Nonparametric Estimation in a Compound Mixture Model and False Discovery Rate Control with Auxiliary Information

by

Zhaoyang Tian

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Wenguang Sun Associate Professor (University of Southern California)
Supervisors:	Dr. Pengfei Li Professor Dr. Kun Liang Associate Professor
Internal Member:	Dr. Paul Marriott Professor Dr. Mary Thompson Professor Emerita
Internal-External Member:	Dr. Dinghai Xu Associate Professor (Department of Economics)

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

A version of Chapter 2 of this thesis has been submitted for review in *Journal of Nonparametric Statistics*. A version of Chapter 3 of this thesis has been accepted for publication in *Biometrics*. Both papers are co-authored with my supervisors, and my contributions include deriving main theorems, implementing the proposed methods through R programming, performing the simulations and real data analysis, and writing the initial draft.

Abstract

In this thesis, we focus on two important statistical problems. The first is the nonparametric estimation in a compound mixture model with application to the malaria study. The second is the control of the false discovery rate in multiple hypothesis testing applications with auxiliary information.

Malaria can be diagnosed by the presence of parasites and symptoms (usually fever) due to parasites. In endemic areas, however, an individual may have fever attributable either to malaria or to other causes. Thus, the parasite level of an individual with fever follows a two-component mixture distribution, with the two components corresponding to malaria and nonmalaria individuals. Furthermore, the parasite levels of nonmalaria individuals can be characterized as a mixture of a zero component and a positive distribution, while the parasite levels of malaria individuals can only be positive. Therefore, the parasite level of an individual with fever follows a compound mixture model. In Chapter 2, we propose a maximum multinomial likelihood approach for estimating the unknown parameters/functions using parasite-level data from two groups of individuals: the first group only contains the malaria individuals, while the second group is a mixture of malaria and nonmalaria individuals. We develop an EM-algorithm to numerically calculate the maximum multinomial likelihood estimates and further establish their convergence rates. Simulation results show that the proposed maximum multinomial likelihood estimators are more efficient than existing nonparametric estimators. The proposed method is used to analyze a malaria survey data.

In many multiple hypothesis testing applications, thousands of null hypotheses are tested simultaneously. For each null hypothesis, usually a test statistic and the corresponding p-value are calculated. Traditional rejection rules work on p-values and hence

ignore the signs of the test statistics in two-sided tests. However, the signs may carry useful directional information in two-group comparison settings. In Chapter 3, we introduce a novel procedure, the signed-knockoff procedure, to utilize the directional information and control the false discovery rate in finite samples. We demonstrate the power advantage of our procedure through simulation studies and two real applications.

In Chapter 4, we further extend the signed-knockoff procedure to incorporate additional information from covariates, which are subject to missing. We propose a new procedure, the covariate and direction adaptive knockoff procedure, and show that our procedure can control the false discovery rate in finite samples. Simulation studies and real data analysis show that our procedure is competitive to existing covariate-adaptive methods.

In Chapter 5, we summarize our contributions and outline several interesting topics worthy of further exploration in the future.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. Pengfei Li and Prof. Kun Liang. I feel so grateful for their time and effort spent on me in the past five years. Thanks to their motivation and insightful guidance, I have learned so much during the PhD period.

Besides, I would like to thank the rest of my thesis committee members, Prof. Paul Marriott, Prof. Mary Thompson, Prof. Dinghai Xu and Prof. Wenguang Sun, for their patient review of my work. I also want to express my gratitude to all the faculty members and staff of the Department of Statistics and Actuarial Science, for their help during my PhD study.

Finally, I would like to thank my family and friends, for their kind support and accompaniment in my wonderful university life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

This thesis contributes to the nonparametric estimation in a compound mixture model and the false discovery rate (FDR) control in multiple hypothesis testing problem with auxiliary information such as the directional and covariate information. We first introduce the research background by giving three illustrating examples.

Example 1.1. *The first example is about the malaria study. Despite the major progress in the fight against malaria, it remains an acute public health problem, particularly in sub-Saharan Africa. Millions of people are at risk of malaria each year throughout the world (Bhatt et al., 2015). Malaria can be diagnosed by the presence of parasites and symptoms (usually fever) due to the parasites. However, malaria is not the only disease that is associated with fever. Further, in areas of high endemicity, asymptomatic parasitaemia is very common, and it should not be identified as clinical malaria. Thus, it is challenging to develop accurate diagnosis methods for malaria.*

The data from a cross-sectional survey of parasitaemia and fever for children less than a year old in a village in the Kilombero district of Tanzania (Kitua et al., 1996) is an excellent illustrating and motivating example. In one study of the survey, parasite levels in blood were collected in two seasons: dry season and wet season. The prevalence of malaria varies between these seasons. In the dry season, the prevalence is very low, and the parasite levels collected can be considered to come from nonmalaria individuals. In the wet season, the prevalence is high due to the presence of mosquitoes. However, children can tolerate parasites without symptoms and may have fever from other causes. Hence, the parasite levels collected in the wet season can be viewed as a mixture of those from people infected with malaria and those from nonmalaria individuals. One problem of interest is to estimate the malaria frequency in the wet season based on the parasite levels collected from both the wet and dry seasons (Smith, Schellenberg, and Hayes, 1994).

A unique feature of the data is that the parasite levels of some nonmalaria individuals are exactly zero. The parasite levels from individuals in the dry season (nonmalaria individuals) therefore follow a mixture of zero and a positive distribution. To ensure the model identifiability, we assume that the parasite levels for nonmalaria individuals have the same distribution in the two seasons. Then the parasite levels from individuals in the wet season follow a compound mixture (Qin and Leung, 2005). Specifically, let $X_1, \ldots, X_m$ be the parasite levels from the nonmalaria individuals (i.e. individuals from the dry season) and $Y_1, \ldots, Y_n$ be the parasite levels from the mixture of malaria and nonmalaria individuals (i.e. individuals from the wet season). Then we have

$$X_1, ..., X_m \sim F_X(x) = pI(x \geq 0) + (1-p)F_1(x), \tag{1.1}$$

$$Y_1, ..., Y_n \sim F_Y(y) = (1-\lambda)\left\{pI(y \geq 0) + (1-p)F_1(y)\right\} + \lambda F_2(y), \tag{1.2}$$

where p is the proportion of nonmalaria individuals with zero parasite levels, λ is the

proportion of malaria individuals in the wet season, and F_1 and F_2 are the cumulative distribution functions (cdfs) of positive parasite levels for the nonmalaria and malaria populations, respectively. We wish to estimate (λ, p, F_1, F_2) without making additional assumptions on F_1 and F_2. The details will be covered in Chapter 2.

Example 1.2. *The second example is from the prostate cancer study. Singh et al. (2002) present a dataset consisting of the gene expression levels of 6033 genes for 102 male individuals, including 52 prostate cancer patients and 50 normal control subjects. The goal of the study is to find the genes that are differentially expressed (DE) between the two groups of subjects. The identified genes might be further investigated for causal link to prostate cancer development (Efron, 2010).*

Let X_{ij} be the gene expression level for the ith gene and jth individual, $i = 1, \ldots, 6033$ and $j = 1, \ldots, 102$, with $j = 1, \ldots, 52$ for the cancer patients and $j = 53, \ldots, 102$ for the normal controls. Further, define $\bar{X}_{i1} = \frac{1}{52}\sum_{j=1}^{52} X_{ij}$ and $\bar{X}_{i2} = \frac{1}{50}\sum_{j=53}^{102} X_{ij}$, $i = 1, \ldots, 6033$. To check if the ith gene is DE between the two groups of subjects, we test

$$H_{0i} : \textit{gene } i \textit{ is not DE in the two samples.}$$

The classic two-sample t-test statistic is

$$t_i = \frac{\bar{X}_{i1} - \bar{X}_{i2}}{s_i},$$

where s_i is the estimate of the standard error of the numerator, that is,

$$s_i^2 = \frac{\sum_{j=1}^{52}\left(X_{ij} - \bar{X}_{i1}\right)^2 + \sum_{j=53}^{102}\left(X_{ij} - \bar{X}_{i2}\right)^2}{100}\left(\frac{1}{50} + \frac{1}{52}\right).$$

Under H_{0i}, t_i follows the t-distribution with 100 degrees of freedom and the corresponding p-value for its observed value is

$$p_i = 2\left\{1 - F_{100}\left(|t_i|\right)\right\},$$

where F_{100} is the cdf of t-distribution with 100 degrees of freedom.

To identify the DE genes in this prostate cancer dataset, we need to test more than 6000 hypotheses simultaneously and decide which null hypotheses to reject. Instead of controlling the type I error rate individually on each null hypothesis, scientists are interested in controlling the proportion of type I errors among the rejections. The expectation of the false discovery proportion is the FDR, proposed by Benjamini and Hochberg (1995).

As we can see in Example 1.2, the sign of t_i is ignored when we calculate the corresponding p-value. However, the signs may carry important directional information. A negative sign indicates that expression of the corresponding gene is potentially suppressed in prostate tumors, while a positive sign indicates a potentially elevated expression level. We aim to develop a powerful FDR control procedure which utilizes the directional information, and details are presented in Chapter 3.

Example 1.3. *The third example is a genetic study of psoriasis vulgaris disease. Jabbari et al. (2012) present a dataset consisting of the gene expression levels of 18151 genes measured on three pairs of lesional and nonlesional skin samples. Each pair of samples is collected from one patient by using the RNA sequencing (RNA-seq) method. Similarly to Example 1.2, the main goal of the study is to find the DE genes between lesional and nonlesional skin samples. Paired two-sample t-tests can be used to produce a list of t-test statistics $t_1, \ldots, t_{18151}$ with 8.8 effective degrees of freedom.*

In addition to the RNA-seq t-statistics, Liang (2019) identify two useful covariates to detect the DE genes. The first covariate is the length of the gene coding region, which is available for all 18151 genes. The second covariate is the test statistic from a related microarray study in Gudjonsson et al. (2010). Although there are many differences between these two studies such as the platform difference and different patient enrollment criteria,

the two studies share the same design of paired lesional and nonlesional skin samples. The microarray test statistics are shown to be very informative not only for the magnitude of the DE signals but also for the directions of the DE genes. Note that due to the platform difference, some RNA-seq informative genes are not measured by the microarray, causing missing data in the second covariate.

The procedure developed in Chapter 3 cannot incorporate the covariate information available in this example. We will address this challenge in Chapter 4.

Throughout the thesis, the finite mixture models and the FDR play important roles. Hence we review them in the next two sections.

1.2 Finite mixture models

In this section, we first give the formal definition of finite mixture models.

Definition (Finite mixture model). Suppose $f_j(\boldsymbol{x})$ is a probability density function (pdf) or probability mass function (pmf), $j = 1, \ldots, g$. If the pdf or the pmf of a random vector $\boldsymbol{X}$ can be written as

$$\sum_{j=1}^{g} \pi_j f_j(\boldsymbol{x}), \tag{1.3}$$

then $\boldsymbol{X}$ follows a g-component finite mixture model. Here $\pi_j \geq 0$ and $\sum_{j=1}^{g} \pi_j = 1$. We call g the number of components, π_j's the mixing proportions, and $f_j(\boldsymbol{x})$'s the component density or mass functions.

The finite mixture model can be interpreted as follows. Assume that $\boldsymbol{X}$ is drawn from a population consisting of g subpopulations. We use a categorical random variable Z to

denote the subpopulation membership. Here the possible values for Z are $1, \ldots, g$ and

$$\mathrm{P}(Z = j) = \pi_j, \quad j = 1, \ldots, g.$$

Further assume that the conditional pdf or pmf of $\boldsymbol{X}$ given $Z = j$ is $f_j(\boldsymbol{x})$, $j = 1, \ldots, g$. If Z is not observable, then the marginal distribution of $\boldsymbol{X}$ is just the mixture model in (1.3).

1.2.1 Parametric finite mixture model

When $f_j(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}_j)$ is known up to some unknown parameter $\boldsymbol{\theta}_j$, the finite mixture model in (1.3) becomes

$$\sum_{j=1}^{g} \pi_j f(\boldsymbol{x}; \boldsymbol{\theta}_j) \tag{1.4}$$

and is called a parametric mixture model. For example, if $f_j(x) = f(x; \mu_j, \sigma_j^2)$ is the pdf of normal distribution with mean μ_j and variance σ_j^2, then we get a g-component normal mixture model.

An important concept associated with the finite mixture model is identifiability, which is the foundation for estimation problem. If the parameters $(\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ are not identifiable, then their estimation becomes meaningless. The identifiability of parametric finite mixture models requires that for any two sets of parameters $(\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ and $(\pi_1^*, \ldots, \pi_{g*}^*, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{g*}^*)$ such that $\pi_j > 0$, $\pi_l^* > 0$, $\boldsymbol{\theta}_j$'s are distinct, and $\boldsymbol{\theta}_l^*$'s are distinct, if

$$\sum_{j=1}^{g} \pi_j f(\boldsymbol{x}; \boldsymbol{\theta}_j) = \sum_{j=1}^{g*} \pi_j^* f(\boldsymbol{x}; \boldsymbol{\theta}_j^*),$$

then $g = g^*$, $(\pi_1, \ldots, \pi_g) = (\pi_1^*, \ldots, \pi_{g*}^*)$, and $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g) = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{g*}^*)$ after permuting the component labels. Many commonly used finite mixture models are identifiable, such

as mixtures of normal distributions. We refer to Titterington, Smith, and Makov (1985) and McLachlan and Peel (2000) for more details.

In applications, the parameters $(\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g)$ are in general unknown and we need to estimate them based on some estimation method with the observed data. The likelihood-based method is one of the most popular methods for such purpose. Given n independent observations $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_i\}_{i=1}^n$ from the finite mixture model (1.3), the log-likelihood function is

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{g} \pi_j f(\boldsymbol{X}_i; \boldsymbol{\theta}_j) \right\},$$

where $\boldsymbol{\beta} = \{(\pi_j, \boldsymbol{\theta}_j)\}_{j=1}^g$. The maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is then defined as

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} l_n(\boldsymbol{\beta}).$$

Due to the complicated structure of the mixture model (1.3), in general the explicit form of MLE is not available. The expectation-maximization (EM) algorithm, proposed by Dempster, Laird, and Rubin (1977), is one of the most popular methods for numerically calculating the MLE. In the following, we provide some details on this algorithm.

In the EM framework, we view that the data consists of two parts, the observations $\boldsymbol{X}_i$ and the unobserved subpopulation memberships Z_i. Note that $Z_1, \ldots, Z_n$ can be seen as a random sample from the multinomial distribution with the cell probabilities being $(\pi_1, \ldots, \pi_g)$. So the complete log-likelihood function based on $\boldsymbol{\mathcal{Z}} = \{Z_i\}_{i=1}^n$ and $\boldsymbol{\mathcal{X}}$ is

$$l_n^c(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{g} I(Z_i = j) \left\{ \log \pi_j + \log f(\boldsymbol{X}_i; \boldsymbol{\theta}_j) \right\}.$$

The EM-algorithm is an iterative method. Each EM-iteration consists of an E-step and an M-step. Let $\boldsymbol{\beta}^{(0)} = \{(\pi_j^{(0)}, \boldsymbol{\theta}_j^{(0)})\}_{j=1}^g$ be the initial value of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{(k)} = \{(\pi_j^{(k)}, \boldsymbol{\theta}_j^{(k)})\}_{j=1}^g$

be the estimate of $\boldsymbol{\beta}$ after k iterations, $k = 1, 2, \ldots$.

