f(x+c) - f(x-c) \ . \quad (3.16)$$

Note that because f is symmetric and unimodal at zero, $\forall x < 0, c > 0$,

$$|x+c| < |x-c| \Rightarrow f(x+c) \geq f(x-c) \Rightarrow I''(c) \geq 0 \ . \quad (3.17)$$

Thus,

$$I''(c) \geq 0, \ I'(0) = 0 \Rightarrow I'(c) \geq 0 \ , \quad (3.18)$$

and

$$I'(c) \geq 0,\ I(0) = 0 \tag{3.19}$$

$$\Rightarrow I(c) \geq 0,\ \forall c > 0 \tag{3.20}$$

$$\Rightarrow \Pr(\theta \leq x) = \frac{1}{2c}\int_{x-c}^{x+c} F(y)\mathrm{d}y \geq F(x) = \Pr(\alpha \leq x)\ ,\ \forall c > 0\ . \tag{3.21}$$

If either or both of F and G have a point mass at zero, then the distribution will become a mixture of δ_0 and a continuous one that is symmetric and unimodal at zero. Because the convolution of all possible combinations of one component of F and one component of G results in a distribution that is stochastically stronger than that component of F, θ will be stochastically stronger than α. At last, θ is also stochastically stronger than β by the same argument. □

Using this modeling device, $\{\beta_{1i}\}$ and g become the main objects of interest, and our problem of determining if h is stochastically stronger than f becomes one of determining if $g = \delta_0$. This idea is illustrated in Figure 3.2.

3.2.2 The EB approach

For clarity, our framework can be more concisely re-written as follows.

$$X_{1i} \overset{\text{ind}}{\sim} N(\beta_{1i} + \alpha_{1i}, s_{1i}^2) \tag{3.22}$$

$$X_{2j} \overset{\text{ind}}{\sim} N(\alpha_{2j}, s_{2j}^2) \tag{3.23}$$

$$\beta_{1i} \overset{\text{iid}}{\sim} g \tag{3.24}$$

$$\alpha_{1i}, \alpha_{2j} \overset{\text{iid}}{\sim} f \tag{3.25}$$

Although many possible assumptions on g, f are possible, here we assume g and f to be scale mixtures of zero-mean Gaussians, following the flexible "adaptive shrinkage" approach

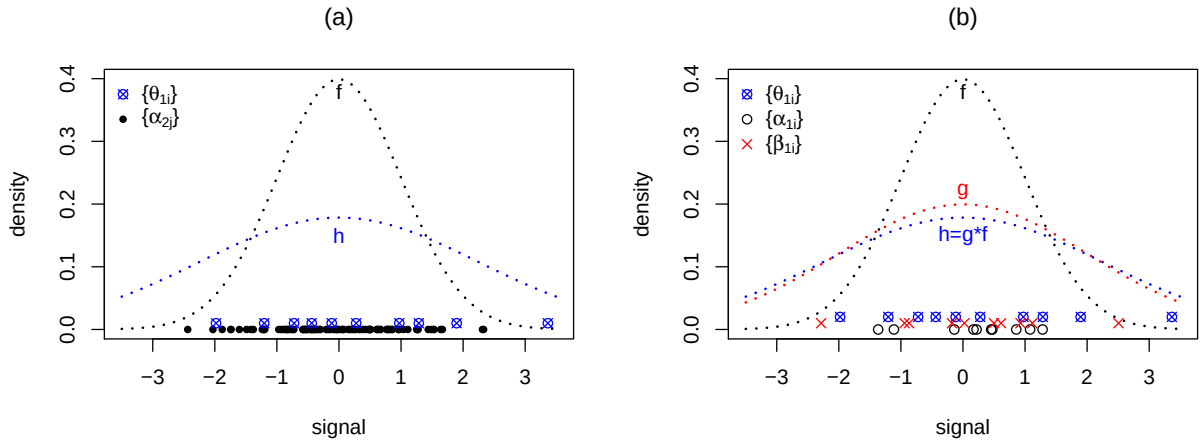

Figure 3.2: Illustration of modeling h being stochastically stronger than f. In this example we have 10 signals $\{\theta_{1i}\}$ in Group 1, and 100 $\{\alpha_{2j}\}$ in Group 2. Panel (a) shows both groups and their respective distributions h and f. Panel (b) separates each θ_{1i} in Group 1 to α_{1i} and β_{1i}, where α_{1i} follows the same distribution as α_{2j} in Group 2, and plots these $\{\theta_{1i}\}$, $\{\alpha_{1i}\}$, $\{\beta_{1i}\}$ with their respective distributions. In this setting, we can determine whether h is stochastically stronger than f by determining whether g is considerably different from δ_0.

in Stephens (2017), and we model them as finite mixtures:

$$g(\cdot;\pi) = \pi_0\delta_0(\cdot) + \sum_{k=1}^{K} \pi_k N(\cdot;0,\sigma_k^2) \ ; \tag{3.26}$$

$$f(\cdot;\omega) = \omega_0\delta_0(\cdot) + \sum_{l=1}^{L} \omega_l N(\cdot;0,\tau_l^2) \ , \tag{3.27}$$

Here the mixture proportions $\pi := \{\pi_0, \pi_1, \ldots, \pi_K\}$, $\omega := \{\omega_0, \omega_1, \ldots, \omega_L\}$ are non-negative and sum to 1, and are to be estimated, whereas the component standard deviations $\sigma_1 < \cdots < \sigma_K$, $\tau_1 < \cdots < \tau_L$ are fixed pre-specified grids of values. π_0, ω_0, in specific, are null proportions. By using a sufficiently wide and dense grid of standard deviations, a finite mixture like this can approximate, to arbitrary accuracy, any scale mixture of zero-mean Gaussians. Using this device, estimating g, f becomes estimating π, ω.

Combining (3.22)-(3.27) and integrating out $\{\beta_{1i}, \alpha_{1i}, \alpha_{2j}\}$, the marginal log-likelihood

for π, ω is

$$\mathcal{L}(\pi,\omega) := \log\left(\prod_{i=1}^{p_1} p(X_{1i}|\pi,\omega)\prod_{j=1}^{p_2} p(X_{2j}|\pi,\omega)\right)$$
$$= \sum_{i=1}^{p_1}\log\left(\sum_{k=0}^{K}\sum_{l=0}^{L}\pi_k\omega_l N(X_{1i};0,\sigma_k^2+\tau_l^2+s_{1i}^2)\right)+\sum_{j=1}^{p_2}\log\left(\sum_{l=0}^{L}\omega_l N(X_{2j};0,\tau_l^2+s_{2j}^2)\right) \tag{3.28}$$

3.2.3 Fitting the model

The usual EB approach to estimating π, ω would be to maximize $\mathcal{L}(\pi,\omega)$. Here we modify this approach by adding penalization. Specifically we use the penalty on π used by Stephens (2017) to encourage conservative (over-)estimation of the null proportion π_0 (to induce conservative estimation of g, our main object of interest). Thus, we solve

$$\begin{aligned}\hat{\pi},\hat{\omega} = \arg\max_{\pi,\omega}\ & \mathcal{L}(\pi,\omega)+\lambda_0\log(\pi_0) \\ \text{subject to } & \sum_{k=0}^{K}\pi_k = 1 \\ & \pi_k \geq 0, \quad k=0,1,\ldots,K \\ & \sum_{l=0}^{L}\omega_l = 1 \\ & \omega_l \geq 0, \quad l=0,1,\ldots,L\ ,\end{aligned} \tag{3.29}$$

where $\lambda_0 \geq 0$ is set to 10 by default.

Problem (3.29) is biconvex. That is, given a feasible $\hat{\pi}$, the optimization over ω is convex; and given a feasible $\hat{\omega}$, the optimization over π is convex. The optimization over either π or ω can be solved using the EM algorithm, or more efficiently and stably using convex optimization methods (Koenker and Mizera, 2014; Koenker and Gu, 2017; Kim et al., 2018).

Here we use `mixsqp` (Kim et al., 2018) to perform convex optimization in each step, and to solve (3.29) we simply iterate between these two steps until convergence. We implemented these methods in `biashr` ("biconvex adaptive shrinkage in `R`"). With $p_1 \simeq 10^2, p_2 \simeq 10^4, K \approx L \simeq 20$, the problem is solved on average within 2 seconds on a personal computer (Apple iMac, 3.2 GHz, Intel Core i5).

3.2.4 Assessing the enrichment

In gene set enrichment analysis, we often have *a priori* hundreds or thousands of candidate gene sets to choose from. For each gene set, we obtain its maximized marginal log-likelihood $\mathcal{L}(\hat{\pi}, \hat{\omega})$ from estimated $\hat{\pi}, \hat{\omega}$. We define the *enrichment score* for this gene set as the average generalized marginal log-likelihood ratio between the estimated model and the null model with $g = \delta_0$ (or equivalently, $\pi_0 = 1$).

$$\frac{1}{p_1}[\mathcal{L}(\hat{\pi}, \hat{\omega}) - \mathcal{L}(\pi_0 = 1, \hat{\omega})] = \frac{1}{p_1} \sum_{i=1}^{p_1} \log \left(\frac{\sum_{k=0}^{K} \sum_{l=0}^{L} \hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)}{\sum_{l=0}^{L} \hat{\omega}_l N(X_{1i}; 0, \tau_l^2 + s_{1i}^2)} \right) , \quad (3.30)$$

which measures the standardized distance between $\hat{g}$ and δ_0. We rank all gene sets by their enrichment scores in descending order. The gene sets ranked on top are believed to show strong evidence for enrichment.

3.2.5 Posterior calculations

The posterior distribution of β_{1i} given $\{X_{1i}, \hat{\pi}, \hat{\omega}\}$ is a mixture of Gaussians:

$$\begin{aligned} &\beta_{1i} | X_{1i}, \hat{\pi}, \hat{\omega} \\ &\sim \sum_k \sum_l \frac{\hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)}{\sum_k \sum_l \hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)} N\left(\frac{\sigma_k^2 X_{1i}}{\sigma_k^2 + \tau_l^2 + s_{1i}^2}, \frac{\sigma_k^2(\tau_l^2 + s_{1i}^2)}{\sigma_k^2 + \tau_l^2 + s_{1i}^2} \right) , \quad (3.31) \end{aligned}$$

from which the analytic forms of key posterior functionals are listed below.

1. The posterior mean of β_{1i}

$$E[\beta_{1i}|X_{1i},\hat{\pi},\hat{\omega}] = \sum_k \sum_l \frac{\hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)}{\sum_k \sum_l \hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)} \cdot \frac{\sigma_k^2 X_{1i}}{\sigma_k^2 + \tau_l^2 + s_{1i}^2} \ . \tag{3.32}$$

   It is used as the point estimate of β_{1i}.

2. The posterior null probability

$$\Pr(\beta_{1i} = 0|X_{1i},\hat{\pi},\hat{\omega}) = \frac{\hat{\pi}_0 \sum_l \hat{\omega}_l N(X_{1i}; 0, \tau_l^2 + s_{1i}^2)}{\sum_k \sum_l \hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)} \ . \tag{3.33}$$

   This quantity is also called the "local false discovery rate (lfdr)." (Efron et al., 2001; Stephens, 2017)

3. The posterior positive probability

$$\begin{aligned} &\Pr(\beta_{1i} > 0|X_{1i},\hat{\pi},\hat{\omega}) \\ &= \sum_{k\geq 1} \sum_l \frac{\hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)}{\sum_k \sum_l \hat{\pi}_k \hat{\omega}_l N(X_{1i}; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)} \Phi\left(\frac{\sigma_k X_{1i}}{\sqrt{\sigma_k^2 + \tau_l^2 + s_{1i}^2}\sqrt{\tau_l^2 + s_{1i}^2}}\right) \ , \end{aligned} \tag{3.34}$$

   where $\Phi(\cdot)$ denotes the CDF of $N(0,1)$. The value of the "local false sign rate (lfsr)" (Stephens, 2017) is defined as

$$\text{lfsr}_j := \min\{\Pr(\beta_{1i} \geq 0 \mid X_{1i},\hat{\pi},\hat{\omega}),\ \Pr(\beta_{1i} \leq 0 \mid X_{1i},\hat{\pi},\hat{\omega})\} \ . \tag{3.35}$$

   Both lfdr and lfsr indicate how likely β_{1i} is non-zero.

3.2.6 Modeling a more general stochastic ordering

The above framework can be readily adapted to account for a wider range of scenarios where researchers postulate whether one distribution is more capable of generating stronger signals than another distribution. Modifying Definition 1, we define a more general stochastic ordering as follows:

Definition 2. *A random variable θ (and its distribution h) is "stochastically stronger" than a random variable α (and its distribution f) if $|\theta|$ is stochastically larger than $|\alpha|$ – that is, if*

$$\forall x > 0 \ , \qquad Pr(|\theta| \geq x) \geq Pr(|\alpha| \geq x) \ . \tag{3.36}$$

With this definition, for example, a gene set is perceived as being enriched if its signal distribution is stochastically stronger than the signal distribution of the genes in the background. In addition, let h_s and f_s be the *symmetrized* distributions of h and f:

$$h^s(\cdot) := \frac{1}{2}(h(\cdot) + h(-\cdot)) \tag{3.37}$$

$$f^s(\cdot) := \frac{1}{2}(f(\cdot) + f(-\cdot)) \tag{3.38}$$

It is easy to show that h is stochastically stronger than f if and only if h^s is stochastically stronger (on both sides) than f^s. Then, the modeling assumptions and devices introduced in Sections 3.2.1-3.2.2 can be applied to model h^s and f^s, as well as their difference, g^s say.

In particular, suppose

$$g^s(\cdot;\pi) = \sum_{k=0}^{K} \pi_k N(\cdot;0,\sigma_k^2) \ , \tag{3.39}$$

$$f^s(\cdot;\omega) = \sum_{l=0}^{L} \omega_l N(\cdot;0,\tau_l^2) \ , \tag{3.40}$$

$$h^s(\cdot;\pi,\omega) = g^s * f^s(\cdot;\pi,\omega) = \sum_{k=0}^{K}\sum_{l=0}^{L} \pi_k\omega_l N(\cdot;0,\sigma_k^2+\tau_l^2) \ , \tag{3.41}$$

where $\sigma_0 = \omega_0 := 0$, $N(\cdot;0,0) := \delta_0(\cdot)$. Under this framework, then, the assessment of whether Group 1 is more likely to contain stronger signals becomes the assessment of whether $g^s = \delta_0$ or $\pi_0 = 1$.

Furthermore, π,ω can be fitted using the absolute values of the observations $\{|X_{1i}|\} := \{|X_{11}|,\ldots,|X_{1p_1}|\}$, $\{|X_{2j}|\} := \{|X_{21}|,\ldots,|X_{2p_2}|\}$. The marginal log-likelihood for π,ω using $\{|X_{1i}|\},\{|X_{2j}|\}$

$$\mathcal{L}^s(\pi,\omega) := \log\left(\prod_{i=1}^{p_1} p_{|X_{1i}|}(|X_{1i}|;\pi,\omega)\prod_{j=1}^{p_2} p_{|X_{2j}|}(|X_{2j}|;\pi,\omega)\right) \ , \tag{3.42}$$

where $p_{|X_{1i}|}(\cdot;\pi,\omega)$ and $p_{|X_{2j}|}(\cdot;\pi,\omega)$ respectively denote the PDF of $|X_{1i}|$ and $|X_{2j}|$ parameterized by π,ω. The analytic form of $\mathcal{L}^s(\pi,\omega)$ is given by the following theorem.

Theorem 2. *Combining* (3.1)-(3.4) *and* (3.39)-(3.42),

$$\begin{aligned}\mathcal{L}^s(\pi,\omega) = C &+ \sum_{i=1}^{p_1}\log\left(\sum_{k=0}^{K}\sum_{l=0}^{L}\pi_k\omega_l N(|X_{1i}|;0,\sigma_k^2+\tau_l^2+s_{1i}^2)\right)\\ &+ \sum_{j=1}^{p_2}\log\left(\sum_{l=0}^{L}\omega_l N(|X_{2j}|;0,\tau_l^2+s_{2j}^2)\right) \ ,\end{aligned} \tag{3.43}$$

where $C = (p_1+p_2)\log 2$.

Proof. Let $F_{|X_{1i}|}(\cdot)$ and $p_{|X_{1i}|}(\cdot)$ denote the marginal CDF and PDF of $|X_{1i}|$ respectively. Let $F_{X_{1i}}(\cdot)$ and $p_{X_{1i}}(\cdot)$ denote the marginal CDF and PDF of X_{1i} respectively. $\forall x > 0$,

$$F_{|X_{1i}|}(x) = \Pr(|X_{1i}| \leq x) = \Pr(-x \leq X_{1i} \leq x) = F_{X_{1i}}(x) - F_{X_{1i}}(-x) \tag{3.44}$$

$$\Rightarrow p_{|X_{1i}|}(x) = p_{X_{1i}}(x) + p_{X_{1i}}(-x) \ . \tag{3.45}$$

From (3.1) and (3.3),

$$p_{X_{1i}}(x) = \int_{\mathbb{R}} h(\theta) N(x; \theta, s_{1i}^2) d\theta = \int_{\mathbb{R}} h(\theta) N(x - \theta; 0, s_{1i}^2) d\theta \tag{3.46}$$

$$\Rightarrow p_{X_{1i}}(-x) = \int_{\mathbb{R}} h(\theta) N(-x - \theta; 0, s_{1i}^2) d\theta = \int_{\mathbb{R}} h(-\theta) N(x - \theta; 0, s_{1i}^2) d\theta \tag{3.47}$$

$$\Rightarrow p_{|X_{1i}|}(x) = \int_{\mathbb{R}} (h(\theta) + h(-\theta)) N(x - \theta; 0, s_{1i}^2) d\theta = \int_{\mathbb{R}} 2h^s(\theta) N(x - \theta; 0, s_{1i}^2) d\theta \ . \tag{3.48}$$

Applying (3.41),

$$\begin{aligned} p_{|X_{1i}|}(x; \pi, \omega) &= \int_{\mathbb{R}} 2 \sum_{k=0}^{K} \sum_{l=0}^{L} \pi_k \omega_l N(\theta; 0, \sigma_k^2 + \tau_l^2) N(x - \theta; 0, s_{1i}^2) d\theta \\ &= 2 \sum_{k=0}^{K} \sum_{l=0}^{L} \pi_k \omega_l N(x; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2) \end{aligned} \tag{3.49}$$

Similarly, from (3.2), (3.4), and (3.40),

$$p_{|X_{2j}|}(x; \pi, \omega) = 2 \sum_{l=0}^{L} \omega_l N(x; 0, \tau_l^2 + s_{2j}^2) \tag{3.50}$$

Finally, (3.43) results from applying (3.49) and (3.50) to (3.42). □

Our EB approach fits π, ω by maximizing the penalized marginal log-likelihood

$$\begin{aligned}
\hat{\pi}, \hat{\omega} = \arg\max_{\pi,\omega} \; & \mathcal{L}^s(\pi, \omega) + \lambda_0 \log(\pi_0) \\
\text{subject to } & \sum_{k=0}^{K} \pi_k = 1 \\
& \pi_k \geq 0, \quad k = 0, 1, \ldots, K \\
& \sum_{l=0}^{L} \omega_l = 1 \\
& \omega_l \geq 0, \quad l = 0, 1, \ldots, L \ ,
\end{aligned} \tag{3.51}$$

where $\lambda_0 \geq 0$ is set to 10 by default. Note that the constrained optimization problems (3.29) and (3.51) will give exactly the same results.

In this context the enrichment score based on $\{|X_{1i}|\}$, $\{|X_{2j}|\}$ becomes

$$\frac{1}{p_1}[\mathcal{L}^s(\hat{\pi}, \hat{\omega}) - \mathcal{L}^s(\pi_0 = 1, \hat{\omega})] = \frac{1}{p_1} \sum_{i=1}^{p_1} \log \left(\frac{\sum_{k=0}^{K} \sum_{l=0}^{L} \hat{\pi}_k \hat{\omega}_l N(|X_{1i}|; 0, \sigma_k^2 + \tau_l^2 + s_{1i}^2)}{\sum_{l=0}^{L} \hat{\omega}_l N(|X_{1i}|; 0, \tau_l^2 + s_{1i}^2)} \right) , \tag{3.52}$$

which measures the standardized distance between $\hat{g}^s$ and δ_0. Note again that the enrichment score computed this way will be numerically the same as the one obtained from the original formation in (3.30). Hence the ranking of the gene sets will also stay the same. In other words, using the absolute values of the observations, we can model in a more general sense the prospect of one distribution being more capable of generating stronger signals than the other, applying essentially the same tools and procedures under less stringent assumptions. It is worth noting that this perspective highlights that our approach is effectively ignoring the information in the data about the signs of the observations. Whether that information is

relevant for our task, and if it is, how to improve `biashr` by incorporating that information, remains a topic for future study.

3.3 Numerical Examples

3.3.1 Gene set enrichment ranking

We now use the well-studied p53 data to illustrate the application of `biashr`. The data set, analyzed previously by Subramanian et al. (2005) and Efron and Tibshirani (2007), contains microarray gene expression data of $p = 10100$ genes from $n = 50$ samples. The samples are under two different conditions: 17 of them are classified as normal and the other 33 carry mutations in the genes. For each gene, the standard limma (Ritchie et al., 2015) – voom (Law et al., 2014) analysis protocol (Smyth, 2004, also discussed in Chapter 4) produces a point estimate, as well as an estimated standard deviation, for the true $\log_2$-fold change in gene expression between the two conditions. The protocol also gives a p-value and a z-score as an indication of the differential expression strength. The distribution of all z-scores, plotted in Figure 3.3, is almost $N(0, 1)$, except for two genes, CDKN1A and BAX, whose z-scores are greater than 5, well above the others. In a study focused only on individual genes, it would be difficult to obtain interesting results besides these two obvious outliers. The data set provides 522 candidate gene sets for consideration. The histogram of the number of genes in each gene set (p_1) is shown in Figure 3.3; most gene sets are composed of less than 50 genes. We now apply `biashr` to perform gene set enrichment analysis on these data.