In the E-step of the kth EM-iteration, we need to calculate

$$Q(\boldsymbol{\beta}; \boldsymbol{\beta}^{(k-1)}) = \mathrm{E}\left\{ l_n^c(\boldsymbol{\beta}) \,|\boldsymbol{\mathcal{X}}, \boldsymbol{\beta}^{(k-1)} \right\} = \sum_{i=1}^{n} \sum_{j=1}^{g} \mathrm{P}(Z_i = j|\boldsymbol{\mathcal{X}}, \boldsymbol{\beta}^{(k-1)}) \left\{ \log \pi_j + \log f(\boldsymbol{X}_i; \boldsymbol{\theta}_j) \right\},$$

where the expectations and probabilities are with respect to the conditional distribution of $\boldsymbol{\mathcal{Z}}$ given $\boldsymbol{\mathcal{X}}$ and substituting $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{(k-1)}$. As the complete log-likelihood function is a linear function of $I(Z_i = j)$, it suffices to calculate the posterior probability

$$w_{ij}^{(k)} = \mathrm{P}(Z_i = j|\boldsymbol{\mathcal{X}}, \boldsymbol{\beta}^{(k-1)}) = \frac{\pi_j^{(k-1)} f(\boldsymbol{X}_i; \boldsymbol{\theta}_j^{(k-1)})}{\sum_{l=1}^{g} \pi_l^{(k-1)} f(\boldsymbol{X}_i; \boldsymbol{\theta}_l^{(k-1)})}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, g$.

In the M-step of the kth EM-iteration, we update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(k)} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}; \boldsymbol{\beta}^{(k-1)}).$$

Due to the convenient structure of $Q(\boldsymbol{\beta}; \boldsymbol{\beta}^{(k-1)})$,

$$\pi_j^{(k)} = \sum_{i=1}^{n} w_{ij}^{(k)} / n,$$

$$\boldsymbol{\theta}_j^{(k)} = \arg\max_{\boldsymbol{\theta}_j} \sum_{i=1}^{n} w_{ij}^{(k)} \log f(\boldsymbol{X}_i; \boldsymbol{\theta}_j).$$

The E-step and M-step are then iteratively applied until

$$|l_n(\boldsymbol{\beta}^{(k+1)}) - l_n(\boldsymbol{\beta}^{(k)})|$$

is less than a pre-specified threshold.

A nice property of EM-algorithm is the so-called monotonicity property. That is,

$$l_n(\boldsymbol{\beta}^{(k+1)}) \geq l_n(\boldsymbol{\beta}^{(k)}).$$

This property guarantees that the EM-algorithm will eventually converge under some very general conditions. Wu (1983) further proves that under some regularity conditions, $\{\boldsymbol{\beta}^{(k)}\}_{k=0}^{\infty}$ will converge to a local maximum.

1.2.2 Extensions of parametric finite mixture models

Although the parametric finite mixture models are easy to interpret and their identifiability has been well studied, the parametric assumptions on $f_j(\boldsymbol{x})$ in some situations may be unrealistic. The model misspecification on $f_j(\boldsymbol{x})$ may lead to misleading results (Pommeret and Vandekerkhove, 2018). Because of that, researchers have tried to relax the parametric assumption on $f_j(\boldsymbol{x})$.

In the univariate case, Bordes, Mottelet, and Vandekerkhove (2006) consider the assumption that $f_j(x) = f(x - \theta_j)$, where the form of $f(x)$ is unknown. Under this assumption, the mixture model in (1.3) becomes

$$\sum_{j=1}^{g} \pi_j f(x - \theta_j). \tag{1.5}$$

The identifiability of (1.5) is discussed in Bordes, Mottelet, and Vandekerkhove (2006) and Hunter, Wang, and Hettmansperger (2007). Bordes, Chauveau, and Vandekerkhove (2007) propose a kernel-based method to estimate $(\pi_1, \ldots, \pi_g, \theta_1, \ldots, \theta_g)$ and f. However, its asymptotic properties still remain unknown. Hunter, Wang, and Hettmansperger (2007) develop a minimum distance estimator of $(\pi_1, \ldots, \pi_g, \theta_1, \ldots, \theta_g)$. Its asymptotic normality is studied in Balabdaoui (2017). More recent development can be found in Xiang, Yao, and Yang (2019).

Qin et al. (2014) and Yu, Li, and Qin (2019) consider a class of univariate nonparametric mixture models, in which each observation is from a finite mixture model with the mixing

proportions known. More specifically, let $(X_i, \pi_{i1}, \ldots, \pi_{ig})$, $i = 1, \ldots, n$ be the independent and identically distributed (iid) observations such that $\pi_{ij} \geq 0$ and $\sum_{j=1}^{g} \pi_{ij} = 1$. Further, the pdf of X_i conditioning on $(\pi_{i1}, \ldots, \pi_{ig})$ is given by

$$\sum_{j=1}^{g} \pi_{ij} f_j(x). \tag{1.6}$$

Here $f_1(x), \ldots, f_g(x)$ are pdfs and their forms are unknown. This class of models have wide applications in genetic and epidemiology studies. Yu, Li, and Qin (2019) provide the sufficient conditions to ensure the identifiability of (1.6) and propose a maximum smoothed likelihood method to estimate $f_1, \ldots, f_g$. They further study the L_1 convergence of the proposed estimators. Qin et al. (2014) propose a maximum binomial likelihood to estimate the component cdfs. Their method can also be applied to the censored data. However the asymptotic properties of their estimator still remain unknown.

In the multivariate case, a commonly used assumption is that $f_j(\boldsymbol{x})$ is equal to the product of its marginal densities. That is, conditional on the subpopulation membership Z, the coordinates of $\boldsymbol{X}$, denoted as $X_1, \ldots, X_p$, are independent. Under this assumption, the pdf of the random vector $\boldsymbol{X}$ can be written as

$$f(\boldsymbol{x}) = \sum_{j=1}^{g} \pi_j \prod_{k=1}^{p} f_{jk}(x_k). \tag{1.7}$$

Allman, Matias, and Rhodes (2009) show that Model (1.7) is identifiable when $p \geq 3$ and $f_{1k}, \ldots, f_{gk}$ are linearly independent for $k = 1, \ldots, p$. Hall and Zhou (2003) propose an minimum distance method to estimate $\pi_j's$ and $f_{jk}'s$ when $g = 2$ and study the asymptotic properties of the corresponding estimators. Benaglia, Chauveau, and Hunter (2009) suggest an EM-like algorithm to obtain the estimates of $\pi_j's$ and $f_{jk}'s$. However, they do not discuss the asymptotic properties of the estimators. Levine, Hunter, and Chauveau (2011) further notice that the EM-like algorithm does not have the monotonicity property. Motivated

by this fact, Levine, Hunter, and Chauveau (2011) introduce the maximum smoothed likelihood estimators for $\pi_j's$ and $f_{jk}'s$ and further propose a maximization-minimization algorithm to numerically calculate the estimates. Recent developments along Model (1.7) and other generalizations of Model (1.7) can be found in Chauveau, Hunter, and Levine (2015) and Zheng and Wu (2019).

An alternative and useful tool to extend and generalize the parametric finite mixture model in (1.4) is the use of local mixture model (Marriott, 2002), which allows for the unknown number of discrete components or continuous mixtures. For illustration purpose, we will concentrate on the univariate case, in which the mixture model in (1.4) can be written as

$$\sum_{j=1}^{g} \pi_j f(\boldsymbol{x}; \theta_j) = \int f(x; \theta) dQ(\theta),$$

where

$$Q(\theta) = \sum_{j=1}^{g} \pi_j I(\theta_j \leq \theta).$$

The main idea of the local mixture model is to assume that $Q(\theta)$ is close to a point mass function at some fixed point θ_0 and then approximate $f(x; Q)$ by

$$\int f(x; \theta) dQ(\theta) \approx f(x; \theta_0) + \sum_{k=2}^{K} \alpha_k \frac{\partial^k f(x; \theta_0)}{\partial \theta^k}, \tag{1.8}$$

where $\alpha_2, \ldots, \alpha_K$ are unknown parameters. The right hand side of (1.8) is called the local mixture model of $\int f(x; \theta) dQ(\theta)$ with order K and it only involves K unknown parameters. The traditional methods can then be applied to estimate the unknown parameters. After Marriott (2002), the local mixture model has wide applications in many areas such as measurement error modelling, Bayesian prediction, and influence analysis. We refer to Maroufy (2016) and Maroufy and Marriott (2017) for the recent development of local mixture models and the generalization of the local mixture models.

1.3 Multiple hypothesis testing and false discovery rate

1.3.1 False discovery rate and family-wise error rate

In multiple hypothesis testing problems, such as in Examples 1.2 and 1.3, a large number of null hypotheses are tested simultaneously. The concept of the FDR plays a very important role in such problems, and we give a brief introduction to the FDR in this section.

Suppose there are n null hypotheses which are tested at the same time. Table 1.1 summarizes the outcome after applying a certain rejection rule to the n hypothesis tests. Here V is the number of true null hypotheses which are falsely rejected, i.e., the number of type I errors, S is the number of false null hypotheses which are correctly rejected, U is the number of true null hypotheses which are correctly accepted, and T is the number of false null hypotheses which are falsely accepted, i.e., the number of type II errors.

Table 1.1: Results of a multiple hypothesis testing problem

	True null hypotheses	False null hypotheses
Rejection	V	S
Acceptance	U	T

In multiple testing problems, if we only control the type I error rate individually for each null hypothesis, the expected number of false positives, $\mathrm{E}(V)$, will increase linearly with the total number of hypotheses n. Hence a sensible multiple hypothesis testing error measure is required to control this multiplicity effect. In the literature, there are two popular measures for such purpose. The traditional one is the family–wise error rate (FWER), which is the

probability of making at least one type I error among n hypothesis tests, i.e.,

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

The more recent one is the FDR (Benjamini and Hochberg, 1995), which is the expectation of the false discovery proportion (FDP), the proportion of type I errors among all rejected null hypotheses, i.e.,

$$\text{FDR} = \text{E}\left(\text{FDP}\right) = \text{E}\left\{\frac{V}{\max(V+S,1)}\right\}.$$

The effect of "$\max(V+S,1)$" is to avoid 0 in the denominator.

In the literature, there are many discussions regarding the application scope of both measures, and we summarize them below. More details can be found in Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001).

(a) The control of the FWER is more important when the overall conclusion is likely to be wrong due to a type I error committed by one of the hypothesis tests. For example, consider the case where we compare multiple new treatments with the standard one and conduct one hypothesis test for each comparison. The goal of the study is to check if any of the new treatments performs better than the standard treatment. If any of the hypothesis tests commits a type I error, a wrong conclusion will be made that one of the new treatments shows better performance than the standard one. Controlling the FWER avoids exaggerating difference in treatment effects.

(b) As the FDR is always no larger than the FWER (Benjamini and Hochberg, 1995), a procedure controls the FDR can be less stringent compared with methods that control the FWER at the same level. That is, we can expect a power gain from controlling the FDR instead of the FWER. For example, consider the case where we compare

characteristics of a treatment group and a control group to investigate the potential treatment effect. Multiple experiments are conducted to test various aspects of the treatment effect. We want to find as many aspects of the treatment effect as possible rather than to draw an overall conclusion about the existence of *any* treatment effect. In this case, controlling the FDR will be more appropriate.

In my thesis, we are interested in controlling the FDR. The goals of the prostate cancer study in Example 1.2 and the psoriasis vulgaris study in Example 1.3 are to identify the DE genes, and each gene can be viewed as one aspect of the human transcriptome. We want to identify as many DE genes as possible to serve as potential biomarkers for disease diagnostic purpose. The identified DE genes will be subject to further investigation, and we wish to have a large proportion of the DE genes among declared discoveries for future research. Controlling the FWER, the probability of any discovery being false, is too stringent for the scientific purpose.

1.3.2 Existing false discovery rate control methods

Benjamini and Hochberg (1995) not only formally define the concept of the FDR, but also propose a linear step-up procedure, which we will refer to as the Benjamini–Hochberg (BH) procedure. This procedure can control the FDR in finite samples under the condition that the true null test statistics are independent of each other and independent of alternative test statistics. The independence condition, denoted as the *null independence condition* in this thesis, is used by many other procedures in the literature, such as Storey, Taylor, and Siegmund (2004) and Liang and Nettleton (2012), as well as our methods in Chapters 3 and 4.

Let π_0 be the proportion of true null hypotheses. The BH procedure actually controls the FDR at the level $\pi_0\alpha$. This suggests a natural way to improve power by applying the BH procedure at a higher nominal FDR level of $\alpha/\hat{\pi}_0$, where $\hat{\pi}_0$ is an estimate of π_0. Let n be the total number of null hypotheses. Storey, Taylor, and Siegmund (2004) propose a π_0-estimator as

$$\hat{\pi}_0(\lambda) = \frac{\#\{i:\ p_i > \lambda\} + 1}{n(1-\lambda)},$$

where p_i are the p-values. The estimator is based on a fixed tuning parameter λ between 0 and 1. They show that an adaptive procedure based on this π_0-estimator can control the FDR and is more powerful than the BH procedure. Liang and Nettleton (2012) propose a class of dynamic adaptive procedures whose tuning parameters are determined by the data. Recently, MacDonald, Liang, and Janssen (2019) show that these dynamic adaptive procedures control the FDR in finite samples.

In practical applications, there are typical auxiliary information available in addition to a list of p-values. One type of the auxiliary information is the directional information illustrated in Example 1.2. Sun and Cai (2007) show that from a decision-theoretic point of view, the p-value based procedures are inadmissible. They propose to work with z-values, a statistic which has a one-to-one relationship with the test statistic that preserve the directional information. Subsequently, some direction-adaptive methods such as Orr, Liu, and Nettleton (2014) and Zhao and Fung (2016) are developed by grouping the p-values according to the signs of the corresponding test statistics. Orr, Liu, and Nettleton (2014) propose to control the FDR in the positive and negative groups separately and combine the two rejection results. Zhao and Fung (2016) suggest a weighted BH procedure with different weights for the two groups. Simulation studies show that by utilizing the directional information, the direction-adaptive methods outperform the traditional p-value based methods. However, none of them are shown to control the FDR in finite samples.

Other types of auxiliary information also abound. Li and Barber (2017) consider the multiple testing problem where the null hypotheses are ordered for rejection. Li and Barber (2019) study various structured settings. For example, in a multiple testing application where each null hypothesis corresponds to a gene, true signals may tend to co-occur in the same genetic pathway. Li and Barber (2019) can utilize this structure information by grouping genes according to their genetic pathway and assigning different group weights. Recently, methods are developed to handle generic covariates. For example, Ignatiadis and Huber (2018) can utilize a single covariate that is either categorical or continuous. Lei and Fithian (2018) propose an adaptive p-value thresholding procedure that can incorporate multiple covariates. All of these methods above are shown to control the FDR in finite samples. Note that it is possible for the group-adaptive methods, such as Li and Barber (2019) and Lei and Fithian (2018), to utilize the directional information by grouping null hypotheses according to the signs of test statistics. However, it is not clear whether they can efficiently use the directional information.