Before analysis, we first remove the two outlier genes from our data, because their possible presence in a given gene set would heavily skew the estimation of the signal distribution of that gene set. Then we apply `biashr` for each gene set and calculate its enrichment score. The enrichment scores of all 522 gene sets are plotted in descending order in Figure 3.4. It appears that the largest ten of them stand out from the rest, suggesting strong evidence for

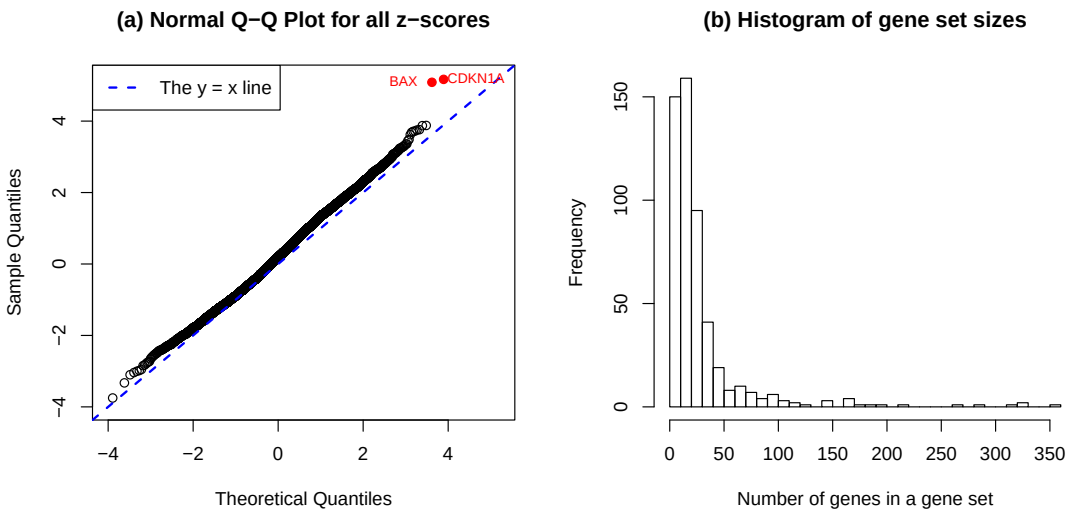

Figure 3.3: Summary of the p53 data. Panel (a) shows that the distribution of all z-scores are close to $N(0, 1)$, with slight mean shift, except for two genes, CDKN1A and BAX, on the right tail. Panel (b) plots the distribution of the gene set sizes and shows that most gene sets consist of less than 50 genes.

enrichment. These gene sets, listed in Table 3.1, are consistent with the results given in both Table 2 in Subramanian et al. (2005) and Table 1 in Efron and Tibshirani (2007). Specifically, six were listed by at least one of the previous studies, while the other four are chosen by `biashr` only. Two gene sets, ngf pathway and ras pathway, were identified by at least one of the previous studies but not `biashr`. To illustrate the commonality and difference in the results given by `biashr` and other methods in more detail, we plot the z-scores of the genes in these gene sets, along with the histogram of all z-scores, in Figure 3.5. The six common discoveries all contain well-dispersed z-scores. Most of them, in particular, contain one or both of the two outliers. The four gene sets exclusively identified by `biashr`, however, do not have those extremely large z-scores, but their z-score dispersions are comparable to that of hsp27 pathway, one of the common discoveries. In contrast, the z-scores in the two gene sets identified by other methods but not by `biashr` are not more dispersed than those in the background, although they appear to be disproportionately negative, indicating some

difference in the signal distributions. Because of our modeling assumptions on g and f, `biashr` is perhaps less well adapted to handling asymmetry like this.

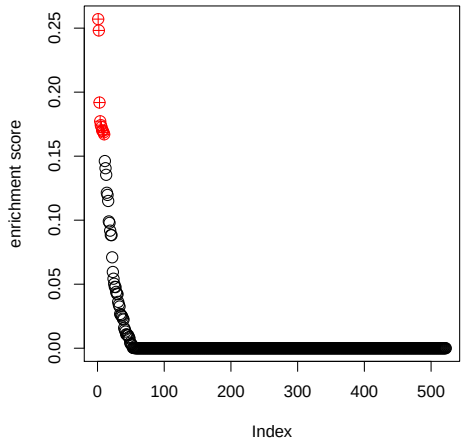

Figure 3.4: Illustration of the gene sets ranked by their enrichment scores. Heuristically the top 10 stand out.

1. p53 up **	6. SA G1 and S phases *
2. hsp27 pathway **	7. il4 pathway
3. arf pathway	8. SA FAS signaling
4. radiation sensitivity **	9. SA DAG1
5. p53 pathway **	10. p53 hypoxia pathway **

Table 3.1: The 10 most enriched gene sets identified by `biashr`. * indicates gene sets identified by Efron and Tibshirani (2007), and ** by both Subramanian et al. (2005) and Efron and Tibshirani (2007). The other gene sets (in green) are identified by `biashr` only.

3.3.2 Confounding correction using control genes

Most existing approaches to account for unobserved confounding factors (Leek and Storey, 2007; Sun et al., 2012; Gagnon-Bartsch and Speed, 2012; Wang et al., 2017; Gerard and Stephens, 2019, 2020) assume the unwanted variations can be captured by the following

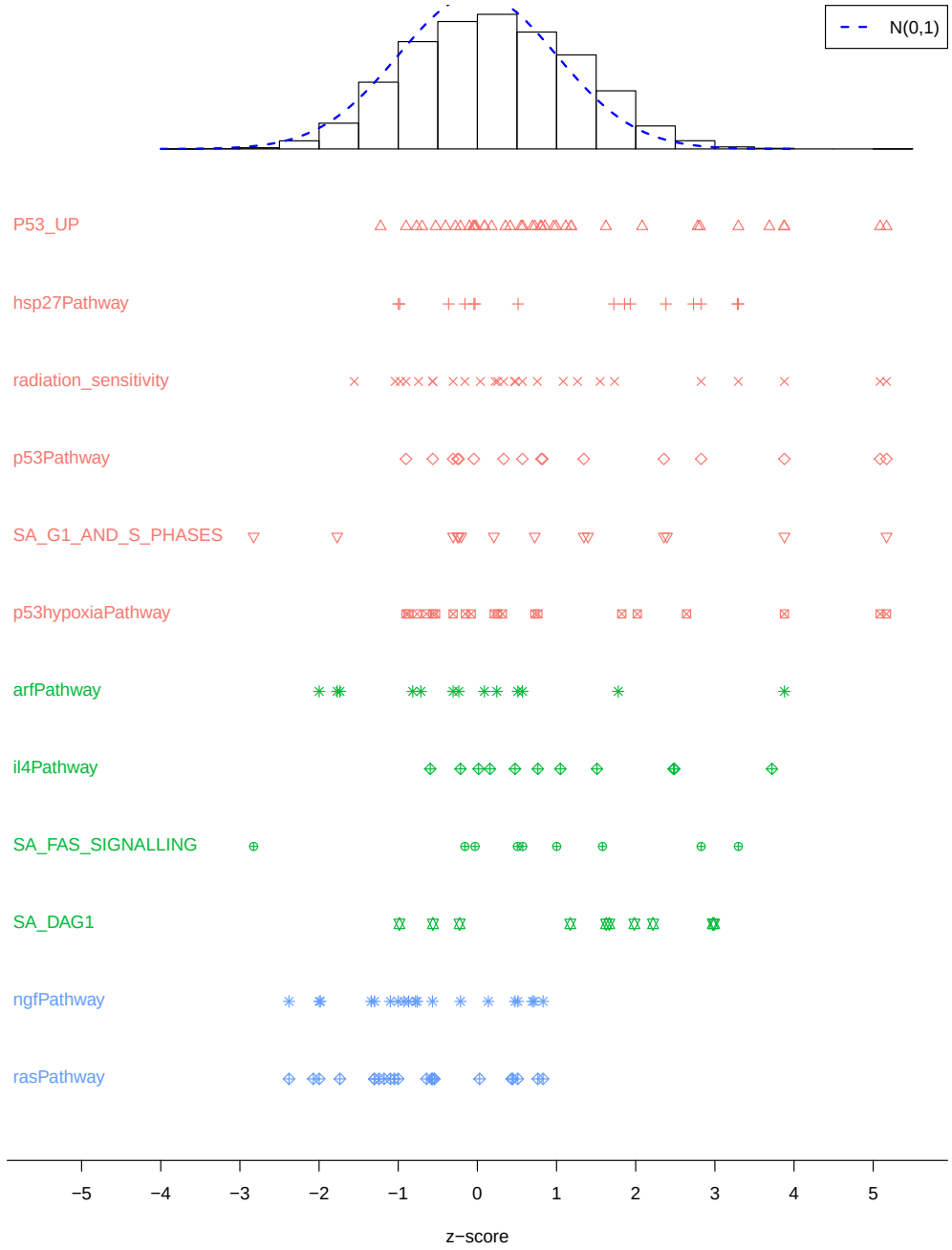

Figure 3.5: Comparison of the z-score distributions in the enriched gene sets. Compared with the distribution of all z-scores (the histogram at the top), the z-scores in the first six gene sets identified by both `biashr` and at least one of the previous two methods (Subramanian et al., 2005; Efron and Tibshirani, 2007) are all well-dispersed. Most contain one or both of the two outlier z-scores. The gene sets identified exclusively by `biashr` but not by the other two also contain well-dispersed z-scores, comparable to those in hsp27Pathway, a common discovery, but no extremely large ones. The z-scores in gene sets identified by other methods but not `biashr`, however, are not more dispersed, though arguably more asymmetric, than z-scores in general.

low-rank factor model,

$$Y_{n\times p} = X_{n\times k}B_{k\times p} + Z_{n\times q}A_{q\times p} + E_{n\times p} \ , \tag{3.53}$$

where Y consists of the expression levels of $p = p_1 + p_2$ genes (including p_1 non-control genes and p_2 control genes) in n samples, X the observed covariates, B the coefficients or "effects" of X, Z the unobserved covariates or confounders, A the coefficients or "loadings" of Z, and E the independent Gaussian noise with mean 0 and gene-specific variance. In the context of gene differential expression studies, we typically have $p \gg n > k$, $p \gg q$, and we assume there is only one covariate of interest in X, usually the condition or treatment assignment, and that covariate is, without loss of generality, the kth column of X. The research goal is to make inference on its corresponding effects, i.e., the kth row of B, denoted as $[\beta_m] := [\beta_1, \ldots, \beta_p]$, without directly estimating A or Z.

Employing the QR decomposition of X, Wang et al. (2017) and Gerard and Stephens (2020) transformed (3.53) into a simplified model

$$\hat{\beta}_m \overset{\text{ind}}{\sim} N(\beta_m + (A^T\tilde{z})_m, s_m^2) := N(\beta_m + \alpha_m, s_m^2) \ , \qquad m = 1, \ldots, p \ , \tag{3.54}$$

where $\hat{\beta}_m$ is the OLS estimate of β_m obtained by regressing the mth column of Y on X and taking the kth element of the result, $\alpha_m := (A^T\tilde{z})_m$ is the unwanted variation to be removed, $\tilde{z}$ is derived from X and Z, and s_m is the standard deviation of $\hat{\beta}_m$. (See the online supplementary material to Gerard and Stephens, 2020, for more details.) Using "1" as a subscript to denote the set of non-control genes as Group 1 and "2" that of control genes as Group 2, (3.54) can be written separately for p_1 non-control genes and p_2 control genes in the same form as (3.22)-(3.23), following the fact that the effects of control genes $\beta_{2j} \equiv 0$ by definition. Assuming that the loadings for non-control genes and for control genes are similarly distributed (an implicit assumption also in Gerard and Stephens, 2019), the

distributions of $\{\alpha_{1i}\}$ and $\{\alpha_{2j}\}$ should be similar. Therefore, we can apply the same `biashr` framework to solve the problem in two steps: first estimate g, f and then make inference on $\{\beta_{1i}\}$ based on its posterior distribution given $\hat{g}, \hat{f}$ and $\hat{\beta}_{1i}$ (or X_{1i} in (3.31)).

To assess the performance of `biashr` in confounding correction using control genes, we compare it with other cutting-edge methods, including `ruv2` (Gagnon-Bartsch and Speed, 2012), `ruv3` (Gerard and Stephens, 2019), and `cate` (Wang et al., 2017), all of which assume a low-rank factor model and attempt to estimate the confounding structure using control genes. We also apply basic OLS to provide a baseline. To make the confounding artifacts realistic, we use real RNA-seq gene expression data in human muscle tissues (The GTEx Consortium, 2015, 2017), similar to the simulation schemes employed in Lu (2018); Gerard and Stephens (2019, 2020) and discussed in Gerard (2019). In each simulation, we randomly select $p = 1100$ genes out of the 10^4 most expressed genes from $n = 10$ random tissue samples, and randomly assign 5 samples as under treatment and the other 5 as controls. Hence, no genes should have differential expression between the two conditions. Then we randomly designate $p_2 = 100$ genes as control genes. For the remaining $p_1 = 1000$ non-control genes, we use the `R` package `seqgendiff` (Gerard, 2019) to add, between the two conditions, synthetic effects $\{\beta_{1i}\}$ from the distribution

$$g = \pi_0 \delta_0 + (1 - \pi_0) N(0, 0.8^2) \ , \tag{3.55}$$

for three choices of $\pi_0 \in \{0.1, 0.5, 0.9\}$. For each π_0, 1000 simulations are run. The simulated data, the condition assignment, and the set of control genes are made available to all methods, while the added effects are withheld. The analysis goal is to estimate $\{\beta_{1i}\}$ and determine which of them are more likely to be non-zero. To assess the accuracy of estimating $\{\beta_{1i}\}$, we use MSE $:= \frac{1}{p_1} \sum_{i=1}^{p_1} (\hat{\beta}_{1i} - \beta_{1i})^2$ as a criterion, computed from point estimates given by OLS, `ruv2`, `ruv3`, `cate` and the posterior mean by `biashr`, lower MSE indicating better performance. To assess the ability for distinguishing non-null effects from null non-effects, we

calculate the area under the receiver operating characteristic curve (AUC) for each method, using p-values given by OLS, `ruv2`, `ruv3`, `cate` and lfsr by `biashr`, higher AUC indicating better performance.

Figure 3.6 compares the performance of all methods. On estimation accuracy, according to MSE, `biashr` outperforms `ruv2`, `ruv3`, `cate` considerably in all scenarios, while the latter three produce similar results and are all better than basic OLS. The superior performance of `biashr` in this realm partly comes from the risk-reducing shrinkage property of EB. On identifying non-null effects, according to AUC, `biashr`'s results are similar to, and sometimes slightly better than, those of `ruv2`, `ruv3`, `cate`, all of which are better than basic OLS. Figure 3.6 provides evidence that `biashr` can satisfactorily remove the unwanted variations, without having to estimate the confounding structures directly.

3.4 Discussion

We have introduced a framework to compare the stochastic ordering of two groups of signals, where one group is perceived to be stochastically stronger than the other. The framework can be applied to a variety of problems, including gene set enrichment analysis and confounding correction using control genes. We have also presented an EB approach to solve this problem and exploited convex optimization techniques to provide an efficient implementation. Real data examples and realistic simulations show that `biashr` produce comparable, and in some aspects, superior performance compared with other commonly-used methodologies.

One underlying assumption of `biashr` is that the standard deviations $\{s_{1i}\}$, $\{s_{2j}\}$ are known. In practice, they are usually estimated. In gene differential expression studies, the `voom-limma` pipeline has been shown to be capable of producing fairly satisfactory estimates. (See Rocke et al., 2015; Gerard and Stephens, 2019, and Chapter 4 for discussions.) Moreover, the main object of interest in our framework is g, which represents the difference between two signal distributions. Even if the estimates of $\{s_{1i}\}$, $\{s_{2j}\}$ are biased, so along as they

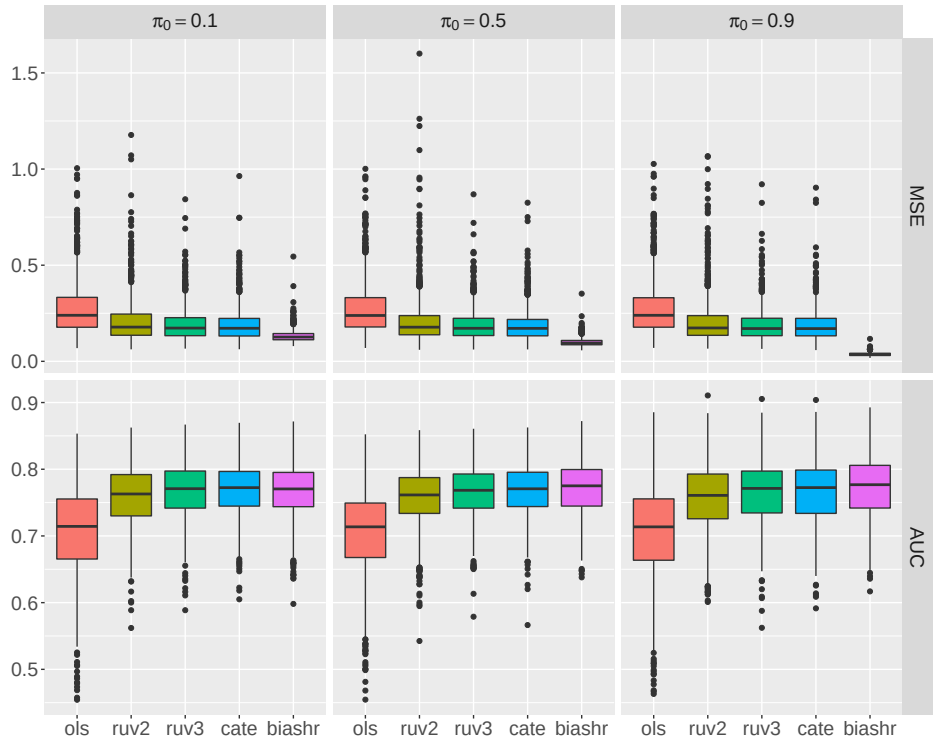

Figure 3.6: Illustration of the performance of `biashr` on estimating effects and identifying non-null effects. Results are shown for three different null proportions ($\pi_0 \in \{0.1, 0.5, 0.9\}$) of the true effect distribution. Two criteria are used: MSE measures the estimation accuracy, while AUC measures the ability to distinguish non-null effects from null non-effects. In all scenarios, compared with factor analysis-based methods (`ruv2`, `ruv3`, `cate`), `biashr` produces substantially lower MSE and comparable AUC, without directly estimating the underlying factor structures. All these methods see clearly improvement over basic OLS.

are biased in a systemic way, the bias can be absorbed into f and thus will not necessarily affect the inference results regarding g.

In gene set enrichment analysis, `biashr` gives a ranking of gene sets according to their enrichment scores, but does not provide a significance threshold. Such threshold can be produced by permutation-based resampling methods in the following way:

1. For each gene set, create a large number of artificial non-enriched gene sets of the same size through random permutation, and calculate their enrichment scores.

2. Compare the enrichment score of the original gene set with those of the artificial ones,

and obtain a p-value for enrichment.

3. Apply multiple testing procedures to these p-values to determine a significance threshold according to a certain Type-I error criterion.

We did not pursue this line but instead used visual heuristics to identify enriched gene sets, as plots such as Figure 3.4 turned out to be quite informative for our task.

So far f and g in `biashr` are assumed to be scale mixtures of zero-mean Gaussians. These assumptions might be too restrictive for some applications, and can be modified to provide more flexible modeling options, including asymmetry or an non-zero mode. The implementation of such extensions could be an area for future exploration.

Our focus here in this chapter is on a specific difference between two signal distributions, namely, the disparity of signal strength in terms of whether or not one being stochastically stronger than the other. It does not include all kinds of difference researchers may find interesting. For example, if Group 1 contains signals generated from $h = N(0, 1)$, while Group 2 from f being a half-normal distribution with scale 1, then h and f have no difference with respect to stochastic strength according to our definition, and so it will be difficult for `biashr` to produce significant results, although researchers may hope to be able to identify the apparent difference between h and f from data. Another related example is the arguably problematic non-identification of ngfPathway and rasPathway by `biashr` in Figure 3.5. Therefore, generalizing our framework or combining our methods with other methods of distribution comparison to tackle similar scenarios can be of further research interest.

CHAPTER 4

SOLVING THE EMPIRICAL BAYES NORMAL MEANS PROBLEM WITH CORRELATED NOISE

4.1 Introduction

We consider the following EBNM problem,

$$X_j \sim N(\theta_j, s_j^2) \ , \qquad j = 1, \ldots, p \ , \tag{4.1}$$

$$\theta_j \overset{\text{iid}}{\sim} g(\cdot) \ . \tag{4.2}$$

The EBNM approach provides attractive features such as "borrowing strength" across parallel observations (Johnstone and Silverman, 2004), performing risk-reducing shrinkage estimation (Efron and Morris, 1972; Berger, 1985), and demonstrating good "post-selection" properties (Dawid, 1994; Stephens, 2017). One application that we pay particular attention to later in this chapter is large-scale multiple testing, and estimation/control of the False Discovery Rate (FDR; Benjamini and Hochberg, 1995; Efron, 2010b; Muralidharan, 2010; Stephens, 2017; Gerard and Stephens, 2020).