1.4 Contribution and outline of our work

In this thesis, we propose novel statistical methods with solid theoretical support for the research problems motivated by Examples 1.1, 1.2 and 1.3.

Chapter 2 focuses on the research problem motivated by Example 1.1 and considers the situation where there is no assumptions imposed on F_1 and F_2 except that they are cdfs. As we will review in Section 2.1, the existing methods for analyzing the data from Models (1.1) and (1.2) are based on some parametric or semi-parametric assumption on the densities of F_1 and F_2. They may not be robust to model misspecification. The existing nonparametric methods, reviewed in Section 1.2.2, do not take the special structure of (1.1)

and (1.2) into consideration. We propose a maximum multinomial likelihood approach to estimate the unknown parameters and functions (λ, p, F_1, F_2) in Models (1.1) and (1.2). An efficient EM-algorithm is further developed to compute the estimates. We establish the asymptotic properties of the proposed maximum multinomial likelihood estimator of (λ, p, F_1, F_2). Simulation studies and the analysis of Example 1.1 show superiority of the proposed method over its competitors.

Chapter 3 focuses on the control of the FDR with the directional information. Traditional methods are based on p-values and ignore the directional information. In Chapter 3, we propose a novel method, the signed-knockoff procedure, to utilize the directional information and control the FDR in finite samples. Simulation studies show our power advantage over competing methods. We also analyze Example 1.2 to demonstrate the advantages of the signed-knockoff procedure.

In Chapter 4, we extend the signed-knockoff procedure to incorporate the additional information from the covariates, which are subject to missing. We also show that the new method, the covariate and direction adaptive knockoff procedure, can control the FDR in finite samples. Simulation studies show that our procedure has a better power performance than existing covariate-adaptive methods. We further analyze Example 1.3 and show the superiority of our method over existing methods.

Chapter 5 concludes the thesis with a brief summary and provides some possible directions for future research.

Chapter 2

Maximum multinomial likelihood estimation in a compound mixture model

2.1 Introduction

Chapter 2 is devoted to developing the statistical method for the research problem motivated by Example 1.1. Recall that the dataset contains parasite levels collected in two different seasons. The parasite levels collected in the dry season are denoted as $X_1, \ldots, X_m$, and those in the wet season are denoted as $Y_1, \ldots, Y_n$. They can be modeled with

$$X_1, ..., X_m \sim F_X(x) = pI(x \geq 0) + (1-p)F_1(x), \tag{2.1}$$

$$Y_1, ..., Y_n \sim F_Y(y) = (1-\lambda)\left\{pI(y \geq 0) + (1-p)F_1(y)\right\} + \lambda F_2(y), \tag{2.2}$$

where p is the proportion of nonmalaria individuals with zero parasite levels, λ is the proportion of malaria individuals in the wet season, and F_1 and F_2 are the cdfs of positive parasite levels for the nonmalaria and malaria populations, respectively. In Section 2.6.1, we show that (λ, p, F_1, F_2) are identifiable if $0 < \lambda, p < 1$ and $F_1(x) = F_2(x) = 0$ for any $x \leq 0$. We wish to estimate (λ, p, F_1, F_2) without making additional assumptions.

There are at least two important applications for the estimation of (λ, p, F_1, F_2). First, the estimation of λ, the estimated malaria frequency in the mixture sample, will be helpful for the implementation of intervention programmes (Vounatsou, Smith, and Smith, 1998). Second, the estimator of (λ, p, F_1, F_2) can be used to construct an estimator of the posterior probability that an individual in the mixture sample is infected with malaria given his/her positive parasite level. See Section 2.5 for more details.

The compound structure seen in (2.1) and (2.2) does not appear only in the malaria study. It also appears in biomedical research and diagnostic practice, especially when there is no gold standard for the true positives. See Smith, Schellenberg, and Hayes (1994) and Qin (2017) for more examples. Using (2.1) and (2.2), Smith, Schellenberg, and Hayes (1994) consider a model-based approach in which the relationship between the parasite level and malaria status (malaria or nonmalaria) is modelled through a logistic regression. Qin and Leung (2005) observe that this relationship is equivalent to a density ratio model assumption on the pdfs of F_1 and F_2. They further propose a semiparametric likelihood method to estimate (λ, p) under the density ratio model assumption. Vounatsou, Smith, and Smith (1998) consider the estimation of (λ, p) under a setup where the pdf ratio of F_2 and F_1 is a monotonically increasing function of the parasite level. They suggest first grouping the positive parasite levels into several ordered categories and then estimating the unknown parameters with a Bayesian method. In summary, the existing methods rely on certain model assumptions for the pdfs of F_1 and F_2, which may not be robust to model

misspecification on F_1 and F_2. Further, Vounatsou, Smith, and Smith (1998)'s method requires ad-hoc grouping of the positive parasite levels.

In this chapter, we concentrate on the situation where there is no assumption on F_1 and F_2 except their identifiability conditions. Under this setup, (λ, p) can be naturally estimated by the binomial estimator (Qin and Leung, 2005) as follows:

$$\tilde{p} = \frac{m_0}{m}, \ \tilde{\lambda} = 1 - \frac{n_0}{n\tilde{p}}, \tag{2.3}$$

where m_0 and n_0 are the number of zeros in $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$, respectively. With the estimator of (λ, p), (F_1, F_2) can then be naturally estimated with

$$\tilde{F}_1(x) = \tilde{F}_{X+}(x), \quad \tilde{F}_2^*(x) = (\tilde{\lambda}^*)^{-1} \left\{ \tilde{F}_{Y+}(x) - (1 - \tilde{\lambda}^*) \tilde{F}_1(x) \right\}, \tag{2.4}$$

where $\tilde{\lambda}^* = \tilde{\lambda}/\{1 - \tilde{p}(1 - \tilde{\lambda})\}$ and $\tilde{F}_{X+}(x)$ and $\tilde{F}_{Y+}(x)$ are the empirical cdfs of the positive values in $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$, respectively. To ensure the monotonicity of the estimator of $F_2(x)$, one can apply isotonic regression to $\tilde{F}_2^*$ to obtain an estimator $\tilde{F}_2(x)$ of F_2, which is guaranteed to be an increasing function. See Section 2.4 for more details.

The estimator $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$ is quite natural and simple to implement. However, the estimator $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1)$ does not use the compound structure in (2.2). Further, the estimator $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$ does not have a likelihood interpretation and may not be fully efficient. There is room for improvement. Motivated by the idea of the binomial likelihood (Qin et al., 2014), we propose a maximum multinomial likelihood approach to estimate (λ, p, F_1, F_2) simultaneously. This method takes the zero-inflated and compound structure in (2.1) and (2.2) into account. The resulting maximum multinomial likelihood estimator is shown to be consistent. Its asymptotic properties are also investigated. Simulation studies show that the proposed estimator is more efficient than $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$.

The rest of Chapter 2 is organized as follows. In Section 2.2, we discuss the application of the empirical likelihood method (Owen, 2001) and show that it fails to produce

a consistent estimator for (λ, F_2). We further review the idea of the binomial likelihood and its recent development. In Section 2.3, we propose the maximum multinomial likelihood estimator for (λ, p, F_1, F_2) and establish its asymptotic properties. We also develop an EM-algorithm to numerically calculate the estimate. In Section 2.4, we use simulation studies to compare the proposed maximum multinomial likelihood estimator with the binomial estimator $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$, the maximum empirical likelihood estimator of (λ, p, F_1, F_2), Qin and Leung (2005)'s semiparametric estimator and the parametric estimator, and then demonstrate the superiority of our method. Section 2.5 applies our method to the real dataset in Example 1.1. For convenience of presentation, all the technical details are given in Section 2.6.

2.2 Empirical likelihood and maximum binomial likelihood

2.2.1 Failure of empirical likelihood method

As a nonparametric likelihood method, the empirical likelihood method (Owen, 2001) seems a natural choice to estimate (λ, p, F_1, F_2). However, we will show that it fails to produce a consistent estimator for (λ, F_2).

Let m_+ and n_+ be the numbers of positive values in $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$, respectively. Without loss of generality, we use $X_1, \ldots, X_{m_+}$ to denote positive values in $\{X_1, \ldots, X_m\}$ and $Y_1, \ldots, Y_{n+}$ to denote positive values in $\{Y_1, \ldots, Y_n\}$. For illustration, we consider the case where there is no tie in $\{X_1, \ldots, X_{m_+}, Y_1, \ldots, Y_{n_+}\}$.

Denote $a_i = dF_1(X_i)$ for $i = 1, \ldots, m_+$ and $a_{j+m_+} = dF_1(Y_j)$ and $b_j = dF_2(Y_j)$ for

$j = 1, \ldots, n_+$. Let $\mathbf{a} = (a_1, \ldots, a_{m_+ + n_+})^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_{n_+})^{\mathrm{T}}$. Following the empirical likelihood principle (Owen, 2001), the log empirical likelihood of $(\lambda, p, \mathbf{a}, \mathbf{b})$ is

$$\begin{aligned}&\tilde{l}_{el}(\lambda, p, \mathbf{a}, \mathbf{b}) \\ =&m_0 \log p + \sum_{i=1}^{m_+} \log\{(1-p)a_i\} + n_0 \log\{(1-\lambda)p\} + \sum_{j=1}^{n_+} \log\left\{(1-\lambda)(1-p)a_{j+m_+} + \lambda b_j\right\}.\end{aligned}$$

The maximum empirical likelihood estimator of $(\lambda, p, \mathbf{a}, \mathbf{b})$ is

$$(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{\mathbf{a}}_{el}, \tilde{\mathbf{b}}_{el}) = \arg \max_{\lambda,\ p,\ \mathbf{a},\ \mathbf{b}} \tilde{l}_{el}(\lambda, p, \mathbf{a}, \mathbf{b})$$

subject to the constraints

$$\lambda \in [0,1], \quad p \in [0,1], \quad a_i \geq 0, \quad b_j \geq 0, \quad \sum_{i=1}^{m_+ + n_+} a_i = 1, \quad \sum_{j=1}^{n_+} b_j = 1.$$

The maximum empirical likelihood estimators of $F_1(x)$ and $F_2(x)$ are

$$\tilde{F}_{1,el}(x) = \sum_{i=1}^{m_+} \tilde{a}_{i,el} I(X_i \leq x) + \sum_{j=1}^{n_+} \tilde{a}_{j+m_+,el} I(Y_j \leq x), \quad \tilde{F}_{2,el}(x) = \sum_{j=1}^{n_+} \tilde{b}_{j,el} I(Y_j \leq x).$$

The following proposition gives the closed form of $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{F}_{1,el}, \tilde{F}_{2,el})$. The proof is in Section 2.6.2.

Proposition 2.1. *Suppose there is no tie in the positive observations in* $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. *Then we have*

$$\tilde{\lambda}_{el} = \frac{n_+}{n}, \quad \tilde{p}_{el} = \frac{m_0 + n_0}{m + n_0},$$

and

$$\tilde{F}_{1,el}(x) = \frac{1}{m_+} \sum_{i=1}^{m_+} I(X_i \leq x), \quad \tilde{F}_{2,el}(x) = \frac{1}{n_+} \sum_{j=1}^{n_+} I(Y_j \leq x).$$

Proposition 2.1 implies that

$$\tilde{\lambda}_{el} \to (1-\lambda)(1-p)+\lambda, \quad \tilde{F}_{2,el}(x) \to (1-\lambda^*)F_1(x)+\lambda^* F_2(x)$$

in probability, where $\lambda^* = \lambda/\{1-p(1-\lambda)\}$. Therefore, when $0<\lambda<1$ and $0<p<1$, $\tilde{\lambda}_{el}$ and $\tilde{F}_{2,el}(x)$ are not consistent estimators for λ and $F_2(x)$.

2.2.2 General idea of binomial likelihood

Our method is motivated by the binomial likelihood approach proposed by Qin et al. (2014). We briefly review the main idea by considering a simple case. Suppose we have a random sample $Z_1, \ldots, Z_n$ from a population with cdf $G(x)$. The binomial likelihood is motivated by the fact that for any given t,

$$I(Z_i \le t) \sim \text{Bin}\Big(1; G(t)\Big),$$

where "Bin" denotes the binomial distribution. Summing the log-likelihood of $\{I(Z_i \le t)\}_{i=1}^n$ for the given t gives

$$\tilde{l}_B(G;t) = \sum_{i=1}^{n} I\left(Z_i \le t\right) \log G(t) + \left\{ n - \sum_{i=1}^{n} I\left(Z_i \le t\right) \right\} \log \bar{G}(t),$$

where $\bar{G}(t) = 1 - G(t)$.

We can arbitrarily choose the values of t and then take the sum of $\tilde{l}_B(G;t)$ over the chosen values of t or we can integrate $\tilde{l}_B(G;t)$ to get an objective function which does not depend on t. Let $Z_{(1)} < \ldots < Z_{(J)}$ be distinct values of the observed random sample $Z_1, \ldots, Z_n$. Qin et al. (2014) suggested choosing the values of t to be $\{Z_{(j)},\ j = 1, \ldots, J\}$ and then taking the sum of $\tilde{l}_B(G;t)$ over these J values. This gives the binomial likelihood as follows:

$$l_B(G) = \sum_{j=1}^{J} \tilde{l}_B(G; Z_{(j)}).$$

The maximum binomial likelihood estimator of $G(x)$ maximises $l_B(G)$ subject to the constraint that $G(x)$ is a cdf. Qin et al. (2014) show that the empirical cdf of $X_1, \ldots, X_n$ is a maximum binomial likelihood estimator of $G(x)$. They further apply this method to estimate the component cdfs of mixture models with known proportions with censored data. Lee et al. (2016) apply the binomial likelihood method to estimate the distributional treatment effects in instrumental variable models. To avoid dealing with potentially vanishing probabilities in the boundaries, they recommend using a subset of $Z_{(1)} < \ldots < Z_{(J)}$. They also establish the asymptotic properties of the maximum binomial likelihood estimator.

Naturally, we may apply the binomial likelihood method to the data generated from Models (2.1) and (2.2). However, the binomial likelihood does not take the zero-inflated structure into account. Specifically, if we apply the binomial likelihood method to the data from Model (2.1), for any given $t > 0$

$$P(X_i \leq t) = p + (1 - p)F_1(t),$$

which implies that p and F_1 are tangled together in the binomial likelihood. A similar problem occurs when the binomial likelihood method is applied to the data from Model (2.2).

2.3 Maximum multinomial likelihood estimation for compound mixtures

2.3.1 Maximum multinomial likelihood

Motivated by the idea of the binomial likelihood, we now propose a maximum multinomial likelihood to estimate (λ, p, F_1, F_2) simultaneously. This method takes the zero-inflated

and compound structure in (2.1) and (2.2) into account. We first introduce some notation. For any $t > 0$, let

$$\mathbf{M}_i(t) = \big(m_{i0}, m_{i1}(t), m_{i2}(t)\big)^{\mathrm{T}} = \Big(I(X_i = 0), I(0 < X_i \le t), I(X_i > t)\Big)^{\mathrm{T}}, \qquad (2.5)$$

for $i = 1, \ldots, m$, and

$$\mathbf{N}_j(t) = \Big(n_{j0}, n_{j1}(t), n_{j2}(t)\big)^{\mathrm{T}} = \big(I(Y_j = 0), I(0 < Y_j \le t), I(Y_j > t)\big)^{\mathrm{T}}, \qquad (2.6)$$

for $j = 1, \ldots, n$.

The multinomial likelihood method is motivated by the fact that

$$\mathbf{M}_i(t) \sim \mathrm{Multi}\Big(1; p, (1-p)F_1(t), (1-p)\bar{F}_1(t)\Big), \quad i = 1, \ldots, m$$

and

$$\mathbf{N}_j(t) \sim \mathrm{Multi}\Big(1; (1-\lambda)p, (1-\lambda)(1-p)F_1(t)+\lambda F_2(t), (1-\lambda)(1-p)\bar{F}_1(t)+\lambda\bar{F}_2(t)\Big), \ j = 1, \ldots, n,$$

where "Multi" denotes the multinomial distribution. Summing the log-likelihoods of $\{\mathbf{M}_i(t)\}_{i=1}^m$ and $\{\mathbf{N}_j(t)\}_{j=1}^n$ gives

$$\begin{aligned}\tilde{l}_M(\lambda, p, F_1, F_2; t) = &\sum_{i=1}^{m} \big[m_{i0}\log p + m_{i1}(t)\log\{(1-p)F_1(t)\} + m_{i2}(t)\log\{(1-p)\bar{F}_1(t)\}\big] \\ &+ \sum_{j=1}^{n} \Big[n_{j0}\log\{(1-\lambda)p\} + n_{j1}(t)\log\{(1-\lambda)(1-p)F_1(t) + \lambda F_2(t)\} \\ &\qquad + n_{j2}(t)\log\{(1-\lambda)(1-p)\bar{F}_1(t) + \lambda\bar{F}_2(t)\}\Big].\end{aligned}$$

Let $t_1 < \ldots < t_k$ be the distinct values of the positive values in $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$. For a given $q \in (0, 0.5)$, define

$$I_q = \big\{\lceil kq\rceil, \lceil kq\rceil + 1, \ldots, \lceil k(1-q)\rceil\big\},$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x. For asymptotic purposes, we suggest choosing the values of t as $\{t_h | h \in I_q\}$ and then taking the sum of $\tilde{l}_M(\lambda, p, F_1, F_2; t)$ over these values to obtain the multinomial likelihood

$$\begin{aligned} l_M(\lambda, p, F_1, F_2) = & \sum_{h \in I_q} \sum_{i=1}^{m} \left[m_{i0} \log p + m_{i1}(t_h) \log\{(1-p)F_1(t_h)\} + m_{i2}(t_h) \log\{(1-p)\bar{F}_1(t_h)\} \right] \\ & + \sum_{h \in I_q} \sum_{j=1}^{n} \Big[n_{j0} \log\{(1-\lambda)p\} + n_{j1}(t_h) \log\{(1-\lambda)(1-p)F_1(t_h) + \lambda F_2(t_h)\} \\ & \qquad + n_{j2}(t_h) \log\{(1-\lambda)(1-p)\bar{F}_1(t_h) + \lambda \bar{F}_2(t_h)\} \Big]. \end{aligned} \tag{2.7}$$

A more detailed discussion of the choice of the t values will be given in Section 2.3.2. We comment that in the definition of l_M, the dependence structures in $\{\mathbf{M}_i(t_h), h \in I_q\}$ and in $\{\mathbf{N}_j(t_h), h \in I_q\}$ are ignored. This definition is motivated from the composite likelihood in Lindsay (1988). A composite likelihood generally leads to model robustness and a simplified numerical solution.

Further, let k_q be the number of indexes in I_q and $J(x, y) = x \log y + (1-x) \log(1-y)$, $x, y \in [0, 1]$. After some algebra work, the multinomial likelihood can be rewritten as

$$\begin{aligned} l_M(\lambda, p, F_1, F_2) = & \, m k_q J\left(\tilde{p}, p\right) + n k_q J\left((1-\tilde{\lambda})\tilde{p}, (1-\lambda)p\right) \\ & + \sum_{h \in I_q} \left\{ m_+ J\left(\tilde{F}_1(t_h), F_1(t_h)\right) + n_+ J\left(\tilde{F}_{Y+}(t_h), F_{Y+}(t_h)\right) \right\}, \end{aligned} \tag{2.8}$$

where $F_{Y+}(x) = (1-\lambda^*)F_1(x) + \lambda^* F_2(x)$, $\tilde{p}$ and $\tilde{\lambda}$ are defined in (2.3), and $\tilde{F}_1$ and $\tilde{F}_{Y+}$ are the empirical cdfs of the positive values in $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$, respectively.

The maximum multinomial likelihood estimator of (λ, p, F_1, F_2) is defined as

$$(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2) = \arg \max_{(\lambda, p, F_1, F_2) \in \Theta} l_M(\lambda, p, F_1, F_2),$$

where $\Theta = \{(\lambda, p, F_1, F_2) | \lambda, p \in [0, 1], \ F_1 \text{ and } F_2 \text{ are cdfs}\}$.

2.3.2 Asymptotic properties

The asymptotic properties of $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ rely on the following regularity conditions:

A1. $0 < \lambda < 1$ and $0 < p < 1$.

A2. $n/N = \rho \in (0, 1)$, where $N = n + m$.

A3. $F_1(x)$ and $F_2(x)$ have the same support. Further, $F_1(x)$ and $F_2(x)$ are absolutely continuous and strictly increasing on the support.

Theorem 2.1. *Suppose A1–A3 are satisfied. Then we have*

(a) $\hat{\lambda} - \lambda = O_p\left(N^{-1/2}\right), \quad \hat{p} - p = O_p\left(N^{-1/2}\right)$;

(b) $k_q^{-1} \sum_{h \in I_q} \{\hat{F}_1(t_h) - F_1(t_h)\}^2 = O_p(N^{-1}), \quad k_q^{-1} \sum_{h \in I_q} \{\hat{F}_2(t_h) - F_2(t_h)\}^2 = O_p(N^{-1})$.

For convenience of presentation, the proof of Theorem 2.1 is in Section 2.6.3. Here we make some comments on the regularity conditions and the choice of the t values in (2.7).

Condition A1 states that the true values of λ and p are interior points of their parameter space. Condition A2 requires that the sample size ratio in the two groups is a constant. Condition A3 and the choice of the t values as $\{t_h : h \in I_q\}$ together ensure that $\tilde{F}_{Y+}(t_h)$ and $\tilde{F}_1(t_h)$ are uniformly away from 0 and 1 in probability for all $h \in I_q$. This guarantees that $l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ can be bounded above and below by quadratic functions. As a result, we can derive the asymptotic results in Theorem 2.1. A similar idea has been used by Lee et al. (2016), who use the maximum binomial method to estimate the distributional treatment effects in instrumental variable models. In practice, we recommend a small value for q; we used $q = 0.001$ in our simulation study and real-data analysis.

2.3.3 EM-algorithm

In (2.8), $F_1(x)$ and $F_2(x)$ are tangled together, which makes the explicit form of $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ unknown. Recall that $\lambda^* = \lambda/\{1-p(1-\lambda)\}$. Then $l_M(\lambda, p, F_1, F_2)$ can be rewritten as

$$\begin{aligned} l_M(\lambda, p, F_1, F_2) = & k_q \left[m_0 \log p + m_+ \log(1-p) + n_0 \log\{p(1-\lambda)\} + n_+ \log\{1-p(1-\lambda)\} \right] \\ & + \sum_{h \in I_q} \sum_{i=1}^{m_+} \left\{ m_{i1}(t_h) \log F_1(t_h) + m_{i2}(t_h) \log \bar{F}_1(t_h) \right\} \\ & + \sum_{h \in I_q} \sum_{j=1}^{n_+} \log \Big[(1-\lambda^*) \{F_1(t_h)\}^{n_{j1}(t_h)} \{\bar{F}_1(t_h)\}^{n_{j2}(t_h)} \\ & \qquad + \lambda^* \{F_2(t_h)\}^{n_{j1}(t_h)} \{\bar{F}_2(t_h)\}^{n_{j2}(t_h)} \Big], \end{aligned}$$

from which we have that

$$n_{j1}(t_h) \sim (1-\lambda^*) \mathrm{Bin}\Big(1, F_1(t_h)\Big) + \lambda^* \mathrm{Bin}\Big(1, F_2(t_h)\Big).$$

With the mixture structure above, the EM-algorithm naturally fits into our problem. We first define the missing data. For $j = 1, \ldots, n_+$ and $h \in I_q$, let $V_{jh} = 1$ if $n_{j1}(t_h)$ is from $\mathrm{Bin}\Big(1, F_2(t_h)\Big)$, and 0 if $n_{j1}(t_h)$ is from $\mathrm{Bin}\Big(1, F_1(t_h)\Big)$. Further, let $\mathcal{X} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ be the observed data, $\mathcal{V} = \{V_{jh} | j = 1, \ldots, n_+, h \in I_q\}$, and $\boldsymbol{\theta} = (\lambda, p, F_1, F_2)$. Then, based on $\{\mathcal{X}, \mathcal{V}\}$, we can write down the complete multinomial likelihood and derive the EM-algorithm accordingly. For convenience of presentation, we leave the technical details to Section 2.6.4 and present the E-step and M-step of the EM-iterations directly.

Let $(\lambda^{(0)}, p^{(0)}, F_1^{(0)}, F_2^{(0)})$ and $\lambda^{*(0)}$ denote the initial values of (λ, p, F_1, F_2) and λ^*, respectively. Further, we use $\boldsymbol{\theta}^{(r)} = (\lambda^{(r)}, p^{(r)}, F_1^{(r)}, F_2^{(r)})$ and $\lambda^{*(r)}$ to respectively denote the updated values of $\boldsymbol{\theta} = (\lambda, p, F_1, F_2)$ and λ^* after r EM-iterations, $r = 0, 1, 2, \ldots$.

In the E-step of the rth iteration, for $h = 1, \ldots, k$ and $j = 1, \ldots, n_+$, we calculate

$$E(V_{jh}|\mathcal{X}, \boldsymbol{\theta}^{(r-1)}) = \left\{a_h^{(r)}\right\}^{n_{j1}(t_h)} \left\{b_h^{(r)}\right\}^{n_{j2}(t_h)},$$

where

$$a_h^{(r)} = \frac{\lambda^{*(r-1)} F_2^{(r-1)}(t_h)}{\lambda^{*(r-1)} F_2^{(r-1)}(t_h) + (1 - \lambda^{*(r-1)}) F_1^{(r-1)}(t_h)}$$

and

$$b_h^{(r)} = \frac{\lambda^{*(r-1)} \bar{F}_2^{(r-1)}(t_h)}{\lambda^{*(r-1)} \bar{F}_2^{(r-1)}(t_h) + (1 - \lambda^{*(r-1)}) \bar{F}_1^{(r-1)}(t_h)}.$$

For any $t > 0$, let

$$m_1(t) = \sum_{i=1}^{m} m_{i1}(t), \quad m_2(t) = \sum_{i=1}^{m} m_{i2}(t), \quad n_1(t) = \sum_{j=1}^{n} n_{j1}(t), \text{ and } n_2(t) = \sum_{j=1}^{n} n_{j2}(t).$$

In the M-step, we update (λ, p, F_1, F_2) by

$$\begin{aligned}
\lambda^{(r)} &= \frac{1}{k_q n} \sum_{h \in I_q} \left\{ n_1(t_h) a_h^{(r)} + n_2(t_h) b_h^{(r)} \right\}, \\
p^{(r)} &= \frac{m_0 + n_0}{m + n - n\lambda^{(r)}}, \\
F_1^{(r)} &= \arg \min_{F \text{ is a cdf}} \sum_{h \in I_q} \left[m_+ + n_1(t_h)\{1 - a_h^{(r)}\} + n_2(t_h)\{1 - b_h^{(r)}\} \right] \left[\tilde{F}_1^{(r)}(t_h) - F(t_h) \right]^2, \\
F_2^{(r)} &= \arg \min_{F \text{ is a cdf}} \sum_{h \in I_q} \left\{ n_1(t_h) a_h^{(r)} + n_2(t_h) b_h^{(r)} \right\} \left\{ \tilde{F}_2^{(r)}(t_h) - F(t_h) \right\}^2,
\end{aligned}$$

where

$$\tilde{F}_1^{(r)}(t_h) = \frac{m_1(t_h) + n_1(t_h)\{1 - a_h^{(r)}\}}{m_+ + n_1(t_h)\{1 - a_h^{(r)}\} + n_2(t_h)\{1 - b_h^{(r)}\}} \text{ and } \tilde{F}_2^{(r)}(t_h) = \frac{n_1(t_h) a_h^{(r)}}{n_1(t_h) a_h^{(r)} + n_2(t_h) b_h^{(r)}}.$$

Updating F_1 or F_2 is an isotonic regression problem, which can be solved via the pool adjacent violators algorithm (Ayer et al., 1955). The E-step and M-step are iterated until convergence.

We make two remarks about the above EM-algorithm. Following the proof in Dempster, Laird, and Rubin (1977), we can show that the multinomial likelihood $l_M(\lambda, p, F_1, F_2)$ does not decrease after each iteration. That is, for $r \geq 1$

$$l_M(\lambda^{(r)}, p^{(r)}, F_1^{(r)}, F_2^{(r)}) \geq l_M(\lambda^{(r-1)}, p^{(r-1)}, F_1^{(r-1)}, F_2^{(r-1)}).$$

Further, note that $l_M(\lambda, p, F_1, F_2) \leq 0$ and $l_M(\lambda, p, F_1, F_2)$ is a continuous function of all the unknown parameters. Then the sequence $\left\{l_M(\lambda^{(r)}, p^{(r)}, F_1^{(r)}, F_2^{(r)})\right\}_{r\geq 1}$ eventually converges to a stationary value of $l_M(\lambda, p, F_1, F_2)$ for a given initial value $\boldsymbol{\theta}^{(0)}$ (Wu, 1983). However, this stationary value may not be a global maximum. To increase the possibility of finding the global maximum, we recommend using multiple initial values. Our simulation results demonstrate that this method works well. In practice, we may stop the algorithm when the increment in $l_M(\lambda^{(r)}, p^{(r)}, F_1^{(r)}, F_2^{(r)})$ after an iteration is no greater than, say, 10^{-6}.

2.4 Simulation study

In this section, we perform a simulation study to compare the proposed maximum multinomial likelihood estimator $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ with the following candidate methods:

– The binomial estimator $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$. Here $(\tilde{\lambda}, \tilde{p})$ is defined in (2.3) and the estimator $(\tilde{F}_1, \tilde{F}_2^*)$ is defined in (2.4). Since $\tilde{F}_2^*(x)$ may not be monotonically increasing, we apply isotonic regression to $\tilde{F}_2^*$ to obtain $\tilde{F}_2$:

$$\tilde{F}_2 = \arg \min_{F_2 \text{ is a cdf}} \sum_{h=1}^{k} \{\tilde{F}_2^*(x) - F_2(x)\}^2,$$

which is guaranteed to be an increasing function.