So far, almost all existing treatments of the EBNM problem assume that the observations $\{X_j\}$ in (4.1) are independent given $\{\theta_j, s_j\}$. However, this assumption can be grossly violated in practice. Non-negligible correlations are common in real world data sets. Further, as we discuss later, EB approaches to the normal means problem are particularly vulnerable to being misled by pervasive correlations. Specifically, when the average strength of pairwise correlations among observations is non-negligible, the estimate $\hat{g}$ of g can be very inaccurate, and this adversely affects inference for *all* $\{\theta_j\}$. Ironically then, the attractive "borrowing strength" property of the EB approach becomes, in the presence of pervasive correlations, its Achilles' heel.

In this paper we introduce methods for solving the EBNM problem *allowing for unknown correlations* among the observations. More precisely, we re-write (4.1) as

$$X_j = \theta_j + s_j Z_j \ , \tag{4.3}$$

$$Z_j \sim N(0,1) \ , \tag{4.4}$$

and develop methods that allow for unknown correlations among the "noise" $\{Z_j\} := \{Z_1, \ldots, Z_p\}$. Our methods are built on elegant theory from Schwartzman (2010), who shows, in essence, that the limiting empirical distribution, f say, of correlated $N(0,1)$ random variables can be represented using a basis of the standard Gaussian density and its derivatives of increasing order. We use this result, combined with an assumption that $\{Z_j\}$ are exchangeable, to frame solving this "EBNM with correlated noise" problem as a two-step process: first *jointly estimate f and g* from all observations; and second compute the posterior distribution of θ_j given the estimated $\hat{f}, \hat{g}$ (and X_j, s_j). Although many possible assumptions on g are possible, here we assume g to be a scale mixture of zero-mean Gaussians, following the flexible "adaptive shrinkage" approach in Stephens (2017). We have implemented these methods in an `R` package, `cashr` ("correlated adaptive shrinkage in `R`"), available from `https://github.com/LSun/cashr`.

The rest of the paper is organized as follows. In Section 4.2, we illustrate how correlation can derail existing EBNM methods, and review Schwartzman (2010)'s representation of the empirical distribution of correlated $N(0,1)$ random variables. In Section 4.3 we introduce the exchangeable correlated noise (ECN) model, and describe methods to solve the EBNM with correlated noise problem. Section 4.4 provides numeric examples with realistic simulations and real data illustrations. Section 4.5 concludes and discusses future research directions. Technical details are provided in Appendix 4.6.

4.2 Motivation and Background

4.2.1 Correlation distorts empirical distribution and misleads EBNM methods

In essence, the reason correlation can cause problems for EBNM methods is that, even with large samples, the empirical distribution of correlated variables can be quite different from their marginal distribution (e.g. Efron, 2007a). To illustrate this, we generated realistic correlated $N(0,1)$ z-scores using a framework similar to Lu (2018); Gerard and Stephens (2019, 2020). Specifically, we took RNA-seq gene expression data on the 10^4 most highly expressed genes in 119 human liver tissues (The GTEx Consortium, 2015, 2017). In each simulation we randomly drew two groups of five samples (without replacement), and applied a standard RNA-seq analysis pipeline, using the software packages `edgeR` (Robinson et al., 2010), `voom` (Law et al., 2014), and `limma` (Ritchie et al., 2015), to compute, for each gene $j = 1, \ldots, 10^4$, an estimate of the $\log_2$-fold difference in mean expression, X_j, and a corresponding p-value, p_j, testing the null hypothesis that the difference in mean is 0. We converted p_j, X_j to a z-score $z_j := -\text{sign}(X_j)\Phi^{-1}(p/2)$, where Φ is the CDF of $N(0,1)$. We also computed an "effective" standard deviation $s_j := X_j/z_j$ for later use (Figure 4.2 and Section 4.4).

In these simulations, because samples are randomly assigned to the two groups, there are no genuine differences in mean expression. Therefore the z-scores should have marginal distribution $N(0,1)$. And, indeed, empirical checks confirm that the analysis pipeline produces well-calibrated marginally $N(0,1)$ z-scores (Appendix 4.6.1). However, the 10^4 z-scores in each simulated data set are correlated, due to correlations among genes, and such correlations can distort the empirical distribution so that it is very different from $N(0,1)$ (Efron, 2007a, 2010a,b). Figure 4.1 shows four examples, which were chosen to highlight some common patterns. Panels (a-c) all exhibit a feature we call *pseudo-inflation*, where the empirical

distribution is *more* dispersed than $N(0,1)$. Conversely, panel (d) exhibits *pseudo-deflation*, where the empirical distribution is *less* dispersed than $N(0,1)$. Panel (b) also exhibits skew.

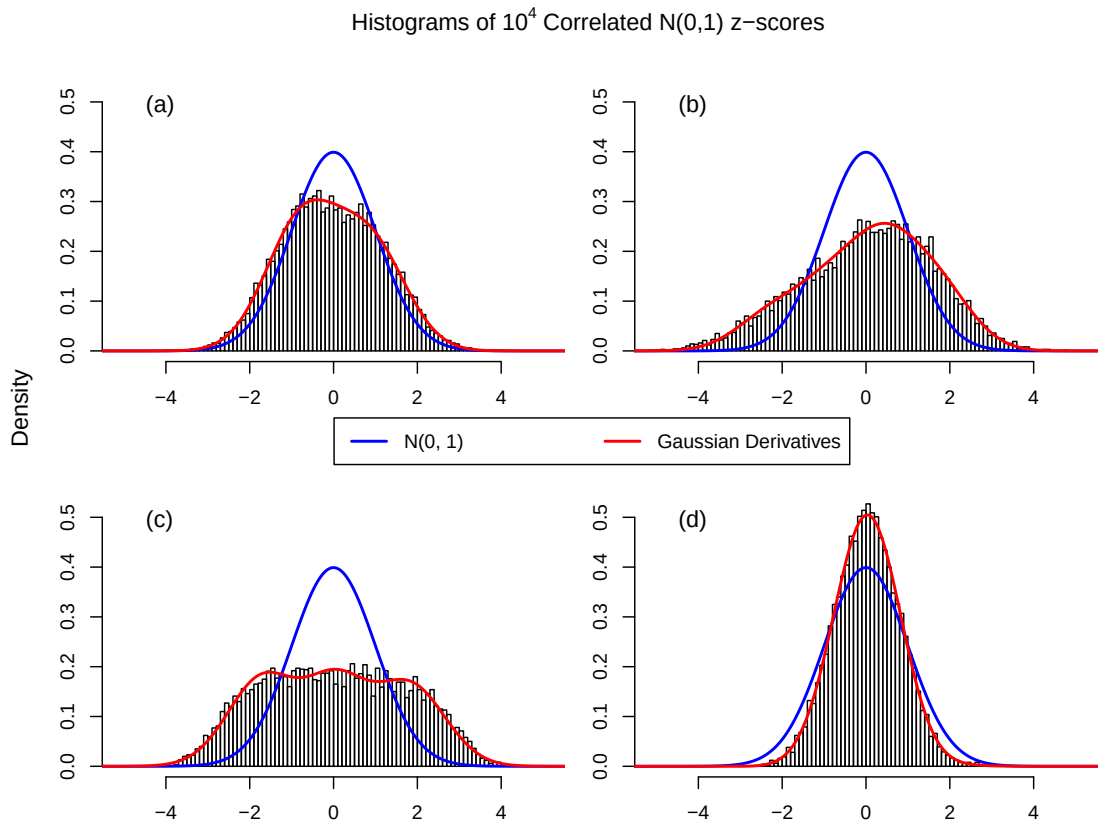

Figure 4.1: Illustration that the empirical distribution of a large number of correlated and marginally $N(0,1)$ null z-scores can deviate substantially from $N(0,1)$. The red lines are fitted densities obtained using our "Exchangeable Correlated Noise" model (Section 4.3.1) which uses a linear combination of the standard Gaussian density and its standardized derivatives.

Such *correlation-induced distortion* of the empirical distribution, if not appropriately addressed, can have serious consequence for EBNM methods. To illustrate this we applied several EBNM methods to five data sets simulated according to (4.2)-(4.4) as follows:

- The $p = 10^4$ normal means $\{\theta_j\}$ are iid samples from the mixture $g(\cdot) = 0.6\delta_0(\cdot) + 0.3N(\cdot; 0, 1) + 0.1N(\cdot; 0, 3^2)$, where $\delta_0(\cdot)$ denotes a point mass on 0 whose coefficient (0.6) is the null proportion, and $N(\cdot; \mu, \sigma^2)$ denotes the Gaussian density with mean μ

and variance σ^2. The same $\{\theta_j\}$ are used in all five data sets.

- In the first four data sets, the noise variables, $\{Z_j\}$, are the correlated null z-scores from the four panels of Figure 4.1. In the fifth data set $\{Z_j\}$ are iid $N(0,1)$ samples.
- The standard deviations $\{s_j\}$ are simulated from the RNA-seq gene expression data as described above, and $\{s_j\}$ are scaled to have $\frac{1}{p}\sum_j s_j^2 = 1$.

We provide the simulated $\{X_j, s_j\}$ values to four existing EBNM methods – `EbayesThresh` (Johnstone and Silverman, 2004, 2005a), `REBayes` (Koenker and Mizera, 2014; Koenker and Gu, 2017), `ashr` (Stephens, 2017), and `deconvolveR` (Efron, 2016; Narasimhan and Efron, 2016) – that all ignore correlation and assume independence among observations. (For `deconvolveR` we set $\{s_j\} \equiv 1$ as its current implementation supports only homoskedastic noise.)

The estimates of g obtained by each method are shown in Figure 4.2. All methods do reasonably well in the fifth data set where $\{Z_j\}$ are indeed independent (panel (e)). However, in the correlated data sets (panels (a-d)) the methods all misbehave in a similar way: over-estimating the dispersion of g under pseudo-inflation, and under-estimating it under pseudo-deflation. Their estimates of the null proportion are particularly inaccurate. These adverse effects are visible even when the distortion appears not severe (e.g. Figure 4.1(a)).

As a taster for what is to come, Figure 4.2 also shows the results from our new method, `cashr`, described later. This new method can account for both pseudo-inflation and pseudo-deflation, and in these examples estimates g consistently well.

4.2.2 Pseudo-inflation is non-Gaussian

In a series of pioneering papers (Efron, 2004, 2007a,b, 2008, 2010a), Efron studied the impact of correlations among z-scores on EB approaches to multiple testing. To account for the

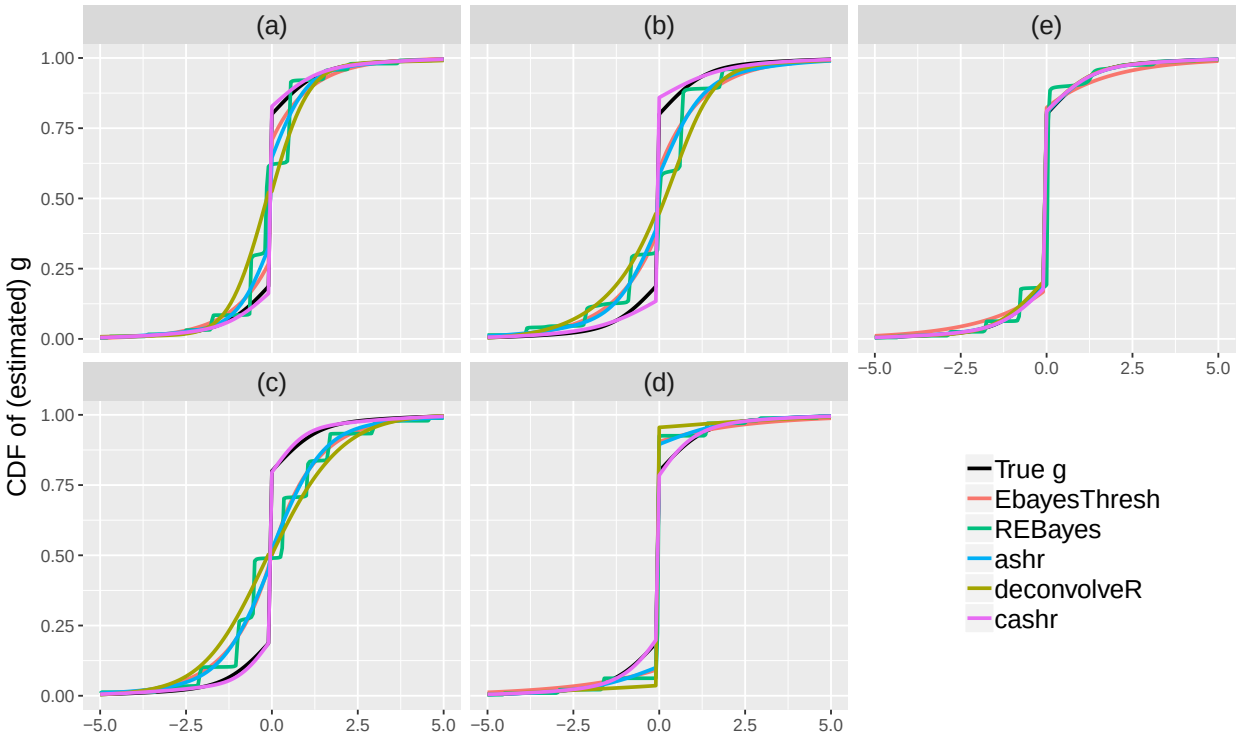

Figure 4.2: Illustration of how correlation can distort $\hat{g}$ estimated by EBNM methods. Each panel compares the true g with the estimated $\hat{g}$ from several EBNM methods applied to the same simulated dataset (see main text for details). In panels (a-d) $\{Z_j\}$ are the correlated null z-scores from the corresponding panels of Figure 4.1. In panel (e) $\{Z_j\}$ are iid $N(0,1)$ samples. Existing EBNM methods (`EbayesThresh`, `REBayes`, `ashr`, `deconvolveR`), which ignore correlation among observations, do reasonably well with iid noise (e). However they do much less well in the correlated cases (a-d): over-estimating the dispersion of g under pseudo-inflation (a-c) and under-estimating it under pseudo-deflation (d). In contrast, our new method `cashr` (Section 4.3) estimates g consistently well.

effects of correlation (pseudo-inflation, pseudo-deflation, and skew) on the empirical distribution of null z-scores he introduced the concept of an "empirical null." In his `locfdr` method (Efron, 2005), the empirical null is assumed to be Gaussian $N(\mu_0, \sigma_0^2)$. However, theory suggests that pseudo-inflation is not well modelled by a Gaussian distribution (Schwartzman, 2010, reviewed in Section 4.2.3), and a closer look at our empirical results here supports this conclusion.

To illustrate, Figure 4.3 shows more detailed analysis of the empirical distribution of

Figure 4.1(c) z-scores. The central part of this z-score distribution could perhaps be modelled by a Gaussian distribution with inflated variance – for example, it matches more closely to a $N(0, 1.6^2)$ than to $N(0, 1)$. However, in the tails, the empirical distribution has much shorter tails than $N(0, 1.6^2)$. For example, 10^4 iid $N(0, 1.6^2)$ samples would be expected to yield 43 p-values exceeding the Bonferroni threshold of $0.05/10^4$, whereas in fact we observe none here. In short, the effects of pseudo-inflation are primarily in the "shoulders" of the distribution, where $|z|$-scores are only moderately large, and not in the tails. (Incidentally, this behavior is far more evident in the histogram of z-scores than in the histogram of corresponding p-values, and we find the z-score histogram generally more helpful for diagnosing potential correlation-induced distortion.)

With hindsight this shoulder-but-not-tail inflation pattern should perhaps be expected. For example, Johnstone and Silverman (1997) illustrates that when p is large, the universal threshold, $\sqrt{2 \log p}$, should be exceedingly difficult for p standard Gaussian noise to pass, independent or otherwise. There are also relevant discussions on "asymptotic independence" in the extreme value theory literature (Sibuya, 1960; De Carvalho and Ramos, 2012). However, this property of pseudo-inflation does suggest that using a Gaussian to describe correlation-induced distortion, as in `locfdr`, is not ideal (more discussion in Section 4.4).

4.2.3 Empirical distribution of correlated standard normal noise

We now summarize an elegant result of Schwartzman (2010), which characterizes the empirical distribution of a large number of correlated $N(0, 1)$ z-scores. This result plays a key role in our work.

On notation: let φ denote the PDF of $N(0, 1)$, and $\varphi^{(l)}$ denote the lth derivative of φ. We refer to the collection of functions $\left\{\frac{1}{\sqrt{l!}}\varphi^{(l)}\right\}_{l=1}^{\infty}$ as the (standardized) Gaussian derivatives, where $\frac{1}{\sqrt{l!}}$ is used to standardize the (weighted) orthogonal Gaussian derivatives so that they are scaled to be orthonormal with respect to the weight function φ.

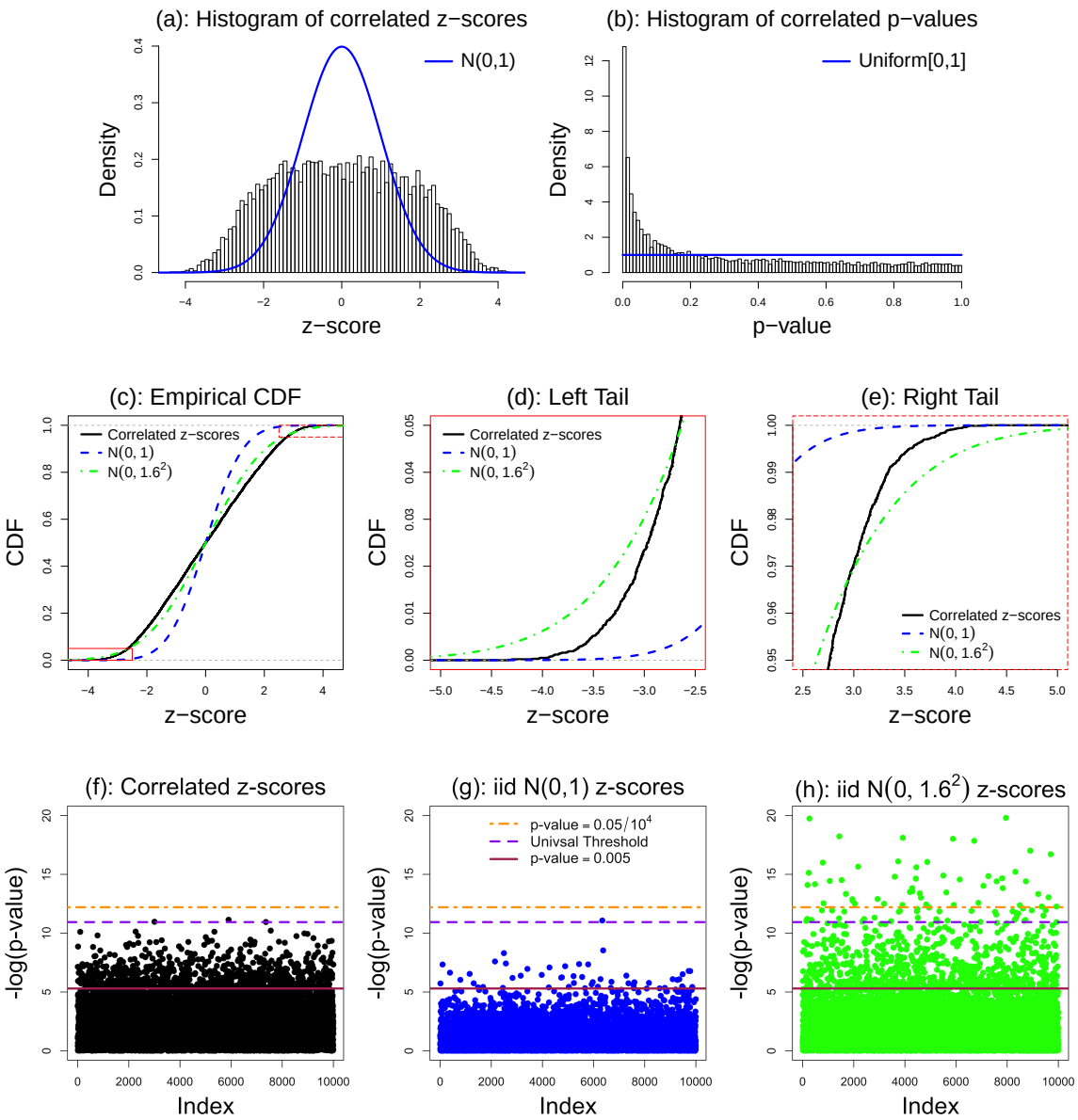

Figure 4.3: Illustration that the effects of pseudo-inflation are primarily in the "shoulders" of the distribution of null z-scores, and not in the tails. Panels (a-b): Histograms of correlated z-scores (from Figure 4.1(c)) and their corresponding p-values. Note that the "shoulder-but-not-tail" inflation is evident in the histogram of z-scores (a) but not in the oft-used histogram of p-values (b). Panels (c-e): Comparison of the empirical CDF of correlated z-scores with the CDF of $N(0,1)$ and $N(0,1.6^2)$. The z-score distribution is closer to $N(0,1.6^2)$ in the center, but closer to $N(0,1)$ in the tails. Panels (f-h): Comparison of correlated p-values with p-values obtained from 10^4 iid $N(0,1)$ and $N(0,1.6^2)$ z-scores. The number of correlated p-values ≤ 0.005 is closer to z-scores from $N(0,1.6^2)$, but the number in the extreme tail (e.g. clearing the Bonferroni or universal thresholds) is closer to $N(0,1)$.

Let $\{Z_j\} := \{Z_1, \ldots, Z_p\}$ be p identically distributed, but not necessarily independent, $N(0,1)$ random variables. Let F_p denote their empirical CDF:

$$F_p(\cdot) := \frac{1}{p}\sum_{j=1}^{p} \mathcal{I}(Z_j \leq \cdot) \ , \tag{4.5}$$

where the indicator function $\mathcal{I}(Z_j \leq \cdot) := \begin{cases} 1 & Z_j \leq \cdot \\ 0 & Z_j > \cdot \end{cases}$. Since $\{Z_j\}$ are random variables, F_p is a random function on $\mathbb{R} \to [0,1]$. Also, because $\{Z_j\}$ are marginally $N(0,1)$, the expectation of F_p is Φ.

Schwartzman (2010) studies the distribution of F_p, and how its deviation from the expectation Φ depends on the correlations among $\{Z_j\}$. Specifically, assuming that each pair $\{Z_i, Z_j\}$ is bivariate normal with correlation ρ_{ij} (which is weaker than the common assumption that all $\{Z_j\}$ are joint multivariate normal), Schwartzman (2010) provides the following representation of F_p when p is large:

$$F_p(\cdot) \approx F(\cdot) := \Phi(\cdot) + \sum_{l=1}^{\infty} W_l \frac{1}{\sqrt{l!}} \varphi^{(l-1)}(\cdot) \ , \tag{4.6}$$

where $W_1, W_2, \ldots$ are uncorrelated random variables with $E[W_l] = 0$, and

$$var(W_l) = \overline{\rho^l} := \frac{1}{p(p-1)} \sum_{i,j:i\neq j} \rho_{ij}^l \ . \tag{4.7}$$

Although uncorrelated, $W_1, W_2, \ldots$ are not independent; they must have higher-order dependence to guarantee that F is non-decreasing. Also here we assume $\overline{\rho^l} \geq 0$ for all $l \in \mathbb{N}$. Note that this assumption should not be too demanding for large p in practice (Schwartzman, 2010, also see Appendix 4.6.4).

Since F is a CDF, its derivative defines a corresponding density:

$$f(\cdot) := F'(\cdot) = \varphi(\cdot) + \sum_{l=1}^{\infty} W_l \frac{1}{\sqrt{l!}} \varphi^{(l)}(\cdot) \ . \tag{4.8}$$

Intuitively, (4.8) characterizes how the (limiting) empirical distribution (histogram) of $\{Z_j\}$ is likely to randomly deviate from the expectation φ, using standardized Gaussian derivatives as basis functions.

The representation (4.8) is crucial to our work here, and provides some remarkable insights. We highlight particularly the following:

1. The expected deviations of f from φ are determined by the variances of the coefficients W_l, which are determined by the mean and higher moments of the pairwise correlations, $\overline{\rho^l}$. In the special case where $\{Z_j\}$ are independent all these terms are 0, and $f = \varphi$.

2. Following from 1, to create a tangible deviation from φ, $\overline{\rho^l}$ must be non-negligible (for some l). This requires *pervasive, but not necessarily strong,* pairwise correlations. For example, pervasive correlations occur if there is an underlying low-rank structure in the data, where all $\{Z_j\}$ are influenced by a small number of underlying random factors, and so are all correlated with one another. In this case $\overline{\rho^l}$ will be non-negligible, and f may deviate from φ. In contrast, there can exist very strong pairwise correlations with negligible effect on f. For example, suppose p is even, and let $\{Z_j\}$ be in $p/2$ pairs, with each pair having correlation one but different pairs being independent. The histogram of $\{Z_j\}$ will look very much like $N(0,1)$, because $\overline{\rho^l} = \frac{1}{p-1} \approx 0$ for large p. In other words, not all kinds of correlations, even large ones, distort the empirical distribution of $\{Z_j\}$.

3. Barring special cases such as $\rho_{ij} = 1$, the moments $\overline{\rho^l}$, and hence the expected magnitude of W_l, will decay quickly with l. Consequently the sum in (4.8) will typically be dominated by the first few terms, and the shape of the first few basis functions

will determine the typical deviation of f from φ. Of the first four basis functions (Figure 4.4), the 2nd and 4th correspond to pseudo-inflation or pseudo-deflation in the shoulders of φ, depending on the signs of their coefficients, whereas the 1st and 3rd correspond to mean shift and skewness. This explains the empirical observation that correlation-induced pseudo-inflation tends to focus in the shoulders, and not the tails. (Also see Appendix 4.6.2 for the special case $\rho_{ij} = 1$.)

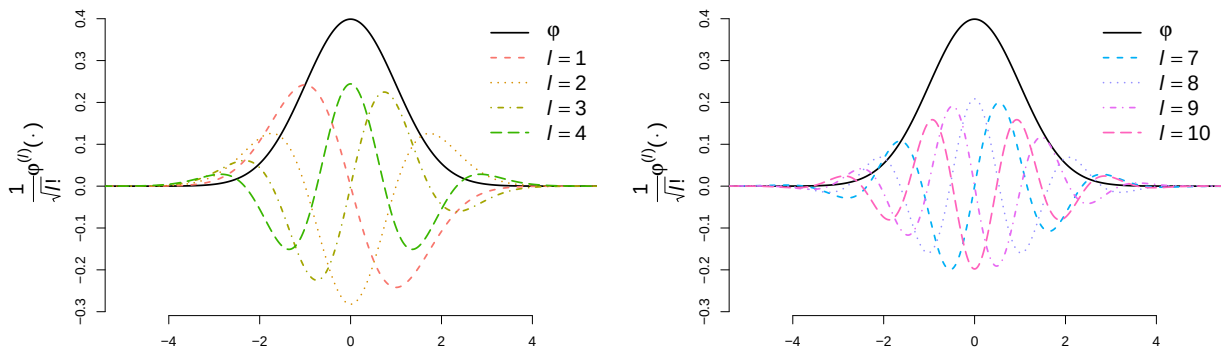

Figure 4.4: Illustration of the standard Gaussian density, φ, and its standardized derivatives. The left panel shows φ and its first four standardized derivatives. The 2nd and 4th derivatives correspond to pseudo-inflation or pseudo-deflation in the shoulders; the 1st and 3rd derivatives correspond to mean shift or skewness. The right panel shows φ and its 7th-10th derivatives. Even for these higher-order derivatives, tails are short, implying that correlation-induced distortion is unlikely to have long tails.

In discussing Efron (2010a), Schwartzman (2010) used this result to argue that "a wide unimodal histogram (of z-scores) may be indication of the presence of true signal, rather than an artifact of correlation." Specifically, by discarding terms for $l \geq 4$ in (4.8), he found that the largest central spread (standard deviation) for f in (4.8) being a unimodal density is approximately 1.3. Along similar lines, we can show (Appendix 4.6.2) that the maximum standard deviation for f being a Gaussian density is $\sqrt{2} \approx 1.4$. The key point here is that the effects of correlation are different from the effects of true signals, so the two can (often) be separated. Our methods here are designed to do exactly that.

4.3 Empirical Bayes Normal Means with Correlated Noise

4.3.1 The Exchangeable Correlated Noise model

As a first step towards allowing for correlated noise in the EBNM problem, we develop methods to fit the representation (4.8) to correlated null z-scores. We do this by treating $\{Z_j\}$ as conditionally iid samples from f in (4.8), parameterized by $\omega := \{\omega_1, \omega_2, \ldots\}$ which are realizations of $W := \{W_1, W_2, \ldots\}$:

$$Z_j \mid \{W = \omega\} \overset{\text{iid}}{\sim} f(\cdot; \omega) := \varphi(\cdot) + \sum_{l=1}^{\infty} \omega_l \frac{1}{\sqrt{l!}} \varphi^{(l)}(\cdot) \ . \tag{4.9}$$

It may seem perverse to model correlated random variables as conditionally iid. However, this treatment can be motivated by assuming $\{Z_j\}$ are exchangeable and appealing to de Finetti's representation theorem (De Finetti, 1937), which says that (infinitely) exchangeable random variables can be represented as being conditionally iid from their empirical distribution. We therefore refer to the model (4.9) as the *exchangeable correlated noise (ECN)* model. We also refer to f as the *correlated noise distribution*.

To fit the ECN model (4.9) with observed $\{Z_j\}$, we estimate ω, essentially by maximum likelihood, but with a couple of additional complications that we now describe. First, since f is a density, we must constrain the parameters ω to ensure that $f(\cdot; \omega)$ is non-negative (note that (4.9) integrates to one for any ω, but is not guaranteed to be non-negative). Ideally f should be non-negative on the whole real line, but this constraint is difficult to work with, so we approximate it using a discrete approximation: we constrain $f(\mathfrak{z}_i; \omega) \geq 0$ on a fine grid $\{\mathfrak{z}_1, \ldots, \mathfrak{z}_m\}$ such as $\{-10, -9.999, -9.998, \ldots, +9.998, +9.999, +10\}$, in addition to $f(Z_j; \omega) \geq 0$ for all j.

Second, to incorporate the prior expectation that the absolute value of ω_l should decay

quickly with l (because $var(W_l) = \overline{\rho^l}$) we introduce a penalty on ω,

$$h(\omega) := \sum_l \gamma_l |\omega_l|, \tag{4.10}$$

where we take the penalty parameters γ_l to be

$$\gamma_l = \begin{cases} 0 & l \text{ is odd} \\ \gamma/\rho^{l/2} & l \text{ is even} \end{cases}, \tag{4.11}$$

where γ represents a common penalty, and ρ some notion of average pairwise correlation. For computational convenience we use only the first $L = 10$ Gaussian derivatives (see Figure 4.4 for 7th-10th standardized Gaussian derivatives) and set $\omega_l = 0$ for $l > 10$. (Recall that $var(W_l) = \overline{\rho^l}$, so W_l's realization ω_l will generally be negligible in practice for $l > 10$. This assumption also matches our empirical results using correlated $\{Z_j\}$ created from real data.) Of course a full Bayesian treatment would attempt to account for uncertainty in ω; in ignoring that here we are making the usual EB compromise.

In numerical simulations, we experimented with combinations of $\gamma \in \{1, 5, 10, 50, 100\}$ and $\rho \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, and found that $\gamma = 10, \rho = 0.5$ performed well in a variety of situations, although results were not very sensitive to the choice of γ and ρ. All results in this paper were obtained with $\gamma = 10, \rho = 0.5$.

In summary, we estimate ω by solving:

$$\begin{aligned} \max_{\omega} \quad & \textstyle\sum_j \log f(Z_j; \omega) - h(\omega) \\ \text{s.t.} \quad & f(Z_j; \omega) \geq 0, && j = 1, \ldots, p\ , \\ & f(\mathfrak{z}_i; \omega) \geq 0, && i = 1, \ldots, m\ . \end{aligned} \tag{4.12}$$

This is a convex optimization and can be solved efficiently and stably using an interior point method; we implemented this using the `R` package `Rmosek` to interface to the MOSEK

commercial solver (MOSEK ApS, 2018). With $p = 10^4$, the problem is solved on average within 0.50 seconds on a personal computer (Apple iMac, 3.2 GHz, Intel Core i5).

Figure 4.1 shows the fitted distributions from the ECN model,

$$\hat{f}(\cdot;\hat{\omega}) := \varphi(\cdot) + \sum_{l=1}^{L} \hat{\omega}_l \frac{1}{\sqrt{l!}} \varphi^{(l)}(\cdot) \ , \tag{4.13}$$

on the four illustrative sets of correlated null z-scores.

4.3.2 The EBNM model with correlated noise

To allow for correlated noise in the EBNM problem, we combine the standard EBNM model (4.2)-(4.4) with the ECN model (4.9):

$$X_j = \theta_j + s_j Z_j \tag{4.14}$$

$$\theta_j \sim g(\cdot) \tag{4.15}$$

$$Z_j \mid \omega \sim f(\cdot;\omega) = \varphi(\cdot) + \sum_{l=1}^{L} \omega_l \frac{1}{\sqrt{l!}} \varphi^{(l)}(\cdot) \ . \tag{4.16}$$

Note that in this model the observations are conditionally independent given f and g.

Following Stephens (2017) we model the prior distribution g by a finite mixture of zero-mean Gaussians:

$$g(\cdot;\pi) = \pi_0 \delta_0(\cdot) + \sum_{k=1}^{K} \pi_k N(\cdot; 0, \sigma_k^2) \ , \tag{4.17}$$

where π_0 is the null proportion. Here the mixture proportions $\pi := \{\pi_0, \pi_1, \ldots, \pi_K\}$ are non-negative and sum to 1, and are to be estimated, whereas the component standard deviations $\sigma_1 < \sigma_2 < \cdots < \sigma_K$ are a fixed pre-specified grid of values. By using a sufficiently wide and dense grid of standard deviations this finite mixture can approximate, to arbitrary accuracy, any scale mixture of zero-mean Gaussians.

The marginal log-likelihood for π, ω, integrating out $\{\theta_j\}$, $\{Z_j\}$, is given by the following Theorem.

Theorem 3. *Combining* (4.14)-(4.17)*, the marginal log-likelihood of* π, ω *is*

$$L(\pi, \omega) := \log\left(\prod_{j=1}^{n} p(X_j|\pi, \omega)\right) = \sum_{j=1}^{n} \log\left(\sum_{k=0}^{K} \pi_k \left(p_{jk0} + \sum_{l=1}^{L} \omega_l p_{jkl}\right)\right), \tag{4.18}$$

where

$$p_{jkl} = \frac{s_j^l}{\sqrt{\sigma_k^2 + s_j^2}^{l+1}} \frac{1}{\sqrt{l!}} \varphi^{(l)}\left(\frac{X_j}{\sqrt{\sigma_k^2 + s_j^2}}\right). \tag{4.19}$$

Proof. The marginal distribution of X_j, denoted as $p(X_j)$, is obtained by integrating out θ_j

$$\begin{aligned}
&p(X_j) \\
&= \int_{\mathbb{R}} g(\theta_j) p(X_j|\theta_j, s_j) \mathrm{d}\theta_j = \int_{\mathbb{R}} g(\theta_j) \frac{1}{s_j} f\left(\frac{X_j - \theta_j}{s_j}\right) \mathrm{d}\theta_j \\
&= \int_{\mathbb{R}} \left[\pi_0 \delta_0(\theta_j) + \sum_{k=1}^{K} \frac{\pi_k}{\sigma_k} \varphi\left(\frac{\theta_j}{\sigma_k}\right)\right] \left[\frac{1}{s_j} \varphi\left(\frac{X_j - \theta_j}{s_j}\right) + \frac{1}{s_j} \sum_{l=1}^{L} \frac{\omega_l}{\sqrt{l!}} \varphi^{(l)}\left(\frac{X_j - \theta_j}{s_j}\right)\right] \mathrm{d}\theta_j \\
&= \sum_{k=0}^{K} \pi_k \left(p_{jk0} + \sum_{l=1}^{L} \omega_l p_{jkl}\right), \qquad (4.20)
\end{aligned}$$

where $p_{jkl} = \int_{\mathbb{R}} \frac{1}{\sigma_k} \varphi\left(\frac{\theta_j}{\sigma_k}\right) \frac{1}{s_j} \frac{1}{\sqrt{l!}} \varphi^{(l)}\left(\frac{X_j - \theta_j}{s_j}\right) \mathrm{d}\theta_j$ is essentially a convolution of φ and $\varphi^{(l)}$ and has an analytic form

$$p_{jkl} = \frac{s_j^l}{\sqrt{\sigma_k^2 + s_j^2}^{l+1}} \frac{1}{\sqrt{l!}} \varphi^{(l)}\left(\frac{X_j}{\sqrt{\sigma_k^2 + s_j^2}}\right).$$

This form is also valid for $k = 0$, $l = 0$. Following (4.20), the marginal log-likelihood of π, ω

is given by

$$\log\left(\prod_{j=1}^{n} p\left(X_j\right)\right) = \sum_{j=1}^{n} \log\left(\sum_{k=0}^{K} \pi_k \left(p_{jk0} + \sum_{l=l}^{L} \omega_l p_{jkl}\right)\right) . \tag{4.21}$$

□

4.3.3 Fitting the model

Following the usual EB approach, we fit the model (4.14)-(4.17) in two steps, first estimating g, f by estimating π, ω and then basing inference for $\{\theta_j\}$ on the (estimated) posterior distribution $p(\{\theta_j\}|\{X_j, s_j\}, \hat{\pi}, \hat{\omega})$. Note that under the model (4.14)-(4.17) $\theta_1, \ldots, \theta_p$ are conditionally independent given $f, g, \{X_j, s_j\}$, so this posterior distribution $p(\{\theta_j\}|\{X_j, s_j\}, \hat{\pi}, \hat{\omega})$ factorizes, and is determined by its marginal distributions $p(\theta_j|X_j, s_j, \hat{\pi}, \hat{\omega})$. The intuition here is that, under the exchangeability assumption, the effects of correlation are captured entirely by the (realized) correlated noise distribution f. Once this distribution is estimated, the inferences for each θ_j become independent, just as in the standard EBNM problem.

The usual EBNM approach to estimating π, ω would be to maximize the likelihood $L(\pi, \omega)$. Here we modify this approach using maximum penalized likelihood. Specifically we use the penalty on ω as in (4.10), and the penalty on π used by Stephens (2017) to encourage conservative (over-)estimation of the null proportion π_0 (to induce conservative estimation of false discovery rates). Thus, we solve

$$\hat{\pi}, \hat{\omega} = \arg\max_{\pi,\omega} \sum_{j=1}^{n} \log\left(\sum_{k=0}^{K} \pi_k \left(p_{jk0} + \sum_{l=1}^{L} \omega_l p_{jkl}\right)\right) + \sum_{k=0}^{K} \lambda_k \log(\pi_k) - \sum_{l=1}^{L} \gamma_l |\omega_l| \tag{4.22}$$

subject to the constraints

$$\sum_{k=0}^{K} \pi_k = 1 \tag{4.23}$$

$$\pi_k \geq 0, \quad k = 0, 1, \ldots, K \tag{4.24}$$

$$\varphi\left(\mathfrak{z}_i\right) + \sum_{l=1}^{L} \omega_l \frac{1}{\sqrt{l!}} \varphi^{(l)}(\mathfrak{z}_i) \geq 0, \quad i = 1, \ldots, m\ . \tag{4.25}$$

In (4.25) we used the same device as in (4.12) to capture non-negativity of f. We set γ_l as in (4.11), use only the first $L = 10$ Gaussian derivatives, and set $\lambda_0 = 10$, $\lambda_1 = \cdots = \lambda_K = 0$ as in Stephens (2017).

Problem (4.22) is biconvex. That is, given a feasible $\hat{\pi}$, the optimization over ω is convex; and given a feasible $\hat{\omega}$, the optimization over π is convex. The optimization over π can be solved using the EM algorithm, or more efficiently using convex optimization methods (Koenker and Mizera, 2014; Koenker and Gu, 2017; Kim et al., 2018). To optimize over ω we use the same approach as in solving (4.12). To solve (4.22) we simply iterate between these two steps until convergence. Although the joint optimization problem is not convex, our empirical results from a diverse set of realistic simulations show that the convergence is usually efficient and satisfactory.

4.3.4 Posterior calculations

For each j, the posterior distribution $p(\theta_j \mid X_j, \hat{\pi}, \hat{\omega})$ is, by Bayes Theorem, given by

$$p(\theta_j \mid X_j, \hat{\pi}, \hat{\omega}) = \frac{\left[\hat{\pi}_0 \delta_0 + \sum\limits_{k=1}^{K} \hat{\pi}_k N(\theta_j; 0, \sigma_k^2)\right] \left[\frac{1}{s_j} \varphi\left(\frac{X_j - \theta_j}{s_j}\right) + \sum\limits_{l=1}^{L} \hat{\omega}_l \frac{1}{s_j} \frac{1}{\sqrt{l!}} \varphi^{(l)}\left(\frac{X_j - \theta_j}{s_j}\right)\right]}{\sum\limits_{k=0}^{K} \hat{\pi}_k \left(p_{jk0} + \sum\limits_{l=1}^{L} \hat{\omega}_l p_{jkl}\right)}. \tag{4.26}$$

Despite the somewhat complex form, some important functionals of this posterior distribution are analytically available.

1. The posterior mean for θ_j

$$E[\theta_j \mid X_j, \hat{\pi}, \hat{\omega}] = \frac{\sum_{k=0}^{K} \hat{\pi}_k \left(m_{jk0} + \sum_{l=1}^{L} \hat{\omega}_l m_{jkl} \right)}{\sum_{k=0}^{K} \hat{\pi}_k \left(p_{jk0} + \sum_{l=1}^{L} \hat{\omega}_l p_{jkl} \right)} , \tag{4.27}$$

   where $m_{jkl} = -\frac{s_j^l \sigma_k^2}{\sqrt{\sigma_k^2 + s_j^2}^{\,l+2}} \frac{1}{\sqrt{l!}} \varphi^{(l+1)} \left(\frac{X_j}{\sqrt{\sigma_k^2 + s_j^2}} \right)$.