– The maximum empirical likelihood estimator $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{F}_{1,el}, \tilde{F}_{2,el})$, which is discussed in Section 2.2.1.

– The semiparametric estimator $(\hat{\lambda}_D, \hat{p}_D, \hat{F}_{1D}, \hat{F}_{2D})$, proposed by Qin and Leung (2005), in which the log-density ratio of $F_1(x)$ and $F_2(x)$ is a linear function of $\log x$.

– The parametric estimator $(\hat{\lambda}_P, \hat{p}_P, \hat{F}_{1P}, \hat{F}_{2P})$. We assume that both F_1 and F_2 are cdfs of log-normal distributions and use the MLE to estimate the unknown parameters.

We comment that the proposed estimator and the first two candidate estimators are obtained under the same model assumptions on F_1 and F_2, and the semiparametric and parametric estimators are obtained under stronger model assumptions on F_1 and F_2.

We consider two scenarios as follows:

Scenario 1: F_1 and F_2 are the cdfs of $\mathrm{LN}(0,1)$ and $\mathrm{LN}(2,1)$, respectively, where $\mathrm{LN}(a,b)$ denotes a log-normal distribution with mean a and variance b both with respect to the log scale (i.e. mean and variance of the associated normal random variable).

Scenario 2: F_1 and F_2 are the cdfs of $\mathrm{Weibull}(1, 0.5)$ and $\mathrm{Weibull}(0.5, 4)$, respectively, where $\mathrm{Weibull}(a,b)$ denotes a Weibull distribution with shape parameter a and scale parameter b.

For Scenario 1 and Scenario 2, we consider two different sample sizes $(m,n) = (100,100)$ and $(m,n) = (150,250)$ and nine combinations of (λ, p) with λ and p in $\{0.45, 0.55, 0.65\}$ and $\{0.35, 0.45, 0.55\}$, respectively. Here the choices of (λ, p) are spread around $(\hat{\lambda}, \hat{p})$ for the real data in Section 2.5. For each combination of (m, n, λ, p), the number of replications is 1000. The relative bias (RB) and mean squared error (MSE) of the estimates of λ and p are used as the basis for comparing the different estimates of λ and p. Following Lee et al.

(2016), the L_2 distance is used as a basis for comparing the different estimates of F_1 and F_2. Here the L_2 distance between the estimated cdf $\hat{F}$ and true cdf F is defined as

$$L_2(\hat{F}, F) = \int_{-\infty}^{\infty} \{\hat{F}(x) - F(x)\}^2 \mathrm{d}F(x).$$

The simulation results are summarized in Tables 2.1–2.6.

In Scenario 1, the model specification on F_1 and F_2 is satisfied for all five methods. As expected, the semiparametric estimator $(\hat{\lambda}_D, \hat{p}_D, \hat{F}_{1D}, \hat{F}_{2D})$ and parametric estimator $(\hat{\lambda}_P, \hat{p}_P, \hat{F}_{1P}, \hat{F}_{2P})$ perform better than the three nonparametric methods: they have smaller MSEs for estimating (λ, p) and smaller L_2-distance for estimating (F_1, F_2). We also see that the proposed estimator $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ outperforms the binomial estimator $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$ as it gives more accurate estimates, especially when λ is small. The improvement of $(\hat{\lambda}, \hat{F}_2)$ over $(\tilde{\lambda}, \tilde{F}_2)$ is more obvious when λ is small, where the information about the component with cdf F_2 is quite limited. We notice that the maximum empirical likelihood estimator $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{F}_{2,el})$ have quite large RBs, which is in line with the Proposition 2.1.

In Scenario 2, the model specification on F_1 and F_2 is met for neither the semiparametric method nor the parametric method. Consequently, the semiparametric estimator $(\hat{\lambda}_D, \hat{p}_D, \hat{F}_{1D}, \hat{F}_{2D})$ and parametric estimator $(\hat{\lambda}_P, \hat{p}_P, \hat{F}_{1P}, \hat{F}_{2P})$ may be biased. The MSEs of $(\hat{\lambda}_D, \hat{p}_D)$ and $(\hat{\lambda}_P, \hat{p}_P)$ are quite bigger than those of the proposed estimators $(\hat{\lambda}, \hat{p})$. The L_2 distances of $(\hat{F}_{1D}, \hat{F}_{2D})$ and $(\hat{F}_{1P}, \hat{F}_{2P})$ are also bigger than those of $(\hat{F}_1, \hat{F}_2)$. The general trend for comparing the three nonparametric methods is similar to Scenario I: the proposed estimator is more efficient than the binomial estimator, and the maximum empirical likelihood method produces a biased estimator for (λ, p, F_2).

Table 2.1: Relative biases (%) and mean squared errors ($\times 1000$) of estimators of λ and p under Scenario 1 with $(m, n) = (100, 100)$.

λ	0.45	0.45	0.45	0.55	0.55	0.55	0.65	0.65	0.65
p	0.35	0.45	0.55	0.35	0.45	0.55	0.35	0.45	0.55
RB of $\hat{\lambda}$	4.24	2.59	2.53	2.78	2.85	0.98	1.68	1.84	1.06
MSE of $\hat{\lambda}$	10.91	8.23	7.03	9.64	7.98	6.13	7.83	6.34	5.30
RB of $\tilde{\lambda}$	-4.10	-2.80	-1.09	-2.53	-0.26	-1.17	-1.13	0.28	-0.02
MSE of $\tilde{\lambda}$	20.79	13.32	9.85	16.97	11.32	8.09	11.83	7.83	6.36
RB of $\tilde{\lambda}_{el}$	79.08	66.86	55.03	52.99	45.19	36.57	34.87	29.92	24.38
MSE of $\tilde{\lambda}_{el}$	128.17	92.31	63.49	86.24	63.46	42.29	52.40	39.07	26.58
RB of $\hat{\lambda}_D$	-1.04	-1.38	0.23	0.24	0.76	-0.17	-0.36	0.71	0.03
MSE of $\hat{\lambda}_D$	9.19	7.30	6.40	7.64	6.68	5.47	6.34	5.50	4.98
RB of $\hat{\lambda}_P$	1.38	0.20	1.01	1.04	1.80	0.46	0.55	1.05	0.46
MSE of $\hat{\lambda}_P$	9.32	7.37	6.17	7.88	6.48	5.31	6.41	5.30	4.68
RB of $\hat{p}$	1.62	0.99	0.77	1.46	1.49	0.72	1.81	1.14	0.89
MSE of $\hat{p}$	2.03	2.19	1.92	2.00	2.29	2.34	2.02	2.29	2.46
RB of $\tilde{p}$	-0.45	-0.37	-0.19	-0.23	0.44	-0.01	0.66	0.46	0.43
MSE of $\tilde{p}$	2.43	2.52	2.15	2.31	2.49	2.57	2.22	2.41	2.62
RB of $\tilde{p}_{el}$	29.66	23.96	18.78	25.08	20.75	16.22	20.86	16.75	13.45
MSE of $\tilde{p}_{el}$	12.77	13.51	12.08	9.67	10.67	9.77	7.30	7.71	7.42
RB of $\hat{p}_D$	-0.01	-0.20	0.07	0.38	0.63	0.24	0.71	0.57	0.37
MSE of $\hat{p}_D$	1.85	2.06	1.82	1.80	2.13	2.16	1.81	2.25	2.37
RB of $\hat{p}_P$	0.70	0.26	0.30	0.69	1.02	0.47	1.16	0.72	0.56
MSE of $\hat{p}_P$	1.88	2.09	1.88	1.80	2.15	2.14	1.85	2.24	2.33

Table 2.2: Relative biases (%) and mean squared errors ($\times 1000$) of estimators of λ and p under Scenario 1 with $(m, n) = (150, 250)$.

λ	0.45	0.45	0.45	0.55	0.55	0.55	0.65	0.65	0.65
p	0.35	0.45	0.55	0.35	0.45	0.55	0.35	0.45	0.55
RB of $\hat{\lambda}$	4.20	1.70	1.82	1.72	1.21	0.62	0.55	-0.08	0.23
MSE of $\hat{\lambda}$	5.93	4.18	3.51	5.00	3.63	2.87	3.76	3.24	2.45
RB of $\tilde{\lambda}$	-0.02	-1.27	-0.17	-0.84	-0.51	-0.39	-0.92	-1.12	-0.29
MSE of $\tilde{\lambda}$	8.96	6.12	4.59	7.56	5.04	3.49	5.10	4.19	2.79
RB of $\tilde{\lambda}_{el}$	79.89	67.11	55.08	53.16	45.00	36.87	34.84	29.44	24.19
MSE of $\tilde{\lambda}_{el}$	129.83	91.90	62.30	86.05	61.88	41.82	51.68	37.17	25.33
RB of $\hat{\lambda}_D$	0.70	0.25	0.21	-0.19	0.29	-0.09	-0.16	-0.03	0.04
MSE of $\hat{\lambda}_D$	3.91	3.14	2.79	3.42	2.70	2.30	2.82	2.41	1.92
RB of $\hat{\lambda}_P$	1.75	0.67	0.72	0.31	0.38	0.06	0.17	0.03	0.09
MSE of $\hat{\lambda}_P$	4.03	2.92	2.68	3.26	2.52	2.24	2.46	2.22	1.81
RB of $\hat{p}$	1.66	1.01	1.00	1.48	1.03	0.45	1.35	0.24	0.50
MSE of $\hat{p}$	1.29	1.44	1.46	1.37	1.49	1.60	1.42	1.62	1.56
RB of $\tilde{p}$	0.22	-0.02	0.29	0.33	0.24	-0.03	0.49	-0.38	0.18
MSE of $\tilde{p}$	1.52	1.64	1.64	1.54	1.68	1.74	1.56	1.78	1.66
RB of $\tilde{p}_{el}$	44.79	35.64	27.54	38.75	30.94	23.79	32.00	25.17	19.98
MSE of $\tilde{p}_{el}$	25.68	26.71	23.76	19.63	20.48	18.12	13.83	14.10	13.12
RB of $\hat{p}_D$	0.11	0.37	0.32	0.35	0.46	0.02	0.72	0.17	0.32
MSE of $\hat{p}_D$	1.03	1.27	1.30	1.23	1.30	1.41	1.19	1.38	1.39
RB of $\hat{p}_P$	0.53	0.52	0.50	0.59	0.49	0.09	0.91	0.19	0.33
MSE of $\hat{p}_P$	1.04	1.22	1.25	1.20	1.26	1.37	1.15	1.37	1.34

Table 2.3: L_2 distance ($\times 100$) between the estimated cumulative distribution function and true cumulative distribution function under Scenario 1.

λ	0.45	0.45	0.45	0.55	0.55	0.55	0.65	0.65	0.65
p	0.35	0.45	0.55	0.35	0.45	0.55	0.35	0.45	0.55
$m = 100$, $n = 100$									
$\hat{F}_1$	0.23	0.28	0.36	0.24	0.29	0.36	0.25	0.28	0.35
$\tilde{F}_1$	0.25	0.31	0.39	0.26	0.31	0.39	0.26	0.30	0.37
$\tilde{F}_{1,el}$	0.25	0.31	0.39	0.26	0.31	0.39	0.26	0.30	0.37
$\hat{F}_{1D}$	0.20	0.24	0.31	0.21	0.25	0.31	0.21	0.24	0.30
$\hat{F}_{1P}$	0.14	0.18	0.23	0.15	0.18	0.23	0.15	0.17	0.22
$\hat{F}_2$	1.21	1.08	0.94	0.81	0.75	0.67	0.59	0.48	0.45
$\tilde{F}_2$	2.78	1.86	1.36	1.55	1.09	0.85	0.87	0.57	0.51
$\tilde{F}_{2,el}$	4.60	3.93	3.06	2.81	2.35	1.87	1.70	1.38	1.10
$\hat{F}_{2D}$	1.04	0.94	0.82	0.67	0.63	0.57	0.51	0.41	0.41
$\hat{F}_{2P}$	0.90	0.83	0.71	0.59	0.55	0.47	0.44	0.36	0.33
$m = 150$, $n = 250$									
$\hat{F}_1$	0.15	0.17	0.23	0.16	0.19	0.23	0.16	0.19	0.24
$\tilde{F}_1$	0.17	0.19	0.25	0.17	0.20	0.25	0.18	0.21	0.25
$\tilde{F}_{1,el}$	0.17	0.19	0.25	0.17	0.20	0.25	0.18	0.21	0.25
$\hat{F}_{1D}$	0.12	0.13	0.18	0.13	0.15	0.18	0.13	0.15	0.19
$\hat{F}_{1P}$	0.09	0.10	0.13	0.09	0.11	0.13	0.09	0.11	0.15
$\hat{F}_2$	0.60	0.52	0.46	0.42	0.34	0.29	0.27	0.24	0.20
$\tilde{F}_2$	1.07	0.79	0.61	0.68	0.48	0.35	0.35	0.30	0.22
$\tilde{F}_{2,el}$	4.44	3.68	2.87	2.76	2.23	1.68	1.57	1.25	0.95
$\hat{F}_{2D}$	0.44	0.37	0.34	0.29	0.24	0.23	0.22	0.18	0.16
$\hat{F}_{2P}$	0.36	0.31	0.27	0.24	0.21	0.19	0.16	0.14	0.13

Table 2.4: Relative biases (%) and mean squared errors ($\times 1000$) of estimators of λ and p under Scenario 2 with $(m, n) = (100, 100)$.

λ	0.45	0.45	0.45	0.55	0.55	0.55	0.65	0.65	0.65
p	0.35	0.45	0.55	0.35	0.45	0.55	0.35	0.45	0.55
RB of $\hat{\lambda}$	0.83	-0.17	-0.39	-0.88	-0.44	-0.21	-0.53	-0.44	-0.62
MSE of $\hat{\lambda}$	15.29	11.17	8.72	12.78	10.06	7.94	11.30	7.66	6.77
RB of $\tilde{\lambda}$	-2.08	-1.69	-1.16	-1.79	-0.94	-0.47	-1.01	-0.57	-0.70
MSE of $\tilde{\lambda}$	20.20	13.39	9.64	14.68	10.95	8.32	12.85	7.89	6.91
RB of $\tilde{\lambda}_{el}$	79.59	67.20	54.88	53.12	45.01	36.86	35.03	29.60	24.07
MSE of $\tilde{\lambda}_{el}$	129.94	93.33	63.10	86.61	62.90	43.06	52.94	38.32	26.09
RB of $\hat{\lambda}_D$	-37.58	-36.26	-32.11	-38.93	-36.45	-31.68	-38.09	-35.82	-30.37
MSE of $\hat{\lambda}_D$	34.51	34.11	29.85	53.33	49.84	41.21	69.15	63.86	52.10
RB of $\hat{\lambda}_P$	9.42	6.11	4.53	6.41	5.69	4.20	5.15	4.22	2.61
MSE of $\hat{\lambda}_P$	18.69	15.39	11.35	17.05	11.86	9.86	15.13	9.77	8.28
RB of $\hat{p}$	0.49	0.25	0.22	0.15	0.25	0.26	0.26	0.19	0.22
MSE of $\hat{p}$	2.18	2.42	2.37	2.20	2.44	2.49	2.34	2.44	2.46
RB of $\tilde{p}$	-0.22	-0.12	0.03	-0.12	0.10	0.18	0.10	0.14	0.19
MSE of $\tilde{p}$	2.36	2.54	2.45	2.27	2.50	2.53	2.40	2.46	2.47
RB of $\tilde{p}_{el}$	29.53	24.04	18.97	25.06	20.55	16.28	20.18	16.65	13.36
MSE of $\tilde{p}_{el}$	12.69	13.61	12.50	9.72	10.53	9.82	7.19	7.71	7.25
RB of $\hat{p}_D$	-9.89	-9.36	-8.19	-12.61	-11.77	-10.27	-15.17	-14.26	-11.99
MSE of $\hat{p}_D$	2.80	3.72	4.32	3.60	4.94	5.94	4.55	6.36	7.49
RB of $\hat{p}_P$	3.11	2.22	1.74	3.02	2.67	2.01	3.18	2.53	1.86
MSE of $\hat{p}_P$	2.40	2.72	2.65	2.32	2.67	2.71	2.64	2.64	2.77

Table 2.5: Relative biases (%) and mean squared errors ($\times 1000$) of estimators of λ and p under Scenario 2 with $(m, n) = (150, 250)$.