2. The local FDR (lfdr; Efron, 2008) is

$$\text{lfdr}_j := \Pr(\theta_j = 0 \mid X_j, \hat{\pi}, \hat{\omega}) = \frac{\hat{\pi}_0 \frac{1}{s_j} \varphi \left(\frac{X_j}{s_j} \right) + \sum_{l=1}^{L} \hat{\omega}_l \frac{1}{s_j} \frac{1}{\sqrt{l!}} \varphi^{(l)} \left(\frac{X_j}{s_j} \right)}{\sum_{k=0}^{K} \hat{\pi}_k \left(p_{jk0} + \sum_{l=1}^{L} \hat{\omega}_l p_{jkl} \right)} . \tag{4.28}$$

   From this, the FDR of any discovery set $\Gamma \subseteq \{1, \ldots, n\}$ can be estimated as

$$\widehat{\text{FDR}}(\Gamma) = \frac{1}{|\Gamma|} \sum_{j \in \Gamma} \text{lfdr}_j , \tag{4.29}$$

   where $|\Gamma|$ denotes the number of elements in Γ. Storey's q-value (Storey, 2003) for each j is defined as

$$q_j := \widehat{\text{FDR}}(\{k : \text{lfdr}_k \leq \text{lfdr}_j\}) . \tag{4.30}$$

3. Stephens (2017) introduced the term "local false sign rate (lfsr)" to refer to the probability of getting the sign of an effect wrong, as well as the false sign rate (FSR) and the s-value, analogous to the local FDR, the FDR, and the q-value, respectively. Making statistical inference about the sign of a parameter, rather than solely focusing on

whether the parameter being zero or not, was also discussed in Tukey (1991); Gelman et al. (2012). The value of lfsr_j is defined as

$$\text{lfsr}_j := \min\{\Pr(\theta_j \geq 0 \mid X_j, \hat{\pi}, \hat{\omega}),\ \Pr(\theta_j \leq 0 \mid X_j, \hat{\pi}, \hat{\omega})\}\ , \tag{4.31}$$

which is easily calculated from lfdr_j and

$$\Pr(\theta_j > 0 \mid X_j, \hat{\pi}, \hat{\omega}) = \frac{\sum_{k=1}^{K} \hat{\pi}_k \left(\hat{\tau}_{jk0} + \sum_{l=1}^{L} \hat{\omega}_l \tau_{jkl} \right)}{\sum_{k=0}^{K} \hat{\pi}_k \left(p_{jk0} + \sum_{l=1}^{L} \hat{\omega}_l p_{jkl} \right)}\ , \tag{4.32}$$

where $\tau_{jkl} = \frac{s_j^l/\sqrt{l!}}{\sqrt{s_j^2+\sigma_k^2}^{\,l+1}} \left[\sum_{m=0}^{l} \binom{l}{m} \left(\frac{\sigma_k}{s_j}\right)^m \varphi^{(m-1)} \left(\frac{X_j}{\sqrt{s_j^2+\sigma_k^2}} \frac{\sigma_k}{s_j} \right) \varphi^{(l-m)} \left(\frac{X_j}{\sqrt{s_j^2+\sigma_k^2}} \right) \right]$.
The FSR and s-value are estimated and defined similarly to the FDR and q-value as

$$\widehat{\text{FSR}}(\Gamma) = \frac{1}{|\Gamma|} \sum_{j \in \Gamma} \text{lfsr}_j\ , \qquad s_j := \widehat{\text{FSR}}(\{k : \text{lfsr}_k \leq \text{lfsr}_j\})\ . \tag{4.33}$$

4.3.5 Software

We implemented both the fitting procedure and posterior calculations in an `R` package `cashr` which is available at `https://github.com/LSun/cashr`. For $p = 10^4$, it takes on average about 6 seconds for model fitting and posterior calculations on a personal computer (Apple iMac, 3.2 GHz, Intel Core i5).

4.4 Numerical Results

We now empirically assess the performance of `cashr` on both simulated and real data. We focus our assessments on the "multiple testing" setting where $\{\theta_j\}$ is sparse and the main goal is to identify "significant" non-zero elements θ_j. This problem can be tackled using EB

methods (Thomas et al., 1985; Greenland and Robins, 1991) and here we compare `cashr` with both `locfdr` (Efron, 2005), which attempts to capture effects of correlation through an empirical null strategy discussed in Section 4.2.2, and `ashr` (Stephens, 2017), which fits the same EBNM model as `cashr` but without allowing for correlation – i.e. `ashr` is equivalent to setting $f = \varphi$ in (4.16). Multiple testing can also be tackled by attempting to control the FDR in the frequentist sense, and so we also compare with the Benjamini-Hochberg procedure (BH; Benjamini and Hochberg, 1995) and `qvalue` (Storey, 2002, 2003). One advantage of the EBNM approach to multiple testing is that it can provide not only FDR assessments, but also point estimates and interval estimates for the effects $\{\theta_j\}$ (Stephens, 2017). However, to keep our comparisons simple we focus here only on FDR assessments.

4.4.1 Realistic simulation with gene expression data

We constructed synthetic data with realistic correlation structure using the simulation framework in Section 4.2.1. The data are simulated according to the EBNM with correlated noise model (4.2)-(4.4) as follows.

- The $p = 10^4$ normal means $\theta_1, \ldots, \theta_p$ are iid samples from

$$g(\cdot) = \pi_0 \delta_0(\cdot) + (1 - \pi_0) g_1(\cdot) \ , \tag{4.34}$$

  for six choices of g_1 and three choices of $\pi_0 \in \{0.5, 0.9, 0.99\}$ (Figure 4.5). The density functions of these six choices of g_1 and other simulation details are in Appendix 4.6.3.

- To make the correlation structure among noise realistic, in each simulation $\{Z_j\}$ are simulated from real gene expression data as in Section 4.2.1.

- The standard deviations $\{s_j\}$ are also simulated from real gene expression data using the same pipeline, and are scaled to have $\frac{1}{p}\sum s_j^2 = 1$.

- The observations are constructed as $X_j = \theta_j + s_j Z_j$, $j = 1, \ldots, p$.

In each simulated data set, this framework generates p correlated observations $\{X_j\}$ of respective normal means $\{\theta_j\}$ with corresponding standard deviations $\{s_j\}$. The data $\{(X_1, s_1), \ldots, (X_p, s_p)\}$ are made available to each method, while the effects $\{\theta_j\}$ are withheld. The analysis goal is to identify which θ_j are significantly different from 0. We applied each method to formulate a discovery set at nominal FDR $= 0.1$, and calculated the empirical false discovery proportion (FDP) for each discovery set. We ran 1000 simulations for each g_1, divided evenly among the three choices of π_0.

Figure 4.5 compares the performance of each method in these simulations. Our first result is that, despite the presence of correlation, most of the methods control FDR in the usual frequentist sense under most scenarios: that is, the mean FDP is usually below the nominal level of 0.1. Indeed, BH is notable in never showing a mean FDP exceeding the nominal level, even though, as far as we are aware, no known theory guarantees this under the realistic patterns of correlation used here (Benjamini and Yekutieli (2001) gives relevant theoretical results under more restrictive assumptions on the correlation). The method most prone to lose control is `ashr`, but even its mean FDP is never above 0.2.

However, despite this frequentist control of FDR, for most methods the FDP for individual data sets can often lie far from the nominal level (see also Owen, 2005; Qiu et al., 2005; Blanchard and Roquain, 2009; Friguet et al., 2009, for example). Arguably, then, frequentist control of FDR is insufficient in practice, since we desire – as far as is possible – to make sound statistical inference for each data set. That is, we might consider a method to perform well if its FDP is consistently close to the nominal level, rather than close on average. By this criterion, `cashr` consistently outperforms other methods (Figure 4.5): it provides uniformly lower root MSE of FDP from the nominal FDR, 0.1, and the whiskers in the boxplots (indicating 5th and 95th percentiles) are narrower. Along with FDP, Figure 4.5 also shows the empirical true discovery proportion (TDP), defined as the proportion of

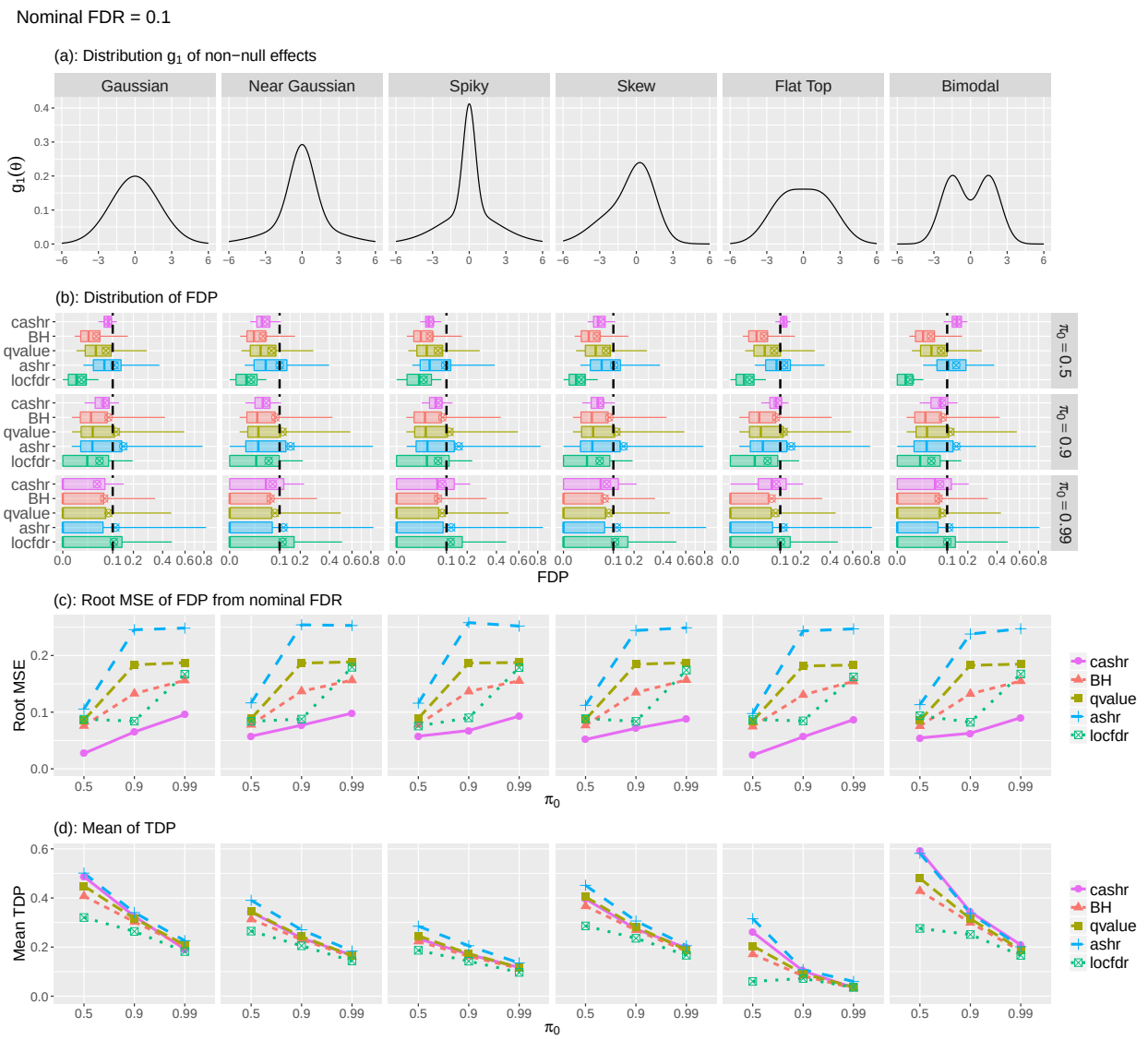

Figure 4.5: Illustration that `cashr` outperforms other methods in producing discovery sets whose FDP are consistently close to the nominal FDR, while maintaining good statistical power. Simulation results are shown for six different distributions for the non-null effect (g_1; panel (a)) and three different values of the null proportion ($\pi_0 \in \{0.5, 0.9, 0.99\}$), stratified by methods. Panel (b): Comparison of the distribution of FDP, summarized as boxplots on square-root scale. The boxplots show the mean (cross), median (line), inter-quartile ranges (box), and 5th and 95th percentiles (whiskers). Panel (c): Comparison of the root MSE of FDP from the nominal FDR of 0.1, defined as $\sqrt{\text{mean}[(\text{FDP} - 0.1)^2]}$. In all scenarios the distribution of FDP for `cashr` is more concentrated near the nominal 0.1 level than other methods. Especially, the root MSE of FDP for `cashr` is uniformly lower than other methods. Panel (d): Comparison of the mean of TDP, as an indication of statistical power. On average, `cashr` maintains good power, only worse than `ashr` in some scenarios, which sometimes finds more true signals at the cost of severely losing control of FDP.

true discoveries out of the number of all non-zero θ_j, as an indication of statistical power. `cashr` maintains good power in that it produces higher TDP than most methods in most scenarios. In some scenarios, `ashr` sometimes finds more true signals than `cashr`, but at the cost of severely losing control of FDP.

We note that `cashr` performs well even in settings that do not fully satisfy its underlying assumptions (e.g. where g_1 is asymmetric or multimodal). Note also that for our choices of g_1, $\pi_0 = 0.99$ is a highly sparse setting, as a large portion of the non-zero θ_j are close to zero. For example, when g_1 is Gaussian, only about 3 out of 10^4 $|\theta_j|$ are expected to be larger than $\sqrt{2 \log p} \approx 4.3$. Therefore, it is understandable that no methods perform particularly well in this difficult setting. But even for this $\pi_0 = 0.99$ setting, although first impressions from the plot may be that `cashr` and BH perform similarly, closer visual inspection shows `cashr` to be better, in that its median FDP tends to be closer to 0.1.

The reason that `cashr` produces more consistently reliable FDP is that, by design, it adapts itself to the particular correlation-induced distortion present in each data set. As illustrated in Figure 4.1, correlation can lead to pseudo-inflation in some data sets and pseudo-deflation in others. `cashr` is able to recognize which pattern is present, and correspondingly modify its behavior – becoming more conservative in the former case and less conservative in the latter. This is illustrated in Figure 4.6, which stratifies the realized data sets according to sample standard deviation of the realized correlated $N(0, 1)$ noise $\{Z_j\}$ in each data set (for the setting where g_1 is Gaussian, $\pi_0 = 0.9$). The bottom 1/3 are categorized as pseudo-deflation, top 1/3 pseudo-inflation, and the others "in-between."

For data sets where $\{Z_j\}$ show no strong distortion ("in-between") all methods give similar and reasonable results, with `cashr` showing only a small improvement. However, when $\{Z_j\}$ are pseudo-inflated, methods ignoring correlation, such as BH, `qvalue`, `ashr`, tend to be anti-conservative; that is, they form discovery sets whose FDP are often much larger than the nominal FDR. In contrast, `cashr` produces conservative FDP near the nominal value;

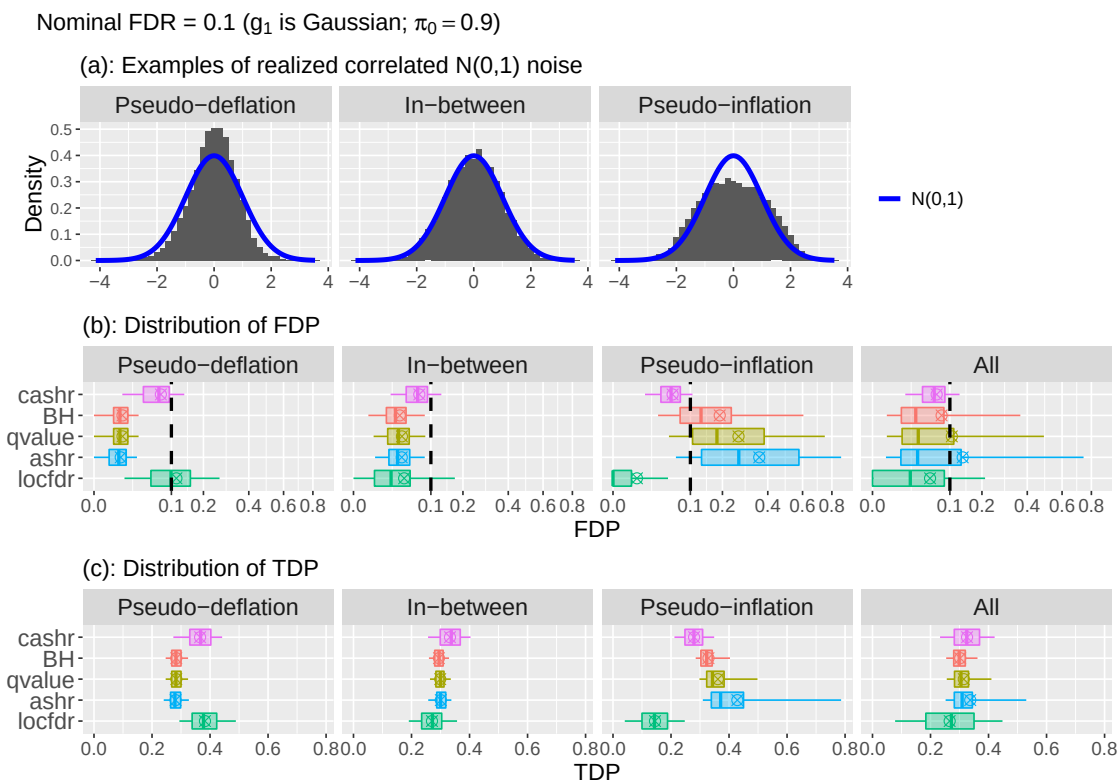

Figure 4.6: Illustration that `cashr` consistently produces reliable FDP under different types of correlation-induced distortion. Here we take the results from a single simulation scenario (g_1 is Gaussian, $\pi_0 = 0.9$) and stratify them into three groups of equal size according to the sample standard deviations of the realized correlated $N(0, 1)$ noise. Methods that ignore correlations among observations (BH, `qvalue`, `ashr`) are generally too conservative under pseudo-deflation and too anti-conservative under pseudo-inflation; `locfdr` tends to be too conservative under pseudo-inflation and consequently lose power; `cashr` maintains good FDR control in all settings. The boxplots show the mean (cross), median (line), inter-quartile ranges (box), and 5th and 95th percentiles (whiskers). FDP are plotted on square-root scale. Other choices of g_1 and π_0 give qualitatively similar results (not shown here).

and `locfdr` is too conservative, consequently losing substantial power (discussed further in Section 4.4.2). Conversely, with pseudo-deflation, methods ignoring correlation are too conservative, producing FDP much smaller than the nominal FDR, losing power compared with `cashr` and `locfdr`.

4.4.2 Real data illustrations

We now use two real data examples to illustrate some of the features of `cashr` (and other methods) that we observed in simulated data. The first example is a well-studied data set from a leukemia study (Golub et al., 1999), comparing gene expression in 47 acute myeloid leukemia vs 25 acute lymphoblastic leukemia samples, which was discussed extensively in Efron (2010a) as a prime example of how correlation can distort empirical distributions. The second example comes from a study on embryonic mouse hearts (Smemo, 2012), comparing gene expression in 2 left ventricle samples vs 2 right ventricle samples. (The number of samples is small, but each sample is a pool of ventricles from 150 mice – necessary to obtain sufficient tissue for the experiments to work well – and so this experiment involved dissection of 300 mouse hearts.)

For each data set we let θ_j denote the true $\log_2$-fold change in gene expression between the two groups for gene j. We use a standard analysis protocol (based on Smyth (2004); see Appendix 4.6.3 for details) to obtain an estimate X_j for θ_j, and a corresponding p-value p_j. As in Section 4.2.1, we convert the p-value to the corresponding z-score z_j and use this to compute an effective standard deviation s_j.

Figure 4.7 shows the empirical distribution of the z-scores for each data set, together with the fitted correlated noise distribution from `cashr` and the fitted empirical null from `locfdr`. In both cases the histogram is substantially more dispersed than $N(0, 1)$. However the two data sets have otherwise quite different patterns of inflation: the leukemia data show inflation in both the shoulders and tails of the distribution, whereas the mouse data show inflation only in the shoulders. This indicates the presence of some strong signals in the leukemia data, whereas the inflation in the mouse data may be primarily pseudo-inflation caused by correlation. Consistent with this, both `locfdr` and `cashr` identify hundreds of significant signals in the leukemia data (at nominal FDR = 0.1), but no significant signals in the mouse data (Table 4.1).

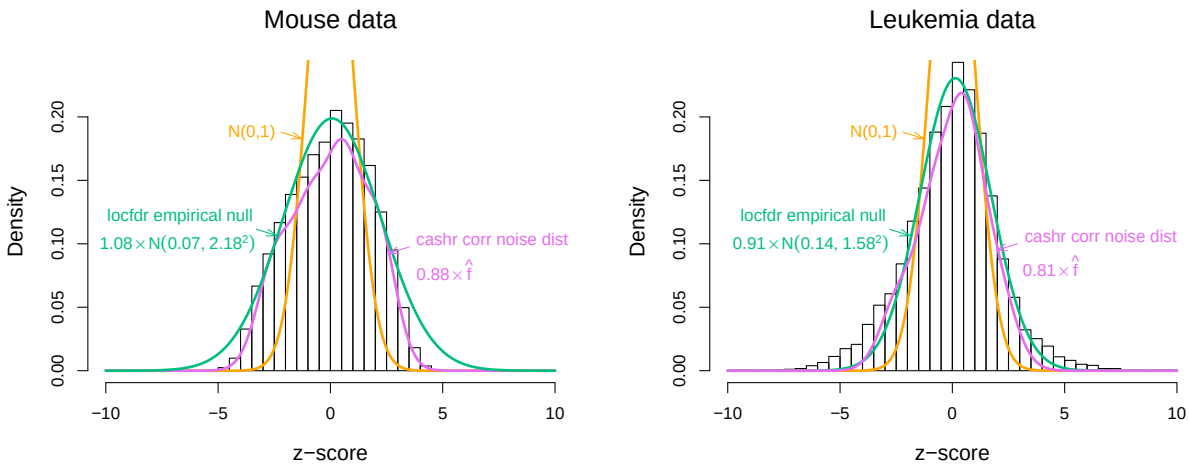

Figure 4.7: Distributions of z-scores from analyzing gene differential expression in two real data sets. In both data sets, for each gene j, a z-score z_j is computed, and $z_j \sim N(0,1)$ under the null hypothesis of no differential expression. Then we compare the histogram of z-scores with $N(0,1)$, the fitted correlated noise distribution from `cashr`, and the fitted empirical null from `locfdr`, scaled by respective estimated null proportions. Both histograms are substantially more dispersed than $N(0,1)$. The mouse data show inflation primarily in the shoulders of the distribution, and the fitted correlated noise distribution from `cashr` appears to be a much better fit than the fitted empirical null from `locfdr`, particularly in the tails. The leukemia data show inflation in both shoulders and tails of the distribution, indicating the presence of some strong signals. Although otherwise similar, the fitted correlated noise distribution from `cashr` has a noticeably shorter right tail than the fitted empirical null from `locfdr`, improving power.