λ	0.45	0.45	0.45	0.55	0.55	0.55	0.65	0.65	0.65
p	0.35	0.45	0.55	0.35	0.45	0.55	0.35	0.45	0.55
RB of $\hat{\lambda}$	0.65	-0.890	-1.61	-0.65	-0.83	-1.00	-0.21	-0.67	-0.01
MSE of $\hat{\lambda}$	7.63	6.34	4.42	6.28	5.06	3.67	5.41	3.66	2.76
RB of $\tilde{\lambda}$	-0.26	-1.47	-1.83	-0.89	-0.96	-1.05	-0.27	-0.70	-0.02
MSE of $\tilde{\lambda}$	8.87	7.13	4.63	6.64	5.22	3.72	5.53	3.68	2.77
RB of $\tilde{\lambda}_{el}$	79.70	67.30	54.45	53.07	45.08	36.60	35.12	29.51	24.44
MSE of $\tilde{\lambda}_{el}$	129.24	92.50	60.84	85.69	62.14	41.27	52.58	37.30	25.86
RB of $\hat{\lambda}_D$	-39.31	-37.23	-34.35	-39.68	-37.14	-33.58	-39.48	-36.70	-30.37
MSE of $\hat{\lambda}_D$	34.58	32.19	28.53	50.73	46.02	39.97	69.23	62.21	47.64
B of $\hat{\lambda}_P$	14.91	10.59	7.87	11.17	8.66	5.83	8.83	5.90	4.60
MSE of $\hat{\lambda}_P$	10.58	8.82	5.29	9.60	6.44	5.02	7.40	5.27	3.96
RB of $\hat{p}$	0.67	-0.26	-0.01	0.55	-0.47	-0.21	0.12	0.06	-0.21
MSE of $\hat{p}$	1.37	1.55	1.68	1.40	1.74	1.69	1.46	1.82	1.61
RB of $\tilde{p}$	0.39	-0.44	-0.08	0.46	-0.52	-0.24	0.09	0.04	-0.21
MSE of $\tilde{p}$	1.44	1.63	1.71	1.42	1.76	1.70	1.47	1.82	1.61
RB of $\tilde{p}_{el}$	45.07	35.26	27.46	38.96	30.30	23.74	31.26	25.45	19.54
MSE of $\tilde{p}_{el}$	25.95	26.12	23.66	19.75	19.77	18.03	13.23	14.43	12.60
RB of $\hat{p}_D$	-13.26	-12.79	-11.29	-16.70	-16.40	-14.56	-21.24	-19.58	-16.95
MSE of $\hat{p}_D$	2.92	4.34	5.38	4.22	6.65	8.25	6.37	9.27	11.21
RB of $\hat{p}_P$	6.49	4.40	3.77	7.11	4.62	3.48	6.56	4.72	3.05
MSE of $\hat{p}_P$	1.76	1.95	1.98	2.03	2.02	2.07	1.91	2.25	1.90

Table 2.6: L_2 distance ($\times 100$) between the estimated cumulative distribution function and true cumulative distribution function under Scenario 2.

λ	0.45	0.45	0.45	0.55	0.55	0.55	0.65	0.65	0.65
p	0.35	0.45	0.55	0.35	0.45	0.55	0.35	0.45	0.55
$m = 100$, $n = 100$									
$\hat{F}_1$	0.25	0.29	0.35	0.24	0.30	0.37	0.26	0.31	0.37
$\tilde{F}_1$	0.27	0.31	0.37	0.25	0.31	0.38	0.26	0.32	0.38
$\tilde{F}_{1,el}$	0.27	0.31	0.37	0.25	0.31	0.38	0.26	0.32	0.38
$\hat{F}_{1D}$	0.24	0.28	0.39	0.25	0.32	0.46	0.30	0.41	0.55
$\hat{F}_{1P}$	0.32	0.34	0.43	0.33	0.37	0.44	0.37	0.42	0.46
$\hat{F}_2$	1.39	1.12	0.89	0.84	0.65	0.58	0.57	0.45	0.43
$\tilde{F}_2$	2.20	1.52	1.03	1.05	0.75	0.62	0.69	0.46	0.44
$\tilde{F}_{2,el}$	1.88	1.64	1.34	1.28	1.08	0.83	0.77	0.65	0.54
$\hat{F}_{2D}$	6.07	5.78	5.04	5.92	5.45	4.76	5.50	5.11	4.24
$\hat{F}_{2P}$	1.85	1.79	1.45	1.44	1.03	0.92	1.01	0.77	0.73
$m = 150$, $n = 250$									
$\hat{F}_1$	0.16	0.19	0.25	0.17	0.18	0.24	0.17	0.20	0.26
$\tilde{F}_1$	0.17	0.20	0.26	0.17	0.19	0.25	0.18	0.21	0.26
$\tilde{F}_{1,el}$	0.17	0.20	0.26	0.17	0.19	0.25	0.18	0.21	0.26
$\hat{F}_{1D}$	0.18	0.24	0.37	0.22	0.29	0.47	0.28	0.40	0.68
$\hat{F}_{1P}$	0.22	0.27	0.32	0.26	0.29	0.35	0.26	0.32	0.42
$\hat{F}_2$	0.63	0.57	0.41	0.36	0.33	0.24	0.23	0.18	0.16
$\tilde{F}_2$	0.84	0.69	0.44	0.39	0.35	0.25	0.24	0.19	0.16
$\tilde{F}_{2,el}$	1.81	1.51	1.19	1.11	0.91	0.72	0.67	0.53	0.42
$\hat{F}_{2D}$	5.72	5.34	4.72	5.59	5.13	4.35	5.28	4.77	3.69
$\hat{F}_{2P}$	0.96	1.00	0.71	0.75	0.60	0.57	0.52	0.49	0.45

2.5 Application to malaria data

In this section, we analyze the dataset in Example 1.1. Similarly to Qin and Leung (2005), we consider a subset of this dataset for children aged between six and nine months collected in January–June, the wet season, and in July–December, the dry season. The measurements are the parasite levels per μl, ranging from 0 to 399952.1. Among these measurements, there are $n = 264$ observations from the mixture sample (i.e. from the wet season) and $m = 144$ observations for nonmalaria individuals from the dry season.

We apply all the five methods discussed in Section 2.4 to the data above. For the parametric method, we use a two-component log-normal mixture to model F_1 since the histogram of the logarithm of the positive parasite levels in the nonmalaria population shows two modes. The estimates of (λ, p) are provided in Table 2.7 and the estimates of (F_1, F_2) are plotted in Figure 2.1. To evaluate the variability of the estimators of λ and p, we use the nonparametric bootstrap method (Efron, 1981) with 1000 bootstrap samples to calculate 95% bootstrap percentile confidence intervals for λ and p. The confidence intervals are provided in Table 2.7. The maximum empirical likelihood estimate $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{F}_{2,el})$ are quite different from other four estimates, and their confidence intervals do not overlap with confidence intervals of other estimates. This agrees with our observations in Section 2.4, and it is because $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{F}_{2,el})$ is biased, and hence unreliable. We notice that the two nonparametric estimators, $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ and $(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$, give quite close results. The proposed estimate $(\hat{\lambda}, \hat{p})$ of (λ, p) is bigger than the semiparametric estimate $(\hat{\lambda}_D, \hat{p}_D)$ and the parametric estimate $(\hat{\lambda}_P, \hat{p}_P)$. We also see that $\hat{F}_1$ is slightly different from $\hat{F}_{1D}$ and $\hat{F}_{1P}$ when the log positive parasite levels are between 6 to 9, and $\hat{F}_2$ is different from $\hat{F}_{2D}$ and $\hat{F}_{2P}$ when the log positive parasite levels are between 8 to 10. Since the proposed estimator makes less assumptions on F_1 and F_2 than the semiparametric and parametric

estimators and further, it is more efficient than the binomial estimator as shown in Section 2.4, the inference based on $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ may be more reliable for the malaria data.

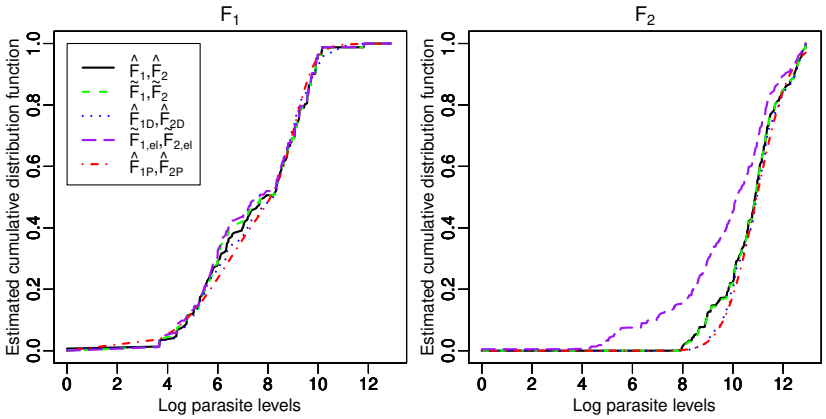

Figure 2.1: Estimates of F_1 and F_2 for the malaria data.

The claim above can be further verified by a simulation experiment according to Models (2.1) and (2.2) with $(m, n) = (144, 264)$ and $(\lambda, p, F_1, F_2) = (\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$. The simulation results based on 1000 repetitions are summarized in Table 2.8. The maximum empirical likelihood estimator of (λ, p) has quite large RB and the largest MSE among five estimators, and also has the largest L_2 distance for estimating F_2 among the five estimators. We further notice that $(\hat{\lambda}, \hat{p})$ has smaller MSEs than $(\tilde{\lambda}_D, \tilde{p}_D)$ and $(\hat{\lambda}_D, \hat{p}_D)$, and the proposed estimator of (F_1, F_2) has the smallest L_2 distance among all five methods. The proposed estimator $(\hat{\lambda}, \hat{p})$ of (λ, p) has the comparable performance as $(\hat{\lambda}_P, \hat{p}_P)$: $(\hat{\lambda}, \hat{p})$ has a smaller RB while $(\hat{\lambda}_P, \hat{p}_P)$ has a relatively smaller MSE. It is worth mentioning that the RBs of $(\hat{\lambda}_P, \hat{p}_P)$ are negative. This explains why $(\hat{\lambda}, \hat{p})$ are bigger than $(\hat{\lambda}_P, \hat{p}_P)$ in the malaria data.

For an individual in the mixture sample, set $D = 1$ if the corresponding subject is infected with malaria and set $D = 0$ otherwise. Let $g_1(x)$ and $g_2(x)$ be the pdfs of the logarithm of the positive parasite levels in the nonmalaria and malaria populations, respectively. Then given the log parasite level x, the posterior probability of $D = 1$ is given

Table 2.7: Point estimates along with 95% bootstrap percentile confidence intervals of λ and p for the malaria data.

	λ		p	
	estimate	95% confidence interval	estimate	95% confidence interval
$(\hat{\lambda},\hat{p})$	0.545	$[0.424, 0.670]$	0.439	$[0.364, 0.520]$
$(\tilde{\lambda},\tilde{p})$	0.541	$[0.382, 0.669]$	0.438	$[0.347, 0.514]$
$(\tilde{\lambda}_{el},\tilde{p}_{el})$	0.791	$[0.745, 0.844]$	0.582	$[0.520, 0.650]$
$(\hat{\lambda}_D,\hat{p}_D)$	0.507	$[0.398, 0.633]$	0.423	$[0.348, 0.503]$
$(\hat{\lambda}_P,\hat{p}_P)$	0.507	$[0.417, 0.642]$	0.423	$[0.357, 0.508]$

Table 2.8: Confirmative simulation for the malaria data.

	RB (%)		MSE ($\times$1000)		L_2 distance ($\times$100)	
	λ	p	λ	p	F_1	F_2
$(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$	1.31	1.06	3.99	1.53	0.20	0.36
$(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2)$	-0.47	0.21	5.38	1.71	0.21	0.49
$(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{F}_{1,el}, \tilde{F}_{2,el})$	35.42	23.94	38.22	12.21	0.23	1.63
$(\hat{\lambda}_D, \hat{p}_D, \hat{F}_{1D}, \hat{F}_{2D})$	-7.73	-3.76	5.44	1.58	0.21	0.65
$(\hat{\lambda}_P, \hat{p}_P, \hat{F}_{1P}, \hat{F}_{2P})$	-5.12	-2.47	3.91	1.40	0.23	0.51

by

$$\eta(x) = \frac{\lambda^* g_2(x)}{(1-\lambda^*)g_1(x) + \lambda^* g_2(x)}.$$

We can use $\hat{F}_1$ and $\hat{F}_2$ to construct an estimator for $\eta(x)$.

We first estimate $g_1(x)$ and $g_2(x)$ as

$$\hat{g}_1(x) = \int K_{h_1}(\log t - x) d\hat{F}_1(t), \quad \hat{g}_2(x) = \int K_{h_2}(\log t - x) d\hat{F}_2(t),$$

where $K_h(x) = K(x/h)/h$, $K(x)$ is a symmetric kernel, and h is the bandwidth. For illustration, we use the standard normal density function for $K(x)$ and choose the bandwidth h by rule of thumb (Silverman, 1986): $h_1 = 1.06\hat{\sigma}_1(m_+ + n_+)^{-1/5}$ and $h_2 = 1.06\hat{\sigma}_2 n_+^{-1/5}$, where

$$\hat{\sigma}_1^2 = \int_x (\log x)^2 d\hat{F}_1(x) - \left\{\int_x \log x d\hat{F}_1(x)\right\}^2, \quad \hat{\sigma}_2^2 = \int_x (\log x)^2 d\hat{F}_2(x) - \left\{\int_x \log x d\hat{F}_2(x)\right\}^2.$$

Since all the positive observations in $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ are used to estimate F_1 and only the positive observations in $\{Y_1, \ldots, Y_n\}$ are used to estimate F_2, $m_+ + n_+$ is used in h_1 and n_+ is used in h_2.