Method	Number of discoveries	
	Leukemia data	Mouse data
`cashr`	385	0
`locfdr`	282	0
BH	1579	4130
`qvalue`	1972	6502
`ashr`	3346	17191

Table 4.1: Numbers of discoveries from different methods at nominal FDR = 0.1. We analyzed 7128 genes in the leukemia data and 17191 genes in the mouse data. In both data sets, the z-score distributions appear to have correlation-induced inflation, and the numbers of significant discoveries declared by methods accounting for correlation (`cashr` and `locfdr`) are much smaller than those ignoring correlation (BH, `qvalue`, `ashr`). For the leukemia data, `cashr` finds 37% more significant genes than `locfdr`.

Although the conclusions from `cashr` and `locfdr` are, here, qualitatively similar, there are some notable differences in their results. First, in the mouse data, the `cashr` correlated noise distribution gives, visually, a much better fit than the `locfdr` empirical null, particularly in the tails (Figure 4.7). This is because the `cashr` correlated noise distribution is ideally suited to capture this "shoulder-but-not-tail" inflation pattern that is symptomatic of correlation-induced inflation. The Gaussian empirical null distribution assumed by `locfdr` is simply inconsistent with these data. Indeed, this inconsistency is reflected in the null proportion estimated by `locfdr` (1.08) which exceeds the theoretical upper bound of 1.

Second, in the leukemia data, `cashr` identifies 37% more significant results than `locfdr` (385 vs 282). This is consistent with the greater power of `cashr` vs `locfdr` in our simulations. One reason that `locfdr` can lose power is that its Gaussian empirical null distribution tends to overestimate inflation in the tails when it tries to fit inflation in the shoulders. We see this feature in the mouse data, and although less obvious, this appears to also be the case for the leukemia data: the estimated standard deviation of the empirical null is 1.58, which is almost certainly too large: a pseudo-inflated Gaussian correlated noise distribution is unlikely to have standard deviation exceeding 1.4 (Appendix 4.6.2). In comparison the fitted correlated noise distribution from `cashr` has a noticeably shorter right tail (e.g. $z \in [4, 5]$) which leads it to categorize more z-scores in the right tail as significant (Figure 4.7). On a side note, `cashr` also experiences the benefits of `ashr` highlighted in Stephens (2017), which can also help increase power. For example, the unimodal assumption on the effects – which allows that some of the z-scores around zero may correspond to true, albeit non-significant, signals – can help improve estimates of π_0, and hence improve power.

Another feature of `cashr`, which distinguishes it from `locfdr`, is that, by estimating g while accounting for correlation-induced distortion, it can provide an estimate on the effect size distribution, g_1. For the mouse data, `cashr` estimates $\hat{\pi}_0 = 0.88$, or 12% of genes may be differentially expressed to some extent, although it is not able to pin down any clear example

of a differentially expressed gene: no gene has an estimated local FDR less than 0.80. One possible explanation for the lack of significant results in this case is lack of power. However, the estimated g_1 from `cashr` suggests that there may simply not exist any large effects to be discovered: 99% of the probability mass of the estimated g_1 is on effect size ≤ 0.26, or a mere 1.2-fold change in gene expression. Thus the signals here, if any, are too weak to be discerned from noise and pseudo-inflation.

We also applied the other methods – BH, `qvalue`, and `ashr` – to both data sets. All three methods find very large numbers of significant results in both data sets (Table 4.1). Although we do not know the truth in these real data, there is a serious concern that many of these results could be false positives, since these methods are all prone to erroneously viewing pseudo-inflation as true signal (Figure 4.6), and Figure 4.7 suggests that pseudo-inflation may be present in both data sets.

4.5 Discussion

We have presented a general approach to accounting for correlations among observations in the widely-used Empirical Bayes Normal Means model. Our strategy exploits theoretical results from Schwartzman (2010) to model the impact of correlation on the empirical distribution of correlated $N(0, 1)$ variables, and convex optimization techniques to produce an efficient implementation. We demonstrated through empirical examples that this strategy can both improve estimation of the underlying distribution of true effects (Figure 4.2) and – in the multiple testing setting – improve estimation of FDR compared with EB methods that ignore correlation (Figures 4.5-4.6). To the best of our knowledge, `cashr` is the first EBNM methodology to deal with correlated noise in this way.

Our empirical results demonstrate some benefits of the EB approach to multiple testing compared with traditional methods. In particular, `cashr` provides, on average, more accurate estimates of the FDP than either BH or `qvalue`. However, although we find these empirical

results encouraging, we do not have theoretical guarantees of (frequentist) FDR control. That said, theoretical guarantees of FDR control under arbitrary correlation structure are lacking even for the widely-studied BH method. BH has been shown to control FDR under certain correlation stuctures (e.g. "positive regression dependence on subsets"; Benjamini and Yekutieli, 2001). The Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) is proved to control FDR under arbitrary dependence, but at the cost of being excessively conservative, and is consequently rarely used in practice.

A key feature of `cashr` is that it requires no information about the actual correlations among observations. This has the important advantage that it can be applied wherever EBNM methods that ignore correlation can be applied. On the other hand, when additional information on correlations is available it clearly may be helpful to incorporate it into analyses. Within our approach such information could be used to estimate the moments of the pairwise correlations, and thus inform estimates of ω in the correlated noise distribution $f(\cdot;\omega)$. Alternatively, one could take a more ambitious approach: explicitly model the whole $p \times p$ correlation matrix, and use this to help inform inference (e.g. Benjamini and Heller, 2007; Wu, 2008; Sun and Cai, 2009; Friguet et al., 2009; Fan et al., 2012). Modeling correlation is likely to provide more efficient inferences when it can be accurately achieved (Hall and Jin, 2010). However, in many situations – particularly involving small sample sizes – reliably modeling correlation may be impossible. Under what circumstances this more ambitious approach produces better inferences could be one area for future investigation.

The main assumptions underlying `cashr` are that the correlated noise is marginally $N(0,1)$, and that the standard deviations are reliably computed. In the multiple testing setting this corresponds to assuming that the test statistics are (marginally) well calibrated. If these conditions do not hold – for example, due to failure of asymptotic theory underlying test statistic computations, or due to confounding factors (such as batch effects in gene expression studies), then `cashr` could give unreliable results. Of course `cashr` is not unique

in this regard – methods like BH and `qvalue` similarly assume that test statistics are well calibrated. Dealing with confounders in gene expression studies is an active area of research, and several approaches exist, many of them based on factor analysis (e.g. Leek and Storey, 2007; Sun et al., 2012; Gagnon-Bartsch and Speed, 2012; Wang et al., 2017; Gerard and Stephens, 2019, 2020). Again, the possibility of combining these ideas with our methods could be a future research direction.

4.6 Appendix

4.6.1 Marginal distributions of the simulated null random noise

Figures 4.8 and 4.9 offer support for the claim that the z-scores simulated in Section 4.2.1 are marginally $N(0,1)$-distributed.

Figure 4.8 compares z-scores simulated as in Section 4.2.1 with z-scores simulated under a modified framework that removes gene-gene correlations, and with iid $N(0,1)$ samples. The modified framework uses exactly the same simulation and analysis pipeline as the original framework of Section 4.2.1, with one important difference: in each simulation, *for each gene independently* we randomly selected two groups of five samples without replacement, hence removing gene-gene correlations.

The empirical CDF of 10^4 data sets simulated as in Section 4.2.1 show a huge amount of variability (panel (a)), presumably due to correlations among genes. In the modified framework, correlation-induced distortion disappears: the empirical CDF of all 10^4 data sets are almost exactly the same as $N(0,1)$ (panel (b)), just as with the iid $N(0,1)$ samples (panel (c)). This demonstrates that without gene-gene correlations, the analysis pipeline used here produces uncorrelated $N(0,1)$ z-scores.

In addition, Figure 4.9 shows that the mean empirical CDF of the 10^4 data sets simulated from the original framework – the average of empirical CDF of Figure 4.8(a) – is very close

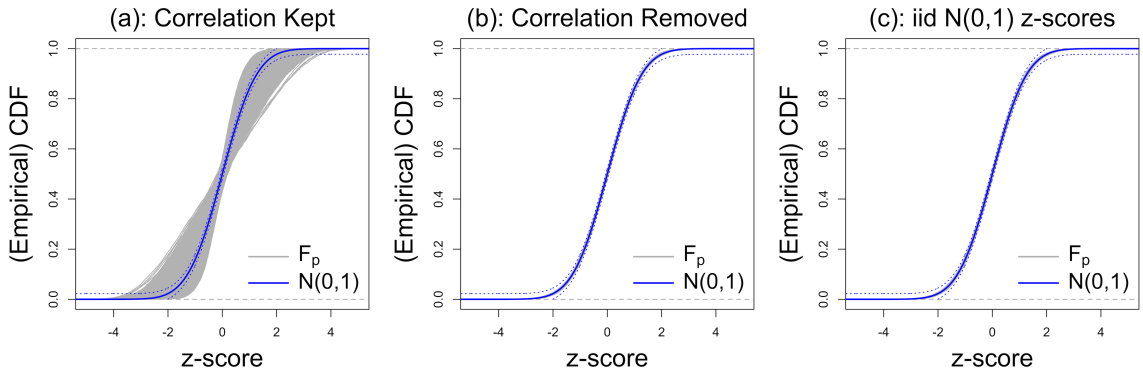

Figure 4.8: Comparison of 10^4 empirical CDF of z-scores (F_p) obtained by applying the same analysis pipeline to data simulated by two different frameworks: the original framework in Section 4.2.1 which keeps gene-gene correlations (panel (a)); and the modified framework to remove gene-gene correlations by randomizing samples for each gene (panel (b)). We also plot 10^4 empirical CDF of iid $N(0,1)$ samples for comparison (panel (c)). The z-scores obtained under the original framework show clear correlation-induced distortion – the variability of empirical CDF is huge. In contrast, when gene-gene correlations are removed under the modified framework, distortion disappears: empirical CDF are almost exactly the same as $N(0,1)$ and the variability is essentially invisible; indeed, they are indistinguishable from 10^4 empirical CDF of iid $N(0,1)$ z-scores. It shows clear evidence that the analysis pipeline can produce well-calibrated null z-scores if no gene-gene correlations. Dotted lines are Dvoretzky-Kiefer-Wolfowitz bounds with $\alpha = 1/10^4$.

to $N(0,1)$. Possible deviation happens only in the far tails ($|z\text{-score}| \in \{5,6\}$). Compared with $N(0,1.05^2)$ and $N(0,1.1^2)$, the deviation is very small even on the logarithmic scale (panels (b-c)), probably caused by numerical constraints as one or two z-scores in this area in a few data sets can make a visible difference.

4.6.2 Decomposing Gaussian by standardized Gaussian derivatives

Proposition 1. *The PDF of $N(\mu,\sigma^2)$ can be decomposed by standard Gaussian and its derivatives in the form of* (4.8) *if and only if $\sigma^2 \leq 2$.*

Proof. Let $h_l(\cdot)$ denote the lth probabilists' Hermite polynomial. The orthogonality and completeness of Hermite polynomials in $L^2(\mathbb{R}, \mathrm{d}\Phi)$ (e.g. Szegõ, 1975) leads to the following

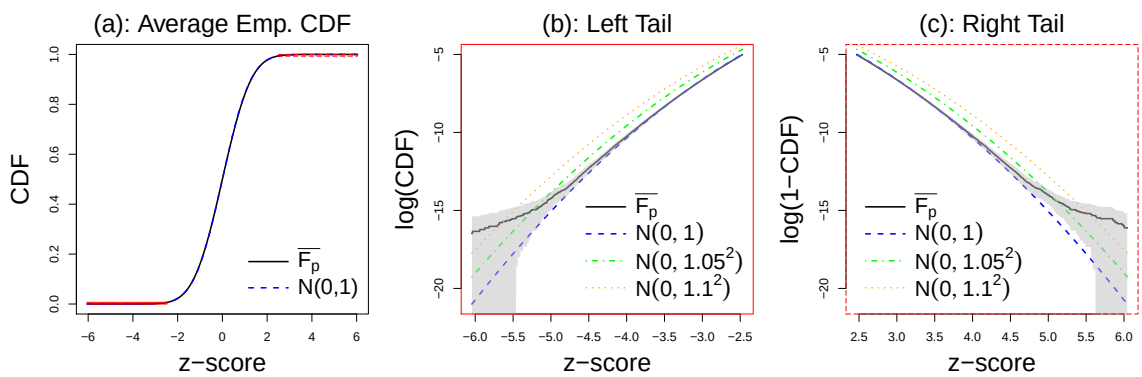

Figure 4.9: Illustration that the average empirical CDF of z-scores ($\overline{F_p}$) simulated as in Section 4.2.1 closely matches $N(0,1)$, aggregated over 10^4 data sets. Left: the average of all empirical CDF in Figure 4.8(a). The average empirical CDF is extremely close to $N(0,1)$. Center and Right: the left and right tails of the average empirical CDF on logarithmic scale. Shaded areas are 99.9% confidence bands. Compared with $N(0, 1.05^2)$ and $N(0, 1.1^2)$, possible deviation from $N(0,1)$ is light even in the far tails.

fact

$$\int_{\mathbb{R}} \frac{1}{\sqrt{m!}} h_m(x) \frac{1}{\sqrt{n!}} \varphi^{(n)}(x) \mathrm{d}x = (-1)^n \delta_{mn} \ , \qquad \forall m, n = 0, 1, 2, \ldots, \tag{4.35}$$

where $\delta_{mn} = \begin{cases} 1 & m = n \\ 0 & \text{otherwise} \end{cases}$. Therefore, if any PDF f can be decomposed in the form of (4.8), the coefficient of the lth-order standardized Gaussian derivative has to be

$$w_l = (-1)^l \int_{\mathbb{R}} \frac{1}{\sqrt{l!}} h_l(x) f(x) \mathrm{d}x = \frac{(-1)^l}{\sqrt{l!}} E_f[h_l] \ , \tag{4.36}$$

where $E_f[h_l]$, sometimes called "Hermite moment," is the expected value of $h_l(\cdot)$ when the PDF of the random variable is f. If f is $N(\mu, \sigma^2)$, we can obtain analytic expressions of these Hermite moments

$$E_f[h_l] = \mu^l + \sum_{k=1}^{\lfloor l/2 \rfloor} \binom{l}{2k} \mu^{l-2k} (\sigma^2 - 1)^k (2k-1)!! := M_l(\mu, \sigma^2 - 1) \ , \tag{4.37}$$

where $n!!$ denotes the double factorial of n, and $M_l(x, y)$ denotes the function of lth-order moment of a Gaussian with mean x and variance y. Putting (4.36)-(4.37) together, the coefficients in (4.8) become

$$w_l = \frac{(-1)^l}{\sqrt{l!}} M_l(\mu, \sigma^2 - 1) \ . \tag{4.38}$$

Note that w_l is not exploding if and only if $|\sigma^2 - 1| \leq 1$ or equivalently, $\sigma^2 \leq 2$. □

This result suggests that a pseudo-inflated Gaussian correlated noise distribution is not likely to have standard deviation greater than $\sqrt{2} \approx 1.4$.

In the special case when $\rho_{ij} = 1$, f becomes δ_z, a point mass on $Z \equiv z$, with z randomly sampled from $N(0, 1)$. It is interesting to note that δ_z can be decomposed in the form of (4.8) as

$$\delta_z(\cdot) = \varphi(\cdot) + \sum_{l=1}^{\infty} \left[\frac{(-1)^l}{\sqrt{l!}} h_l(z) \right] \left[\frac{1}{\sqrt{l!}} \varphi^{(l)}(\cdot) \right] \ , \quad \forall z \in \mathbb{R} \ .$$

4.6.3 Simulation details

Six choices of the non-null effect distribution

Table 4.2 lists the details of the six choices of g_1, the non-null effects in Section 4.4.1. The table also shows the average signal strength, $E[|\cdot|^2]$, and the probability of large signal, $\Pr(|\cdot| \geq \sqrt{2 \log p})$, conditioned on g_1.

Table 4.2: Details of the non-null effect distribution g_1 used in Section 4.1

g_1	PDF	$E[\lvert\cdot\rvert^2]$	$\Pr(\lvert\cdot\rvert \geq \sqrt{2\log p})$
Gaussian	$N(0, 2^2)$	4	0.032
Near Gaussian	$0.6N(0, 1) + 0.4N(0, 3^2)$	4.2	0.061
Spiky	$0.4N(0, 0.5^2) + 0.2N(0, 2^2) + 0.4N(0, 3^2)$	4.5	0.067
Skew	$0.25[N(-2, 2^2) + N(-1, 2^2) + N(0, 1) + N(1, 1)]$	4	0.045
Flat Top	$0.5N(-1.5, 1.5^2) + 0.5N(1.5, 1.5^2)$	4.5	0.031
Bimodal	$0.5N(-1.5, 1) + 0.5N(1.5, 1)$	3.25	0.0026

Implementation of methods

The existing methods we use for comparison in this paper mostly use the default settings in their respective `R` packages. That include `REBayes`, `deconvolveR` for deconvolution (Section 4.2.1), and `qvalue`, `locfdr` for multiple testing (Section 4.4). For `EbayesThresh`, we set `a=NA` to allow the scale parameter of the Laplace distribution to be estimated from the data. For `ashr`, we set `mixcompdist="normal"` to use scale mixture of zero-mean Gaussians to approximate g.

Pipeline for analyzing gene expression data

Let θ_j denote the true $\log_2$-fold change in gene expression for each gene j. The analysis pipeline is used to provide, for each θ_j, an estimate X_j with a standard error s_j, such that X_j can be assumed to be $N(\theta_j, s_j^2)$.

For RNA-seq data such as the mouse data, the analysis pipeline is described in Section 4.2.1.

For microarray data such as the leukemia data, we use a widely-used analysis protocol implemented in the `limma` software (Ritchie et al., 2015). This yield an estimate X_j for θ_j, and a corresponding p-value p_j from a moderated t-statistic (Smyth, 2004). Then as in Section 4.2.1, we convert the p-value to the corresponding z-score z_j and use it to compute the effective standard deviation s_j.

Reproducibility

All the code generating the results and plots in this paper are available at `https://github.com/LSun/cashr_paper`. The RNA-seq gene expression data from human liver tissues we use in this paper were generated by the GTEx Project, which was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this paper

were obtained from the GTEx Portal at `https://www.gtexportal.org`. In particular, the human liver RNA-seq data for realistic simulation are also available at `https://github.com/LSun/cashr_paper`. The leukemia microarray data are available at `http://statweb.stanford.edu/~ckirby/brad/LSI/datasets-and-programs/datasets.html`. The mouse heart RNA-seq data are available at `https://github.com/LSun/cashr_paper`.

4.6.4 Representation of the correlated noise distribution

If $\{Z_j\}$ are independent and p is large then F_p will be close to its mean, Φ. This is guaranteed by well-established results like the Glivenko-Cantelli theorem and the Dvoretzky-Kiefer-Wolfowitz inequality (e.g. Wasserman, 2006). However, when $\{Z_j\}$ are correlated F_p can be grossly different from Φ, as we have seen in Section 4.2.1. The covariance of F_p indicates how far it tends to stray from its mean, Φ, and therefore captures the extent of correlation-induced distortion. Schwartzman (2010) provides the following elegant characterization of the covariance of F_p. For completeness we also put it here.

Proposition 2. *The mean, variance, and covariance functions of F_p (Schwartzman, 2010) Assume $\forall i \neq j$, $\begin{bmatrix} Z_i \\ Z_j \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{bmatrix}\right)$. Let $\overline{\rho^l} := \frac{1}{p(p-1)} \sum_{i,j:i\neq j} \rho_{ij}^l$. Then $\forall x, y \in \mathbb{R}$,*

$$E(F_p(x)) = \Phi(x) \tag{4.39}$$

$$var(F_p(x)) = \left(1 - \frac{1}{p}\right) \sum_{l=1}^{\infty} \overline{\rho^l} \left[\frac{1}{\sqrt{l!}} \varphi^{(l-1)}(x)\right]^2 + \frac{1}{p}\Phi(x)(1 - \Phi(x)) \tag{4.40}$$

$$\begin{aligned} cov(F_p(x), F_p(y)) = &\left(1 - \frac{1}{p}\right) \sum_{l=1}^{\infty} \overline{\rho^l} \left[\frac{1}{\sqrt{l!}} \varphi^{(l-1)}(x)\right] \left[\frac{1}{\sqrt{l!}} \varphi^{(l-1)}(y)\right] \\ &+ \frac{1}{p}[\Phi(\min(x, y)) - \Phi(x)\Phi(y)] \end{aligned} \tag{4.41}$$

Proof. The mean function is straightforward. The covariance function

$$
\begin{aligned}
cov(F_p(x), F_p(y)) &= cov\left(\frac{1}{p}\sum_{i=1}^{p}\mathcal{I}(Z_i \le x), \frac{1}{p}\sum_{j=1}^{p}\mathcal{I}(Z_j \le y)\right) \\
&= E\left[\left(\frac{1}{p}\sum_{i=1}^{p}\mathcal{I}(Z_i \le x)\right)\left(\frac{1}{p}\sum_{j=1}^{p}\mathcal{I}(Z_j \le y)\right)\right] \\
&- E\left[\frac{1}{p}\sum_{i=1}^{p}\mathcal{I}(Z_i \le x)\right] E\left[\frac{1}{p}\sum_{j=1}^{p}\mathcal{I}(Z_j \le y)\right] \\
&= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} E[\mathcal{I}(Z_i \le x)\mathcal{I}(Z_j \le y)] - \Phi(x)\Phi(y) \\
&= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} P(Z_i \le x, Z_j \le y) - \Phi(x)\Phi(y) \\
&= \frac{1}{p^2}\sum_{i\neq j} P(Z_i \le x, Z_j \le y) + \frac{1}{p}\Phi(\min(x,y)) - \Phi(x)\Phi(y) \ . \qquad (4.42)
\end{aligned}
$$

According to Mehler's identity (Kibble, 1945), under the assumption of $\{Z_i, Z_j\}$ being bivariate normal, the joint PDF can be written as

$$
p(x,y) = \varphi(x)\varphi(y) + \sum_{l=1}^{\infty}\rho_{ij}^{l}\left[\frac{1}{\sqrt{l!}}\varphi^{(l)}(x)\right]\left[\frac{1}{\sqrt{l!}}\varphi^{(l)}(y)\right] \ , \qquad (4.43)
$$

so the joint CDF is

$$
P(Z_i \le x, Z_j \le y) = \Phi(x)\Phi(y) + \sum_{l=1}^{\infty}\rho_{ij}^{l}\left[\frac{1}{\sqrt{l!}}\varphi^{(l-1)}(x)\right]\left[\frac{1}{\sqrt{l!}}\varphi^{(l-1)}(y)\right] \ . \qquad (4.44)
$$

(4.42) and (4.44) lead to the covariance function (4.41). Setting $x = y$ gives the variance function (4.40). □

Note that $var(F_p)$ has two parts. The second part $\frac{1}{p}\Phi(z)(1-\Phi(z))$ is the familiar variance function when $\{Z_j\}$ are independent, and it quickly vanishes as p increases. This is why F_p

of iid $N(0,1)$ sample will not deviate much from Φ when p is large. In contrast, the first part

$$\left(1-\frac{1}{p}\right)\sum_{l=1}^{\infty}\overline{\rho^l}\left(\frac{1}{\sqrt{l!}}\varphi^{(l-1)}(x)\right)^2 \tag{4.45}$$

demonstrates the effect of correlation. If $\overline{\rho^l}$ is non-negligible for large p, $var(F_p)$ will be non-vanishing, and so F_p and the histogram of $\{Z_j\}$ are more likely to deviate substantially from $N(0,1)$.

When p is large,

$$cov(F_p(x), F_p(y)) \approx \sum_{l=1}^{\infty}\overline{\rho^l}\left[\frac{1}{\sqrt{l!}}\varphi^{(l-1)}(x)\right]\left[\frac{1}{\sqrt{l!}}\varphi^{(l-1)}(y)\right] \ . \tag{4.46}$$

(4.39) and (4.46) suggest we can characterize F_p as (4.6) (Schwartzman, 2010), assuming $\overline{\rho^l} \geq 0$ for all $l \in \mathbb{N}$. This assumption should not be too demanding for large p in practice. For example, when $l = 1$,

$$\overline{\rho} = \frac{1}{p(p-1)}\sum_{i\neq j}\rho_{ij} = \frac{1}{p(p-1)}(\mathbf{1}^T\Sigma_Z\mathbf{1} - p) \geq \frac{1}{p(p-1)}(-p) = -\frac{1}{p-1} \ , \tag{4.47}$$

following the fact that Σ_Z, the correlation matrix of $\{Z_j\}$, is positive semi-definite.

CHAPTER 5

DISCUSSION AND FUTURE WORK

Research on EB in theories, methodologies, and applications has seen tremendou progress in the past decades. This dissertation contributes to this growing body of work by introducing tools to examine existing EB methods, applying EB to statistically model scientific inquiries, and developing new methodologies to solve problems often encountered in large-scale statistical analysis.

Neither fully frequentist nor fully Bayesian, EB appears to enjoy the best of both worlds – exhibiting Bayesian features such as risk-reducing shrinkage and post-selection adaptivity, and keeping statistics of interest under frequentist control. Building on Jiang and Zhang (2009); Brown and Greenshtein (2009), Efron (2019) introduced the Oracle Bayes (OB) framework to understand this dual nature of EB and explore its potential role in bridging frequentism or Bayesianism, in the context of shrinkage estimation and risk reduction. Applying OB to multiple testing problems and connecting frequentist, Bayesian, and EB approaches could be an interesting research topic. Here we make some preliminary exploration along this line as a potential future direction.

5.1 Oracle Bayes Multiple Testing

Suppose we observe

$$X_j \overset{\text{ind}}{\sim} f_j(\cdot|\theta_j), \quad j = 1, \ldots, p. \tag{5.1}$$

where $\{X_j\} := \{X_1, \ldots, X_p\}$ are observations and $\{\theta_j\} := \{\theta_1, \ldots, \theta_p\}$ are primary parameters of interest. The subscript j in $f_j(\cdot|\theta_j)$ indicates that the difference among these distributions can be more than the difference among their primary parameters. One such example is $f_j(\cdot|\theta_j) = N(\cdot|\theta_j, s_j^2)$, a normal distribution with mean θ_j and variance s_j^2. Our

goal is to simultaneously test

$$H_0^j : \theta_j = 0, \quad j = 1, \ldots, p. \tag{5.2}$$

Under the OB framework (Efron, 2019), suppose an Oracle has told us the order statistics (or equivalently, the empirical distribution) of $\{\theta_j\}$, from which we know

$$\begin{aligned} \theta_{(\kappa_j)} &= 0, \quad j = 1, \ldots, p_0, \\ \theta_{(\kappa_j)} &\neq 0, \quad j = p_0 + 1, \ldots, p, \end{aligned} \tag{5.3}$$

where $\{\kappa_1, \ldots, \kappa_p\}$ is a permutation of $\{1, \ldots, p\}$, and $\theta_{(j)}$ is the jth order statistic of $\{\theta_j\}$. In addition, we define the OB prior $\bar{g}(\theta)$ as the discrete distribution putting probability $1/p$ on each point $\theta_{(\kappa_j)}$,

$$\bar{g}(\theta) = \frac{1}{p} \sum_{j=1}^{p} \delta(\theta - \theta_{(\kappa_j)}), \tag{5.4}$$

where $\delta(\cdot)$ denotes a point mass at zero.

Now we can define the *oracle local false discovery rate (Oracle lfdr)*, l_j, as the posterior probability of θ_j being zero under prior $\bar{g}(\cdot)$,

$$l_j := \Pr(\theta_j = 0 | X_j, \bar{g}) = \frac{\Pr(\theta_j = 0 | \bar{g}) f_j(X_j | \theta_j = 0)}{p_j(X_j | \bar{g})} \tag{5.5}$$

$$= \frac{\frac{p_0}{p} f_j(X_j | 0)}{p_j(X_j | \bar{g})} \tag{5.6}$$

$$= \frac{\frac{p_0}{p} f_j(X_j | 0)}{\frac{p_0}{p} f_j(X_j | 0) + \frac{1}{p} \sum_{j=p_0+1}^{p} f_j(X_j | \theta_{(\kappa_j)})}, \tag{5.7}$$

where $p_j(\cdot | g)$ denotes the marginal distribution of X_j under prior g.

The construction of Oracle lfdr is frequentist in nature. However, it has close connection

to the Bayesian local false discovery rate (Efron et al., 2001), defined upon z-scores. Let $\pi_0 := \frac{p_0}{p}$ denote the proportion of zero among $\{\theta_j\}$, i.e., the oracle null proportion $\Pr(\theta_j = 0|\bar{g})$. In the context of z-scores, $f_j(\cdot|0)$ is the common distribution of all null z-scores, denoted as $f_0(\cdot)$, and $p_j(\cdot|\bar{g})$, the marginal distribution under $\bar{g}$, is assumed to be common for all z-scores and not depending on j, so it can be denoted as $h(\cdot)$. Then (5.6) becomes

$$l_j = \frac{\pi_0 f_0(\cdot)}{h(\cdot)}, \tag{5.8}$$

equivalent to the Bayesian lfdr construct in Efron et al. (2001).

Section 5.2 shows that Oracle lfdr provides good ordering of the p hypotheses being simultaneously tested. Section 5.3 discusses a (frequentist) FDR-controlling procedure based on Oracle lfdr. Preliminary simulation results are provided.

5.2 Ordering Hypotheses by Oracle lfdr

Like Bayesian lfdr, Oracle lfdr naturally orders p hypotheses: intuitively, a hypothesis with a smaller Oracle lfdr should be seen as more "significant" than, and hence be rejected before, a hypothesis with a larger Oracle lfdr. Therefore, a simple decision rule is to reject hypotheses whose Oracle lfdr is less than a threshold λ, $0 < \lambda < 1$. λ is usually determined to maintain a certain kind of Type I error control. In fully Bayesian settings, Müller et al. (2004) discusses this decision rule and concludes that this rule is optimal under several arguably common-sense loss functions.

Thresholding on Oracle lfdr $\leq \lambda$ in (5.7) is equivalent to thresholding on the following quantity

$$\text{MTLR}_j := \frac{\frac{1}{p-p_0}\sum_{j=p_0+1}^{p} f_j(X_j|\theta_{(\kappa_j)})}{f_j(X_j|0)} \geq \tau, \tag{5.9}$$

where $\tau := \frac{p_0(1-\lambda)}{(p-p_0)\lambda}$. Here "MTLR" stands for "Multiple Testing Likelihood Ratio," as this quantity can be seen as the likelihood ratio test statistic for

$$\begin{aligned} &H_0^j : \theta_j = 0, \qquad \text{versus} \\ &H_a^j : \theta_j \text{ is one of } \{\theta_{(\kappa_j)} : j = p_0 + 1, \ldots, p\} \text{ with equal probability.} \end{aligned} \tag{5.10}$$

Incidentally, we can also call this quantity "Oracle Bayes Factor," since it can be viewed as a Bayes Factor under the OB prior $\bar{g}(\cdot)$.

MTLR has some connection to the optimal discovery procedure proposed by Storey (2007), which attempts to develop an optimal multiple testing framework analogous to the Neyman-Pearson Lemma for single hypothesis testing (Neyman et al., 1933). The author proposes a significance thresholding function $S_{\text{ODP}}(\cdot)$ for *all* tests, which in our settings boils down to

$$S_{\text{ODP}}(\cdot) = \frac{f_{(\kappa_{p_0+1})}(\cdot|\theta_{(\kappa_{p_0+1})}) + \cdots + f_{(\kappa_N)}(\cdot|\theta_{(\kappa_N)})}{f_{(\kappa_1)}(\cdot|0) + \cdots + f_{(\kappa_{p_0})}(\cdot|0)}, \tag{5.11}$$

where $f_{(\kappa_j)}(\cdot|\theta_{(\kappa_j)})$ denotes the sampling distribution corresponding to $\theta_{(\kappa_j)}$. The decision rule is to reject H_0^j if and only if $S_{\text{ODP}}(X_j) \geq \tau$. The author defines "simultaneous thresholding procedures" to be those using a *common* thresholding function and a *common* threshold for all tests, and proves that for each fixed τ, "this procedure yields the maximum number of expected true positive results among all simultaneous thresholding procedures" having an equal or smaller number of expected false positive results.

Both MTLR and S_{ODP} attempt to "borrow strength" among tests by accounting for information from all p hypotheses together rather than considering each hypothesis separately. When the primary parameter θ_j is the only difference between sampling distributions $f_j(\cdot|\theta_j)$ so that $\theta_i = \theta_j \Rightarrow f_i(\cdot|\theta_i) = f_j(\cdot|\theta_j)$, thresholding on MTLR is equivalent to thresholding on S_{ODP}. For example, if $f_j(\cdot|\theta_j) = N(\cdot|\theta_j, 1)$, both rules boil down to rejecting H_0^j if and

only if

$$\frac{\varphi(X_j - \theta_{(\kappa_{p_0+1})}) + \cdots + \varphi(X_j - \theta_{(\kappa_N)})}{\varphi(X_j)} \geq \tau, \tag{5.12}$$

where $\varphi(\cdot)$ denotes the PDF of $N(0,1)$. However, if the difference between $f_j(\cdot|\theta_j)$ is more than the difference between θ_j, the two rules are different. The following example illustrates this point.

Example 1 (Heteroskedastic Normal Means)**.** *Suppose we observe*

$$X_j \overset{ind}{\sim} N(\cdot|\theta_j, s_i^2), \quad i = 1, \ldots, N, \tag{5.13}$$

and we have obtained oracle information (5.3)*. Our goal is multiple testing* (5.2)*. Then*

$$MTLR_j = \frac{1}{p - p_0} \frac{N(X_j|\theta_{(\kappa_{p_0+1})}, s_j^2) + \cdots + N(X_j|\theta_{(\kappa_N)}, s_j^2)}{N(X_j|0, s_j^2)}, \tag{5.14}$$

$$S_{ODP}(X_j) = \frac{N(X_j|\theta_{(\kappa_{p_0+1})}, s^2_{(\kappa_{p_0+1})}) + \cdots + N(X_j|\theta_{(\kappa_N)}, s^2_{(\kappa_N)})}{N(X_j|0, s^2_{(\kappa_1)}) + \cdots + N(X_j|0, s^2_{(\kappa_{p_0})})}, \tag{5.15}$$

where $s_{(\kappa_j)}$ denotes the standard deviation corresponding to $\theta_{(\kappa_j)}$. (5.14) *and* (5.15) *can provide different orders for the p hypotheses. Also note that S_{ODP} requires more oracle information than MTLR: in addition to the order statistics of $\{\theta_j\}$, S_{ODP} also needs their corresponding s_j in the same order.*

We now empirically compare these two ordering rules by comparing the areas under their respective receiver operating characteristic curves (AUC), in a simple simulation. Let $N = 10^4$, 90% of these $\{\theta_j\}$ are zero, 2% are $1, 2, 3, 4, 5$ each, and all $\{\theta_j\}$ are in random order. $s_j \overset{\text{ind}}{\sim} \chi^2_{10}$ and normalized to have $\frac{1}{p}\sum_{i=1}^p s_j^2 = 1$ (Figure 5.1). The same $\{\theta_j\}$ and $\{s_j\} := \{s_1, \ldots, s_p\}$ are used for all simulations. In each simulation, $\{X_j\}$ are generated

from (5.13). To compute MTLR and S_{ODP} from these data, the actual values of $\{\theta_j\}$ are withheld, but their respectively needed oracle information is provided. We then compute the AUC produced by MTLR and S_{ODP}, using the true $\{\theta_j\}$.

We also include the case of working with z-scores to avoid heteroskedasticity. From (5.13),

$$z_j := X_j/s_j \overset{\text{ind}}{\sim} N(\cdot|\theta_j/s_j, 1), \quad j = 1, \ldots, p. \tag{5.16}$$

For a fair comparison, suppose we have also obtained the order statistics of $\{\theta_1/s_1, \ldots, \theta_p/s_p\}$ from the Oracle. In each simulation, we compute MTLR from these z-scores, and AUC from these MTLR. Note that with homoskedasticity by using only z-scores in (5.16), MTLR and S_{ODP} have the same AUC.

To provide a baseline, we also compute AUC of p-values, calculated as $p_j = 2\Phi(-|z_j|)$, where $\Phi(\cdot)$ is the CDF of $N(0, 1)$, and of "adjusted" p-values by the Benjamini-Hochberg procedure (BH; Benjamini and Hochberg, 1995).

We run 1000 simulations, and plot in Figure 5.1 the distribution (as boxplots) of AUC from five ordering rules: heteroskedastic MTLR on $\{X_j, s_j\}$, heteroskedastic S_{ODP} on $\{X_j, s_j\}$, homoskedastic MTLR (or equivalently S_{ODP}) on z-scores, as well as p-values and BH. Higher AUC indicates better ordering. Despite small difference, heteroskedastic MTLR produces clearly higher AUC than heteroskedastic S_{ODP}. Second, for MTLR, heteroskedastic modeling is better than homoskedastic modeling. But even the latter is still better than heteroskedastic S_{ODP}. Indeed, for almost every simulated data set, the AUC of heteroskedastic MTLR is higher than that of homoskedastic MTLR, which is higher than that of heteroskedastic S_{ODP}.

It is probably a good sign that MTLR, or equivalently Oracle lfdr, while requiring less oracle information, yields better results than the "optimal discovery procedure." It can happen because MTLR is not a "simultaneous thresholding procedure" S_{ODP} is proved to

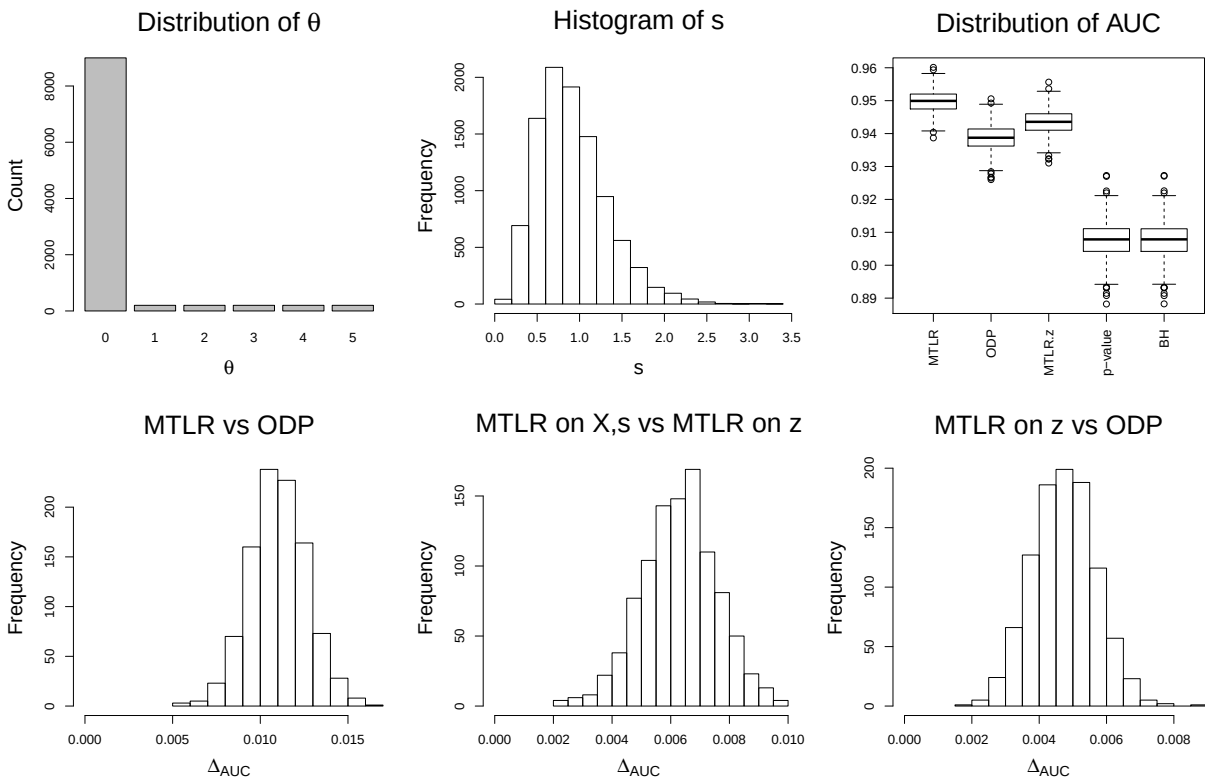

Figure 5.1: Comparison of ordering rules. The same $\{\theta_j\}$ and $\{s_j\}$ are used for all the simulations. The boxplots of AUC show small but clear improvement of MTLR over S_{ODP}. For each simulated data set, we compute Δ_{AUC}, the differences in AUC produced by two procedures, and plot them in the three histograms on the bottom. All $\Delta_{\text{AUC}} > 0$; that is, for each simulated data set, the AUC of MTLR on $\{X_j, s_j\}$ in the heteroskedastic setting is higher than that of MTLR (or equivalently S_{ODP}) on z_j-scores in the homoskedastic setting, which is higher than that of S_{ODP} on $\{X_j, s_j\}$.

outperform in Storey (2007). Indeed, although MTLR still uses a common threshold for all tests, it uses *different* thresholding functions for different tests.

5.3 FDR Control by Oracle lfdr

We can formulate a multiple testing procedure based on Oracle lfdr to control FDR at level q. Let $l_{(1)} \leq \cdots \leq l_{(p)}$ be the ordered Oracle lfdr. Let R be the largest i for which $\frac{1}{i}(l_{(1)} + \cdots + l_{(i)}) \leq q$. Then reject all null hypotheses corresponding to $l_{(1)}, \ldots, l_{(R)}$, with

the convention $R := 0$ and no rejection if $l_{(1)} > q$. Simply put, the discovery set will be the set that contains as many hypotheses as possible whose average Oracle lfdr is less than q.

Let the false discovery proportion (FDP) of this procedure

$$\mathrm{FDP}(\{X_j\}, q|\{\theta_j\}) := \frac{\text{The number of true null hypotheses rejected}}{\max\{R, 1\}}. \tag{5.17}$$

FDR is then defined as the mean of FDP

$$\mathrm{FDR}_\theta(q) = \mathrm{E}[\mathrm{FDP}(\{X_j\}, q|\{\theta_j\})], \tag{5.18}$$

where the expectation is taken over the sample space of $\{X_j\}$.