Figure 2.2 shows the density estimate of the nonmalaria population (i.e. observations from the dry season), the density estimate of the mixture sample (i.e. observations from the wet season), and the estimated posterior probability of having malaria given the logarithm of the positive parasite level in the mixture sample. We can see that the density estimates match the histograms well for the data from both the dry season and the wet season. This indicates that the density estimators $\hat{g}_1$ and $\hat{g}_2$ work well for the malaria data. Further, we can see that the estimated posterior probability of having malaria given the logarithm of the positive parasite level is an increasing function of the log parasite level. This agrees with clinical malaria knowledge (Vounatsou, Smith, and Smith, 1998): the higher the parasite level, the higher the probability that an individual is infected with malaria.

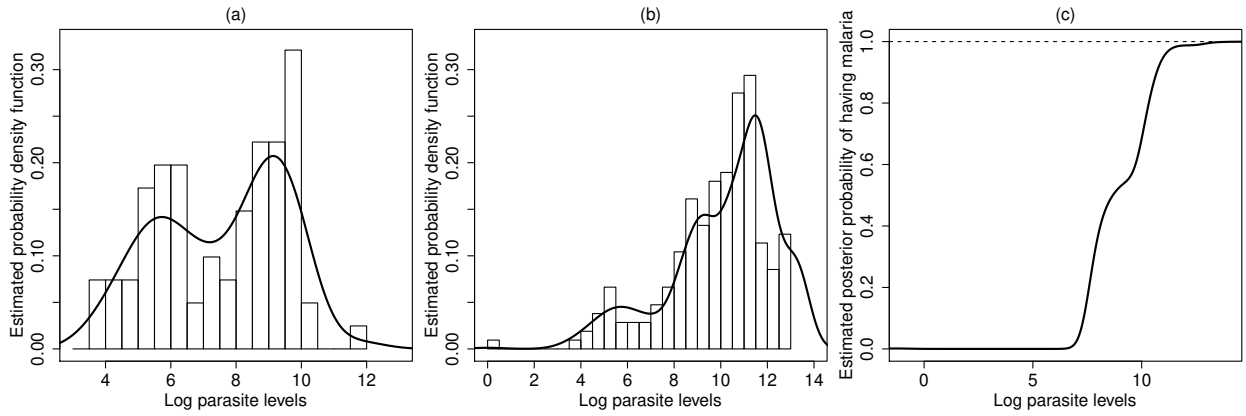

Figure 2.2: Density and posterior probability estimation of the malaria data: Panel (a) plots the histogram of the logarithm of the positive parasite levels in the nonmalaria population and the density estimate $\hat{g}_1$; panel (b) plots the histogram of the logarithm of the positive parasite levels in the mixture sample and the density estimate $(1-\hat{\lambda}^*)\hat{g}_1+\hat{\lambda}^*\hat{g}_2$; and panel (c) plots the estimated posterior probability of having malaria given the logarithm of the positive parasite level in the mixture sample.

2.6 Technical details

2.6.1 Identifiability of (λ, p, F_1, F_2)

In this section, we show that (λ, p, F_1, F_2) is identifiable if $0 < \lambda, p < 1$ and $F_1(x) = F_2(x) = 0$ for any $x \leq 0$. Recall that

$$F_X(x) = pI(x \geq 0) + (1-p)F_1(x),\ F_Y(y) = (1-\lambda)\left\{pI(y \geq 0) + (1-p)F_1(y)\right\} + \lambda F_2(y).$$

Assume that two parameter combinations (λ, p, F_1, F_2) and $(\lambda^*, p^*, F_1^*, F_2^*)$ give the same model (F_X, F_Y). That is, for all x,

$$pI(x \geq 0) + (1-p)F_1(x) = p^*I(x \geq 0) + (1-p^*)F_1^*(x), \tag{2.9}$$

and for all y,

$$\begin{aligned}&(1-\lambda)\left\{pI(y \geq 0) + (1-p)F_1(y)\right\} + \lambda F_2(y)\\ &= (1-\lambda^*)\left\{p^*I(y \geq 0) + (1-p^*)F_1^*(y)\right\} + \lambda^* F_2^*(y).\end{aligned} \tag{2.10}$$

We need to argue that $(\lambda, p, F_1, F_2) = (\lambda^*, p^*, F_1^*, F_2^*)$.

Substituting $x = 0$ into (2.9) and using the fact that $F_1(0) = F_1^*(0) = 0$, we have

$$F_X(0) = p = p^*.$$

This, together with (2.9) and the fact that $0 < p = p^* < 1$, implies that $F_1(x) = F_1^*(x)$. Hence

$$(p, F_1) = (p^*, F_1^*). \tag{2.11}$$

Similarly, substituting $y = 0$ into (2.10), and using the fact that $F_2(0) = F_2^*(0) = 0$ and the result that $(p, F_1) = (p^*, F_1^*)$, we get $\lambda = \lambda^*$. With (2.10), we further have

$$\lambda F_2(y) = \lambda^* F_2^*(y),$$

which, together with the result that $0 < \lambda = \lambda^* < 1$, leads to $F_2 = F_2^*$. Therefore,

$$(\lambda, F_2) = (\lambda^*, F_2^*). \tag{2.12}$$

Combining (2.11) and (2.12), we have $(\lambda, p, F_1, F_2) = (\lambda^*, p^*, F_1^*, F_2^*)$. That is, (λ, p, F_1, F_2) is identifiable.

2.6.2 Proof of Proposition 2.1

For a given vector $\mathbf{a}$ such that $a_i \geq 0$ and $\sum_{i=1}^{m_+ + n_+} a_i = 1$, let $a_+ = \sum_{i=1}^{m_+} a_i$,

$$a_i^* = \frac{a_+}{m_+}, \quad i = 1, \ldots, m_+, \quad a_{j+m_+}^* = \frac{1 - a_+}{n_+}, \quad j = 1, \ldots, n_+,$$

and $\mathbf{a}^* = (a_1^*, \ldots, a_{m_+ + n_+}^*)^{\mathrm{T}}$. Further, let $\mathbf{b}^* = (1/n_+, \ldots, 1/n_+)^{\mathrm{T}}$.

Using Jensen's inequality, we can show that

$$\tilde{l}_{el}(\lambda, p, \mathbf{a}, \mathbf{b}) \leq \tilde{l}_{el}(\lambda, p, \mathbf{a}^*, \mathbf{b}^*) \tag{2.13}$$

with equality holding if and only if $\mathbf{a} = \mathbf{a}^*$ and $\mathbf{b} = \mathbf{b}^*$. Hence, $\tilde{b}_{j,el} = \mathbf{b}^*$. The inequality in (2.13) also implies that we can concentrate on the case where $\mathbf{b} = \mathbf{b}^*$ and

$$a_1 = \ldots = a_{m_+} = a, \quad a_{m_+ + 1} = \ldots = a_{m_+ + n_+} = a^*,$$

where $a \geq 0$, $a^* \geq 0$, and $m_+ a + n_+ a^* = 1$. In this case, $\tilde{l}_{el}(\lambda, p, \mathbf{a}, \mathbf{b})$ becomes

$$\begin{aligned} \tilde{l}_{el}^*(\lambda, p, a, a^*) = {} & m_0 \log p + m_+ \log(1 - p) + m_+ \log a \\ & + n_0 \log(1 - \lambda) + n_0 \log p + n_+ \log \{(1 - \lambda)(1 - p)a^* + \lambda / n_+\} \end{aligned}$$

and the parameter space for (λ, p, a, a^*) is

$$\Theta^* = \{(\lambda, p, a, a^*) | \lambda \in [0, 1], \ p \in [0, 1], \ a \in [0, 1/m_+], \ a^* \in [0, 1/n_+], \ \ m_+ a + n_+ a^* = 1\}.$$

Let

$$(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) = \arg \max_{(\lambda,p,a,a^*)\in\Theta} \tilde{l}^*_{el}(\lambda, p, a, a^*).$$

Define

$$\Theta^*_1 = \{(\lambda, p, a, a^*) \in \Theta^* | \lambda \in (0,1),\ p \in (0,1),\ a \in (0, 1/m_+)\},$$
$$\Theta^*_2 = \{(\lambda, p, a, a^*) \in \Theta^* | a = 1/m_+\},$$
$$\Theta^*_3 = \{(\lambda, p, a, a^*) \in \Theta^* | \lambda = 0\}.$$

Clearly, $\tilde{\lambda}_{el} \neq 1$ and $\tilde{a}_{el} \neq 0$, $\tilde{p}_{el} \neq 0$ or 1. Thus, $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el})$ belongs to Θ^*_1, Θ^*_2, or Θ^*_3. To complete the proof, we need to argue that $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) \in \Theta^*_2$. We first argue that $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) \notin \Theta^*_1$.

Define

$$\begin{aligned}\tilde{L}^*_{el}(\lambda, p, a, a^*) = {} & m_0 \log p + m_+ \log(1-p) + m_+ \log a + n_0 \log(1-\lambda) + n_0 \log p \\ & + n_+ \log\{(1-\lambda)(1-p)a^* + \lambda/n_+\} + s(m_+ a + n_+ a^* - 1),\end{aligned}$$

where s is the Lagrange multiplier. If $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) \in \Theta^*_1$, then $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el})$ will satisfy

$$\frac{\partial \tilde{L}^*_{el}(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el})}{\partial \lambda} = -\frac{n_0}{1-\tilde{\lambda}_{el}} + \frac{1 - n_+(1-\tilde{p}_{el})\tilde{a}^*_{el}}{(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el})\tilde{a}^*_{el} + \tilde{\lambda}_{el}/n_+} = 0, \tag{2.14}$$

$$\frac{\partial \tilde{L}^*_{el}(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el})}{\partial p} = \frac{m_0 + n_0}{\tilde{p}_{el}} - \frac{m_+}{1-\tilde{p}_{el}} - \frac{n_+(1-\tilde{\lambda}_{el})\tilde{a}^*_{el}}{(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el})\tilde{a}^*_{el} + \tilde{\lambda}_{el}/n_+} = 0, \tag{2.15}$$

$$\frac{\partial \tilde{L}^*_{el}(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el})}{\partial a} = \frac{m_+}{\tilde{a}_{el}} + s m_+ = 0, \tag{2.16}$$

$$\frac{\partial \tilde{L}^*_{el}(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el})}{\partial a^*} = \frac{n_+(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el})}{(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el})\tilde{a}^*_{el} + \tilde{\lambda}_{el}/n_+} + s n_+ = 0. \tag{2.17}$$

Further,

$$m_+ \tilde{a}_{el} + n_+ \tilde{a}^*_{el} - 1 = 0. \tag{2.18}$$

Combining (2.16), (2.17), and (2.18) leads to

$$\tilde{a}^*_{el} = \frac{n_+(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el}) - m_+\tilde{\lambda}_{el}}{n_+(m_+ + n_+)(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el})}. \tag{2.19}$$

By (2.14), we have

$$(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el})\tilde{a}^*_{el} + \tilde{\lambda}_{el}/n_+ = 1/n, \tag{2.20}$$

which, together with (2.19), results in

$$(1-\tilde{\lambda}_{el})(1-\tilde{p}_{el}) + \tilde{\lambda}_{el} = \frac{m_+ + n_+}{n}. \tag{2.21}$$

Substituting (2.20) into (2.15), we get

$$(m+n)\tilde{p}_{el} = m_0 + n_0 + n\tilde{\lambda}_{el}. \tag{2.22}$$

Combining (2.21) and (2.22), we have $\tilde{p}_{el} = 1$, which is a contradiction. Hence, $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) \notin \Theta^*_1$ and

$$(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) \in \Theta^*_2 \cup \Theta^*_3.$$

We can show that

$$\max_{(\lambda,p,a,a^*)\in\Theta^*_2} \tilde{l}^*_{el}(\lambda,p,a,a^*) = (m_0+n_0)\log(m_0+n_0) + n_0\log n_0 - n\log n - (m+n_0)\log(m+n_0),$$

$$\max_{(\lambda,p,a,a^*)\in\Theta^*_3} \tilde{l}^*_{el}(\lambda,p,a,a^*) = (m_0+n_0)\log(m_0+n_0) - (m+n)\log(m+n).$$

It can be checked that

$$n\log n + (m+n_0)\log(m+n_0) < n_0\log n_0 + (m+n)\log(m+n),$$

which implies that

$$\max_{(\lambda,p,a,a^*)\in\Theta^*_2} \tilde{l}^*_{el}(\lambda,p,a,a^*) > \max_{(\lambda,p,a,a^*)\in\Theta^*_3} \tilde{l}^*_{el}(\lambda,p,a,a^*).$$

Therefore, $(\tilde{\lambda}_{el}, \tilde{p}_{el}, \tilde{a}_{el}, \tilde{a}^*_{el}) \in \Theta^*_2$.

Note that

$$\arg \max_{(\lambda,p,a,a^*)\in\Theta^*_2} \tilde{l}^*_{el}(\lambda, p, a, a^*) = \left(\frac{n_+}{n}, \frac{m_0 + n_0}{m_0 + n_0 + m_+}, \frac{1}{m_+}, 0 \right).$$

Then

$$\begin{aligned}
\tilde{\lambda}_{el} &= \frac{n_+}{n}, \\
\tilde{p}_{el} &= \frac{m_0 + n_0}{m_0 + n_0 + m_+}, \\
\tilde{a}_{el,i} &= \tilde{a}_{el} = \frac{1}{m_+}, \ i = 1, \cdots, m_+, \\
\tilde{a}_{el,i} &= \tilde{a}^*_{el} = 0, \ i = m_+ + 1, \cdots, m_+ + n_+, \\
\tilde{b}_{el,j} &= \frac{1}{n_+}, \ j = 1, \cdots, n_+.
\end{aligned}$$

This completes the proof of Proposition 2.1.

2.6.3 Proof of Theorem 2.1

Technical Preparation

Since both F_1 and F_2 are absolutely continuous, without loss of generality, we assume throughout the proofs that there are no ties in the positive values of $\{X_i\}_{i=1}^m$ and $\{Y_j\}_{j=1}^n$. Further, define

$$H_k(x) = \frac{1}{k} \sum_{h=1}^{k} I(t_h \le x).$$

Denote $t_L = \min\{t_h : \ h \in I_q\}$ and $t_U = \max\{t_h : \ h \in I_q\}$. The following lemma summarises some useful results for the proof of Theorem 2.1.