We now show the performance of this procedure in a simple simulation. The basic setting is similar to the one in Section 5.2. We use the same $\{\theta_j\}$ and $\{s_j\}$ as in Figure 5.1. In each simulation, we compute Oracle lfdr using simulated data and necessary oracle information. At each nominal FDR q from a dense grid of $q \in (0, 1)$, we form a discovery set according to our procedure, and compute its FDP. We run 1000 simulations, and compute the average FDP at each q, as well as their 2.5% and 97.5% quantiles. Figure 5.2 shows the multiple testing procedure based on Oracle lfdr can indeed control the frequentist FDR very well.

So far, the preliminary simulation results of Oracle Bayes multiple testing procedures are encouraging. They suggest the possible optimality of the MTLR-based multiple testing procedure, the apparent superiority of modeling heteroskedasticity over homoskedasticity, and the ability of the Oracle lfdr-based procedure to control the frequentist FDR. The accuracy of different EB methods in estimating the OB prior is also worth exploring. Together, OB may enhance our understanding of EB's usually good frequentist performance on multiple testing (Efron, 2010b; Muralidharan, 2010; Stephens, 2017).

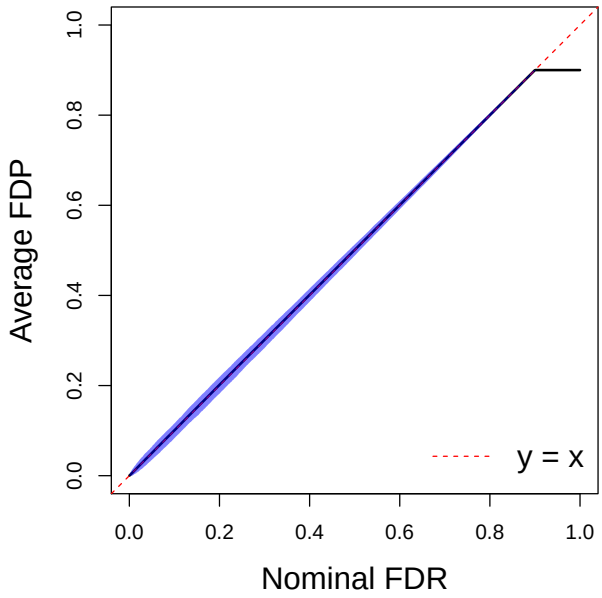

Figure 5.2: Illustration that the Oracle lfdr-based procedure controls the frequentist FDR. We use the same $\{\theta_j\}$ and $\{s_j\}$ in Figure 5.1. The average FDP vs nominal FDR line is very close to the dotted $y = x$ line in most of the area, and converges to the $y = 0.9$ line when the nominal FDR $q \geq 0.9$. The small blue area indicates the 2.5% to 97.5% quantiles of FDP, implying that the variability of FDP is small.

References

Ballard, D., Abraham, C., Cho, J., and Zhao, H. (2010). "Pathway Analysis Comparison Using Crohn's Disease Genome Wide Association Studies." *BMC Medical Genomics*, 3(1): 25.

Benjamini, Y. and Heller, R. (2007). "False Discovery Rates for Spatial Signals." *Journal of the American Statistical Association*, 102(480): 1272–1281.

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 Yekutieli, D. (2001). "The Control of The False Discovery Rate in Multiple Testing Under Dependency." *Annals of Statistics*, 29(4): 1165–1188.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer, 2nd edition.

Blanchard, G. and Roquain, E. (2009). "Adaptive False Discovery Rate Control Under Independence and Dependence." *Journal of Machine Learning Research*, 10: 2837–2871.

Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. Stockholm, Sweden: Almqvist & Wiksell, 1st edition.

Bovy, J., Hogg, D. W., and Roweis, S. T. (2011). "Extreme Deconvolution: Inferring Complete Distribution Functions from Noisy, Heterogeneous and Incomplete Observations." *Annals of Applied Statistics*, 5(2B): 1657–1677.

Brown, L. D. (2008). "In-season Prediction of Batting Averages: A Field Test of Empirical Bayes and Bayes Methodologies." *Annals of Applied Statistics*, 2(1): 113–152.

Brown, L. D. and Greenshtein, E. (2009). "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-dimensional Vector of Normal Means." *Annals of Statistics*, 37(4): 1685–1704.

Carbonetto, P. and Stephens, M. (2013). "Integrated Enrichment Analysis of Variants and Pathways in Genome-Wide Association Studies Indicates Central Role for IL-2 Signaling Genes in Type 1 Diabetes, and Cytokine Signaling Genes in Crohn's Disease." *PLOS Genetics*, 9(10): 1–19.

Clyde, M. and George, E. I. (2000). "Flexible Empirical Bayes Estimation for Wavelets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 681–698.

Cordy, C. B. and Thomas, D. R. (1997). "Deconvolution of a Distribution Function." *Journal of the American Statistical Association*, 92(440): 1459–1465.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., et al. (2015). "Pathway and Network Analysis of Cancer Genomes." *Nature Methods*, 12(7): 615.

Dawid, A. P. (1994). *Selection Paradoxes of Bayesian Inference*, volume 24 of *Lecture Notes–Monograph Series*, 211–220. Hayward, CA: Institute of Mathematical Statistics.

De Carvalho, M. and Ramos, A. (2012). "Bivariate Extreme Statistics, II." *REVSTAT-Statistical Journal*, 10(1): 83–107.

De Finetti, B. (1937). "La prévision: ses lois logiques, ses sources subjectives [Foresight: Its Logical Laws, Its Subjective Sources]." *Annales de l'institut Henri Poincaré*, 7(1): 1–68.

De Leeuw, C. A., Neale, B. M., Heskes, T., and Posthuma, D. (2016). "The Statistical Properties of Gene-Set Analysis." *Nature Reviews Genetics*, 17(6): 353.

Dey, K. K. and Stephens, M. (2018). "CorShrink : Empirical Bayes Shrinkage Estimation of Correlations, with Applications." *bioRxiv*.

Efron, B. (1996). "Empirical Bayes Methods for Combining Likelihoods." *Journal of the American Statistical Association*, 91(434): 538–550.

— (2004). "Large-Scale Simultaneous Hypothesis Testing." *Journal of the American Statistical Association*, 99(465): 96–104.

— (2005). "Local False Discovery Rates." Technical Report 2005-20B/234, Division of Biostatistics, Stanford University.

— (2007a). "Correlation and Large-Scale Simultaneous Significance Testing." *Journal of the American Statistical Association*, 102(477): 93–103.

— (2007b). "Size, Power and False Discovery Rates." *Annals of Statistics*, 35(4): 1351–1377.

— (2008). "Microarrays, Empirical Bayes and the Two-Groups Model." *Statistical Science*, 23(1): 1–22.

— (2010a). "Correlated z-values and the Accuracy of Large-Scale Statistical Estimates." *Journal of the American Statistical Association*, 105(491): 1042–1055.

— (2010b). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. New York, NY: Cambridge University Press, 1st edition.

— (2014). "Two Modeling Strategies for Empirical Bayes Estimation." *Statistical Science*, 29(2): 285–301.

— (2016). "Empirical Bayes Deconvolution Estimates." *Biometrika*, 103(1): 1–20.

— (2018). "Curvature and Inference for Maximum Likelihood Estimates." *Annals of Statistics*, 46(4): 1664–1692.

— (2019). "Bayes, Oracle Bayes and Empirical Bayes." *Statistical Science*, 34(2): 177–201.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. New York, NY: Cambridge University Press, 1st edition.

Efron, B. and Morris, C. (1972). "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part II: The Empirical Bayes Case." *Journal of the American Statistical Association*, 67(337): 130–139.

— (1973). "Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach." *Journal of the American Statistical Association*, 68(341): 117–130.

Efron, B. and Tibshirani, R. (2007). "On Testing the Significance of Sets of Genes." *Annals of Applied Statistics*, 1(1): 107–129.

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.

Eisenberg, E. and Levanon, E. Y. (2013). "Human Housekeeping Genes, Revisited." *Trends in Genetics*, 29(10): 569–574.

Fan, J. (1991). "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems." *Annals of Statistics*, 19(3): 1257–1272.

Fan, J., Han, X., and Gu, W. (2012). "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence." *Journal of the American Statistical Association*, 107(499): 1019–1035.

Franck, W. E. (1984). "A Likelihood Ratio Test for Stochastic Ordering." *Journal of the American Statistical Association*, 79(387): 686–691.

Friguet, C., Kloareg, M., and Causeur, D. (2009). "A Factor Model Approach to Multiple Testing Under Dependence." *Journal of the American Statistical Association*, 104(488): 1406–1415.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). "Using Control Genes to Correct for Unwanted Variation in Microarray Data." *Biostatistics*, 13(3): 539–552.

Gelman, A., Hill, J., and Yajima, M. (2012). "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness*, 5(2): 189–211.

Gerard, D. (2019). "Data-based RNA-seq Simulations by Binomial Thinning." *bioRxiv*, 758524.

Gerard, D. and Stephens, M. (2019). "Unifying and Generalizing Methods for Removing Unwanted Variation Based on Negative Controls." *Statistica Sinica*, in press.

— (2020). "Empirical Bayes Shrinkage and False Discovery Rate Estimation, Allowing for Unwanted Variation." *Biostatistics*, 21(1): 15–32.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*, 286(5439): 531–537.

Greenland, S. and Robins, J. M. (1991). "Empirical-Bayes Adjustments for Multiple Comparisons Are Sometimes Useful." *Epidemiology*, 2(1): 244–251.

Hall, P. and Jin, J. (2010). "Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise." *Annals of Statistics*, 38(3): 1686–1732.

Harter, H. L. (1984). "Another Look at Plotting Positions." *Communications in Statistics—Theory and Methods*, 13(13): 1613–1633.

Hazen, A. (1914). "The Storage to Be Provided in Impounding Reservoirs for Municipal Water Supply." *Transactions of the American Society of Civil Engineers*, 77: 1539–1669.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). "GSEA-SNP: Applying Gene Set Enrichment Analysis to SNP Data from Genome-Wide Association Studies." *Bioinformatics*, 24(23): 2784–2785.

Irizarry, R. A., Wang, C., Zhou, Y., and Speed, T. P. (2009). "Gene Set Enrichment Analysis Made Simple." *Statistical Methods in Medical Research*, 18(6): 565–575.

Jia, P., Wang, L., Fanous, A. H., Chen, X., Kendler, K. S., Zhao, Z., Consortium, I. S., et al. (2012). "A Bias-Reducing Pathway Enrichment Analysis of Genome-Wide Association Data Confirmed Association of the MHC Region with Schizophrenia." *Journal of Medical Genetics*, 49(2): 96–103.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). "Synthetic Spike-in Standards for RNA-seq Experiments." *Genome Research*, 21(9): 1543–1551.

Jiang, W. and Zhang, C.-H. (2009). "General Maximum Likelihood Empirical Bayes Estimation of Normal Means." *Annals of Statistics*, 37(4): 1647–1684.

Johnstone, I. (2019). "Gaussian Estimation: Sequence and Wavelet Models." Unpublished Manuscript.

Johnstone, I. and Silverman, B. (1997). "Wavelet Threshold Estimators for Data with Correlated Noise." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2): 319–351.

— (2004). "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences." *Annals of Statistics*, 32(4): 1594–1649.

— (2005a). "`EbayesThresh`: R Programs for Empirical Bayes Thresholding." *Journal of Statistical Software*, 12(8): 1–38.

— (2005b). "Empirical Bayes Selection of Wavelet Thresholds." *Annals of Statistics*, 33(4): 1700–1752.

Khintchine, A. Y. (1938). "On Unimodal Distributions." *Izvestiya Nauchno-Issledovatel'skogo Instituta Matematiki i Mekhaniki*, 2(2): 1–7.

Kibble, W. F. (1945). "An Extension of a Theorem of Mehler's on Hermite Polynomials." *Mathematical Proceedings of the Cambridge Philosophical Society*, 41(1): 12–15.

Kiefer, J. and Wolfowitz, J. (1956). "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *Annals of Mathematical Statistics*, 27(4): 887–906.

Kim, Y., Carbonetto, P., Stephens, M., and Anitescu, M. (2018). "A Fast Algorithm for Maximum Likelihood Estimation of Mixture Proportions Using Sequential Quadratic Programming." *ArXiv e-prints*.

Kimball, B. F. (1946). "Assignment of Frequencies to a Completely Ordered Set of Sample Data." *Eos, Transactions American Geophysical Union*, 27(6): 843–846.

— (1960). "On the Choice of Plotting Positions on Probability Paper." *Journal of the American Statistical Association*, 55(291): 546–560.

Koenker, R. and Gu, J. (2017). "`REBayes`: An R Package for Empirical Bayes Mixture Methods." *Journal of Statistical Software*, 82(8): 1–26.

Koenker, R. and Mizera, I. (2014). "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules." *Journal of the American Statistical Association*, 109(506): 674–685.

Laird, N. (1978). "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution." *Journal of the American Statistical Association*, 73(364): 805–811.

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). "Fast and Rigorous Computation of Gene and Pathway Scores from SNP-based Summary Statistics." *PLoS Computational Biology*, 12(1): e1004714.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). "`voom`: Precision Weights Unlock Linear Model Analysis Tools for RNA-seq Read Counts." *Genome Biology*, 15(2): R29.

Lebedev, V. V. (1952). *Gidrologiya i Gidrometriya w Zadacah (Hydrology and Hydrometry in Problems)*. Leningrad, USSR: Gidrometeorologiceskoe Izdatel'stvo, 1st edition.

Lee, Y. J. and Wolfe, D. A. (1976). "A Distribution-Free Test for Stochastic Ordering." *Journal of the American Statistical Association*, 71(355): 722–727.

Leek, J. T. and Storey, J. D. (2007). "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis." *PLOS Genetics*, 3(9): 1–12.

Lu, M. (2018). "Generalized Adaptive Shrinkage Methods and Applications in Genomic Studies." Ph.D. Dissertation, University of Chicago.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006). "Sparse Statistical Modelling in Gene Expression Genomics." In Do, K.-A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 155–176. Cambridge: Cambridge University Press.

Makkonen, L. (2008). "Bringing Closure to the Plotting Position Controversy." *Communications in Statistics—Theory and Methods*, 37(3): 460–467.

Mandelboum, S., Manber, Z., Elroy-Stein, O., and Elkon, R. (2019). "Recurrent Functional Misinterpretation of RNA-seq Data Caused by Sample-specific Gene Length Bias." *PLoS Biology*, 17(11).

Mau, J. (1988). "A Generalization of a Nonparametric Test for Stochastically Ordered Distributions to Censored Survival Data." *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3): 403–412.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (2010). "Enrichment Map: A Network-based Method for Gene-Set Enrichment Visualization and Interpretation." *PloS One*, 5(11): e13984.

Mooney, M. A., Nigg, J. T., McWeeney, S. K., and Wilmot, B. (2014). "Functional and Genomic Context in Pathway Analysis of GWAS Data." *Trends in Genetics*, 30(9): 390–400.

Morris, C. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

MOSEK ApS (2018). *MOSEK `Rmosek` Package. Release 8.1.0.51*.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). "Optimal Sample Size for Multiple Testing." *Journal of the American Statistical Association*, 99(468): 990–1001.

Muralidharan, O. (2010). "An Empirical Bayes Mixture Method for Effect Size and False Discovery Rate Estimation." *Annals of Applied Statistics*, 4(1): 422–438.

Narasimhan, B. and Efron, B. (2016). "A G-modeling Program for Deconvolution and Empirical Bayes Estimation." Technical Report 2016-07, Department of Statistics, Stanford University.

Neyman, J., Pearson, E. S., and Pearson, K. (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.

Owen, A. B. (2005). "Variance of the Number of False Discoveries." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3): 411–426.

Petrone, S., Rousseau, J., and Scricciolo, C. (2014). "Bayes and Empirical Bayes: Do They Merge?" *Biometrika*, 101(2): 285–302.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). "Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders." *American Journal of Human Genetics*, 94(5): 677–694.

Qiu, X., Klebanov, L., and Yakovlev, A. (2005). "Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology for Finding Differentially Expressed Genes." *Statistical Applications in Genetics and Molecular Biology*, 4(1): Article 34.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). "Pathway Enrichment Analysis and Visualization of Omics Data Using g:Profiler, GSEA, Cytoscape and EnrichmentMap." *Nature Protocols*, 14(2): 482.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). "`limma` Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies." *Nucleic Acids Research*, 43(7): e47.

Robbins, H. (1951). "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 131–149. Berkeley, California: University of California Press.

— (1956). "An Empirical Bayes Approach to Statistics." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 157–163. Berkeley, California: University of California Press.

— (1964). "The Empirical Bayes Approach to Statistical Decision Problems." *Annals of Mathematical Statistics*, 35(1): 1–20.

Robertson, T. and Wright, F. T. (1981). "Likelihood Ratio Tests for and Against a Stochastic Ordering Between Multinomial Populations." *Annals of Statistics*, 9(6): 1248–1257.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). "`edgeR`: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics*, 26(1): 139–140.

Rocke, D. M., Ruan, L., Gossett, J. J., Durbin-Johnson, B., and Aviran, S. (2015). "Controlling False Positive Rates in Methods for Differential Gene Expression Analysis using RNA-Seq Data." *BioRxiv*.

Rousseau, J. and Szabo, B. (2017). "Asymptotic Behaviour of the Empirical Bayes Posteriors Associated to Maximum Marginal Likelihood Estimator." *Annals of Statistics*, 45(2): 833–865.

Schwartzman, A. (2010). "Comment to Efron (2010)." *Journal of the American Statistical Association*, 105(491): 1059–1063.

Scott, J. G. and Berger, J. O. (2010). "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-selection Problem." *Annals of Statistics*, 38(5): 2587–2619.

Shepp, L. (1962). "Symmetric Random Walk." *Transactions of the American Mathematical Society*, 104(1): 144–153.

Sibuya, M. (1960). "Bivariate Extreme Statistics, I." *Annals of the Institute of Statistical Mathematics*, 11(3): 195–210.

Smemo, S. A. (2012). "Regulation of Heart Development via Transcriptional Enhancers and Epigenetic Modifications." Ph.D. Dissertation, University of Chicago.

Smyth, G. K. (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology*, 3(1): 1–25.

Stefanski, L. A. and Carroll, R. J. (1990). "Deconvolving Kernel Density Estimators." *Statistics*, 21(2): 169–184.

Stephens, M. (2017). "False Discovery Rates: A New Deal." *Biostatistics*, 18(2): 275–294.

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.

— (2003). "The Positive False Discovery Rate: A Bayesian Interpretation and the q-value." *Annals of Statistics*, 31(6): 2013–2035.

— (2007). "The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3): 347–368.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550.

Sun, L. and Stephens, M. (2018). "Solving the Empirical Bayes Normal Means Problem with Correlated Noise." *ArXiv e-prints*.

Sun, W. and Cai, T. T. (2009). "Large-Scale Multiple Testing Under Dependence." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(2): 393–424.

Sun, Y., Zhang, N. R., and Owen, A. B. (2012). "Multiple Hypothesis Testing Adjusted for Latent Variables, with an Application to the AGEMAP Gene Expression Data." *Annals of Applied Statistics*, 6(4): 1664–1688.

Szegõ, G. (1975). *Orthogonal Polynomials*. Providence, RI: American Mathematical Society, 4th edition.

The GTEx Consortium (2015). "The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science*, 348(6235): 648–660.

— (2017). "Genetic Effects on Gene Expression Across Human Tissues." *Nature*, 550: 204–213.

Thomas, D., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., and Armstrong, B. (1985). "The Problem of Multiple Inference in Studies Designed to Generate Hypotheses." *American Journal of Epidemiology*, 122(6): 1080–1095.

Tukey, J. W. (1991). "The Philosophy of Multiple Comparisons." *Statistical Science*, 6(1): 100–116.

Urbut, S. M., Wang, G., Carbonetto, P., and Stephens, M. (2019). "Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions." *Nature Genetics*, 51(1): 187–195.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). "Confounder Adjustment in Multiple Hypothesis Testing." *Annals of Statistics*, 45(5): 1863–1894.

Wang, L., Jia, P., Wolfinger, R. D., Chen, X., and Zhao, Z. (2011). "Gene Set Analysis of Genome-Wide Association Studies: Methodological Issues and Perspectives." *Genomics*, 98(1): 1–8.

Wang, W. and Stephens, M. (2018). "Empirical Bayes Matrix Factorization." *ArXiv e-prints*.

Wang, Y. (1996). "A Likelihood Ratio Test Against Stochastic Ordering in Several Populations." *Journal of the American Statistical Association*, 91(436): 1676–1683.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Secaucus, NJ: Springer-Verlag, 1st edition.

Weibull, W. (1939). "A Statistical Theory of the Strength of Materials." *Ingeniörs Vetenskaps Akademiens Handlingar*, 151.

Wilk, M. B. and Gnanadesikan, R. (1968). "Probability Plotting Methods for the Analysis of Data." *Biometrika*, 55(1): 1–17.

Wu, W. B. (2008). "On False Discovery Control Under Dependence." *Annals of Statistics*, 36(1): 364–380.

Xing, Z., Carbonetto, P., and Stephens, M. (2019). "Flexible Signal Denoising via Flexible Empirical Bayes Shrinkage." *ArXiv e-prints*.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). "Gene Ontology Analysis for RNA-seq: Accounting for Selection Bias." *Genome Biology*, 11(2): R14.

Zhu, X. and Stephens, M. (2018). "Large-Scale Genome-Wide Enrichment Analyses Identify New Trait-Associated Genes and Pathways Across 31 Human Phenotypes." *Nature Communications*, 9(1): 4361.