Lemma 2.1. *Let* $H(x) = (1-\eta)F_1(x) + \eta F_2(x)$, *where*

$$\eta = \frac{\rho\lambda}{(1-\rho)(1-p) + \rho(1-\lambda)(1-p) + \rho\lambda}.$$

Under Conditions A1–A3, we have

(a) $\tilde{\lambda} - \lambda = O_p(N^{-1/2})$, $\tilde{p} - p = O_p(N^{-1/2})$, $\tilde{\lambda}^* - \lambda^* = O_p(N^{-1/2})$;

(b) $t_L \to H^{-1}(q)$ *and* $t_U \to H^{-1}(1-q)$ *in probability;*

(c) $\sup_x |\tilde{F}_1(x) - F_1(x)| = O_p(N^{-1/2})$ *and* $\tilde{F}_1(t_L) \to F_1\big(H^{-1}(q)\big)$ *and* $\tilde{F}_1(t_U) \to F_1\big(H^{-1}(1-q)\big)$ *in probability;*

(d) $\sup_x |\tilde{F}_2^*(x) - F_2(x)| = O_p(N^{-1/2})$ *and* $\tilde{F}_2^*(t_L) \to F_2\big(H^{-1}(q)\big)$ *and* $\tilde{F}_2^*(t_U) \to F_2\big(H^{-1}(1-q)\big)$ *in probability.*

Proof: We start with Part (a). Recall that

$$\tilde{p} = \frac{m_0}{m}, \ \tilde{\lambda} = 1 - \frac{n_0}{n\tilde{p}}.$$

By classical asymptotic theory (Serfling, 2000) and Conditions A1–A2, we have

$$\tilde{p} - p = O_p(N^{-1/2}) \text{ and } (1-\tilde{\lambda})\tilde{p} - (1-\lambda)p = O_p(N^{-1/2}),$$

which together with the delta method imply that $\tilde{\lambda} - \lambda = O_p(N^{-1/2})$ and $\tilde{\lambda}^* - \lambda^* = O_p(N^{-1/2})$. This completes Part (a).

Next we consider Part (b). Note that

$$t_L = H_k^{-1}(q) \quad \text{and} \quad t_U = H_k^{-1}(1-q).$$

To establish the consistency of t_L and t_U, we first argue that

$$\sup_x |H_k(x) - H(x)| = O_p(N^{-1/2}). \tag{2.23}$$

Let

$$F_{X,m}(x) = \frac{1}{m}\sum_{i=1}^{m} I(X_i \le x), \quad F_{Y,n}(x) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \le x).$$

Then for $x > 0$,

$$\begin{aligned} H_k(x) &= \frac{1}{k}\left[m\{F_{X,m}(x) - F_{X,m}(0)\} + n\{F_{Y,n}(x) - F_{Y,n}(0)\}\right] \\ &= \frac{1}{k/(n+m)}\left[(1-\rho)\{F_{X,m}(x) - F_{X,m}(0)\} + \rho\{F_{Y,n}(x) - F_{Y,n}(0)\}\right]. \end{aligned}$$

Note that $H(x)$ can be rewritten as

$$H(x) = \frac{1}{(1-\rho)(1-p) + \rho\{(1-\lambda)(1-p)+\lambda\}}\left[(1-\rho)\{F_X(x) - F_X(0)\} + \rho\{F_Y(x) - F_Y(0)\}\right].$$

By the classical central limit theorem, the delta method, and Condition A2,

$$k/(n+m) - [(1-\rho)(1-p) + \rho\{(1-\lambda)(1-p)+\lambda\}] = O_p(N^{-1/2}). \tag{2.24}$$

By the triangular inequality,

$$\begin{aligned} \sup_x |H_k(x) - H(x)| \le & \left| \frac{1}{k/(n+m)} - \frac{1}{(1-\rho)(1-p) + \rho\{(1-\lambda)(1-p)+\lambda\}} \right| \\ & + \frac{2\sup_x |F_{X,m}(x) - F_X(x)| + 2\sup_x |F_{Y,n}(x) - F_Y(x)|}{(1-\rho)(1-p) + \rho\{(1-\lambda)(1-p)+\lambda\}}. \end{aligned} \tag{2.25}$$

By Conditions A1–A2, the uniform convergence rate of empirical cdfs, and the delta method, (2.24) and (2.25) lead to

$$\sup_x |H_k(x) - H(x)| = O_p(N^{-1/2}).$$

This, together with Condition A3, implies that $t_L \to H^{-1}(q)$ and $t_U \to H^{-1}(1-q)$ in probability. This completes Part (b).

We now consider (c). Similarly to (2.23), we have

$$\sup_x |\tilde{F}_1(x) - F_1(x)| = O_p(N^{-1/2}),$$

which implies that $\tilde{F}_1(t_L) - F_1(t_L) \to 0$ in probability. Further, by the continuous mapping theorem, Part (a), and Condition A3, we have

$$F_1(t_L) - F_1\big(H^{-1}(q)\big) \to 0$$

in probability. Hence,

$$\tilde{F}_1(t_L) - F_1\big(H^{-1}(q)\big) \to 0$$

in probability. Similarly, we can establish the consistency of $\tilde{F}_1(t_U)$. This completes Part (c).

Finally, the proof of Part (d) is similar to that of Part (c), so it is omitted. □

Proof of Theorem 2.1

For convenience of presentation, we define more notation. Let

$$\hat{F}_{Y+} = (1 - \hat{\lambda}^*)\hat{F}_1(x) + \hat{\lambda}^* \hat{F}_2(x),$$

where

$$\hat{\lambda}^* = \frac{\hat{\lambda}}{1 - \hat{p} + \hat{p}\hat{\lambda}}.$$

For $x \in [t_L, t_U]$, let

$$\breve{F}_2^*(x) = \arg \min_{F_2(x) \text{ is increasing}} \sum_{h \in I_q} \{\tilde{F}_2^*(x) - F_2(x)\}^2$$

and

$$\breve{F}_2(x) = \min[\max\{\breve{F}_2^*(x), 0\}, 1].$$

Further, define

$$\breve{F}_{Y+}(x) = (1 - \tilde{\lambda}^*)\tilde{F}_1(x) + \tilde{\lambda}^* \breve{F}_2(x).$$

Let

$$A_{n,m} = \{\tilde{p} \in (0,1), \tilde{\lambda}^* \in (0,1),\ 0 < \tilde{F}_2^*(t_h) < 1,\ h \in I_q\}.$$

By Lemma 2.1 and Condition A1, $P(0 < \tilde{p} < 1) \to 1$, $P(0 < \tilde{\lambda} < 1) \to 1$, and $P(0 < \tilde{F}_2^*(t_h) < 1,\ \ h \in I_q) \to 1$ as $N \to \infty$. We therefore concentrate on the sample points in $A_{n,m}$. Since $\breve{F}_2^*(x)$ is the isotonic regression function of $\tilde{F}_2^*(x)$, we have $\max_{h\in I_q} \breve{F}_2^*(t_h) \leq \max_{h\in I_q} \tilde{F}_2^*(t_h)$ and $\min_{h\in I_q} \breve{F}_2^*(t_h) \geq \min_{h\in I_q} \tilde{F}_2^*(t_h)$. As a result, for every sample point in $A_{n,m}$, we also have $\breve{F}_2^*(t_h) = \breve{F}_2(t_h)$.

The roadmap for Theorem 2.1 is as follows. In the first step, we find a lower bound for $l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$, which is a quadratic function of $\hat{p} - \tilde{p}$, $(1-\hat{\lambda})\hat{p} - (1-\tilde{\lambda})\tilde{p}$, $\hat{F}_1 - \tilde{F}_1$, and $\hat{F}_{Y+} - \tilde{F}_{Y+}$. In the second step, we argue that $l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ is bounded above by an $O_p(N)$ term. The results then follow.

For the first step, note that if x is a fixed number, then (Lee et al., 2016)

$$J(x,x) - J(x,y) \geq 0.5(x-y)^2.$$

Hence,

$$\begin{aligned}
& l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2) \\
\geq\ & \frac{1}{2} m k_q (\hat{p} - \tilde{p})^2 + \frac{1}{2} n k_q \left\{(1-\hat{\lambda})\hat{p} - (1-\tilde{\lambda})\tilde{p}\right\}^2 \\
& + \frac{m_+}{2} \sum_{h\in I_q} \left\{\hat{F}_1(t_h) - \tilde{F}_1(t_h)\right\}^2 + \frac{n_+}{2} \sum_{h\in I_q} \left\{\hat{F}_{Y+}(t_h) - \tilde{F}_{Y+}(t_h)\right\}^2. \qquad (2.26)
\end{aligned}$$

For the second step, we find an upper bound for $l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$. Since $(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2)$ is the maximum multinomial likelihood estimator, we have

$$l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\hat{\lambda}, \hat{p}, \hat{F}_1, \hat{F}_2) \leq l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \breve{F}_2).$$

Therefore, it suffices to find an upper bound for $l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \breve{F}_2)$.

Let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Applying the first-order Taylor expansion, we have

$$\begin{aligned}
& l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \breve{F}_2) \\
=& \sum_{h \in I_q} n_+ \left\{ J\big(\tilde{F}_{Y+}(t_h), \tilde{F}_{Y+}(t_h)\big) - J\big(\tilde{F}_{Y+}(t_h), \breve{F}_{Y+}(t_h)\big) \right\} \\
=& \sum_{h \in I_q} n_+ \left\{ \frac{\tilde{F}_{Y+}(t_h)}{\delta^2(t_h)} + \frac{1 - \tilde{F}_{Y+}(t_h)}{\{1 - \delta(t_h)\}^2} \right\} \left\{ \tilde{F}_{Y+}(t_h) - \breve{F}_{Y+}(t_h) \right\}^2,
\end{aligned}$$

where $\delta(t_h) \in [\tilde{F}_{Y+}(t_h) \wedge \breve{F}_{Y+}(t_h), \tilde{F}_{Y+}(t_h) \vee \breve{F}_{Y+}(t_h)]$.

By the definition of $\tilde{F}_{Y+}(x)$ and $\breve{F}_{Y+}(x)$, we further have for every sample point in $A_{n,m}$,

$$\tilde{F}_{Y+}(t_h) \geq (1 - \tilde{\lambda}^*)\tilde{F}_1(t_L), \quad \breve{F}_{Y+}(t_h) \geq (1 - \tilde{\lambda}^*)\tilde{F}_1(t_L),$$

and

$$1 - \tilde{F}_{Y+}(t_h) \geq (1 - \tilde{\lambda}^*) \left\{ 1 - \tilde{F}_1(t_U) \right\}, \quad 1 - \breve{F}_{Y+}(t_h) \geq (1 - \tilde{\lambda}^*) \left\{ 1 - \tilde{F}_1(t_U) \right\}.$$

Therefore,

$$\begin{aligned}
l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \tilde{F}_2^*) - l_M(\tilde{\lambda}, \tilde{p}, \tilde{F}_1, \breve{F}_2) \leq & \frac{n_+ \left(\tilde{\lambda}^* \right)^2}{(1 - \tilde{\lambda}^*)^2 [\tilde{F}_1(t_L)\{1 - \tilde{F}_1(t_U)\}]^2} \sum_{h \in I_q} \left\{ \tilde{F}_2^*(t_h) - \breve{F}_2(t_h) \right\}^2 \\
=& O_p(N) \sum_{h \in I_q} \left\{ \tilde{F}_2^*(t_h) - \breve{F}_2^*(t_h) \right\}^2 \\
\leq & O_p(N) \sum_{h \in I_q} \left\{ \tilde{F}_2^*(t_h) - F_2(t_h) \right\}^2 \\
=& O_p(N), \qquad (2.27)
\end{aligned}$$

where the last three steps follow from Lemma 2.1 and $\breve{F}_2^*(t_h) = \breve{F}_2(t_h)$ for any sample point in $A_{n,m}$.

Combining (2.26) and (2.27), we have

$$mk_q(\hat{p} - \tilde{p})^2 = O_p(N), \tag{2.28}$$

$$nk_q \left\{(1-\hat{\lambda})\hat{p} - (1-\tilde{\lambda})\tilde{p}\right\}^2 = O_p(N), \tag{2.29}$$

$$\frac{m_+}{2} \sum_{h \in I_q} \left\{\hat{F}_1(t_h) - \tilde{F}_1(t_h)\right\}^2 = O_p(N), \tag{2.30}$$

$$\frac{n_+}{2} \sum_{h \in I_q} \left\{\hat{F}_{Y+}(t_h) - \tilde{F}_{Y+}(t_h)\right\}^2 = O_p(N). \tag{2.31}$$

By Conditions A1 and A2, k_q, m, n, m_+, and n_+ all have the same order as N. Hence, (2.28) and (2.29) lead to

$$\hat{p} - \tilde{p} = O_p(N^{-1/2}), \quad (1-\hat{\lambda})\hat{p} - (1-\tilde{\lambda})\tilde{p} = O_p(N^{-1/2}),$$

which together with Part (a) of Lemma 2.1 implies

$$\hat{p} - p = O_p(N^{-1/2}), \quad \hat{\lambda} - \lambda = O_p(N^{-1/2}).$$

Similarly, (2.30) implies

$$\sum_{h \in I_q} \left\{\hat{F}_1(t_h) - \tilde{F}_1(t_h)\right\}^2 = O_p(1).$$

By Part (c) of Lemma 2.1 and the triangular inequality, we easily conclude that

$$\sum_{h \in I_q} \left\{\hat{F}_1(t_h) - F_1(t_h)\right\}^2 = O_p(1). \tag{2.32}$$

Since k_q has the same order as N, we further have

$$\frac{1}{k_q} \sum_{h \in I_q} \left\{\hat{F}_1(t_h) - F_1(t_h)\right\}^2 = O_p(N^{-1}).$$

Similarly to (2.23), we have

$$\sup_x |\tilde{F}_{Y+}(x) - F_{Y+}(x)| = O_p(N^{-1/2}),$$

which together with (2.31) implies that

$$\sum_{h \in I_q} \left\{ \hat{F}_{Y+}(t_h) - F_{Y+}(t_h) \right\}^2 = O_p(1).$$

With $\hat{p} - p = O_P(N^{-1/2})$, $\hat{\lambda} - \lambda = O_P(N^{-1/2})$, and the delta method, we have $\hat{\lambda}^* - \lambda^* = O_p(N^{-1/2})$. By the triangular inequality and the form of $\hat{F}_{Y+}(t_h)$, we get

$$\sum_{h \in I_q} \left\{ (1 - \hat{\lambda}^*)\hat{F}_1(t_h) + \hat{\lambda}^* \hat{F}_2(t_h) - (1 - \hat{\lambda}^*)F_1(t_h) - \hat{\lambda}^* F_2(t_h) \right\}^2 = O_p(1).$$

By (2.32) and the triangular inequality, we obtain

$$(\hat{\lambda}^*)^2 \sum_{h \in I_q} \left\{ \hat{F}_2(t_h) - F_2(t_h) \right\}^2 = O_p(1),$$

which, together with

$$\hat{\lambda}^* - \lambda^* = O_p(N^{-1/2})$$

and the fact that k_q has the same order as N, implies

$$\sum_{h \in I_q} \left\{ \hat{F}_2(t_h) - F_2(t_h) \right\}^2 = O_p(1).$$

This completes the proof of Theorem 2.1.

2.6.4 Details of EM-algorithm

Based on $\{\mathcal{X}, \mathcal{V}\}$, the complete multinomial likelihood has the following form:

$$l_M^c(\lambda, p, F_1, F_2) = k_q \left[m_0 \log p + m_+ \log(1-p) + n_0 \log\{p(1-\lambda)\} + n_+ \log\{1 - p(1-\lambda)\} \right]$$

$$+\sum_{h\in I_q}\sum_{i=1}^{m_+}\left\{m_{i1}(t_h)\log F_1(t_h)+m_{i2}(t_h)\log\bar{F}_1(t_h)\right\}$$
$$+\sum_{h\in I_q}\sum_{j=1}^{n_+}(1-V_{jh})\left[n_{j1}(t_h)\log\{(1-\lambda^*)F_1(t_h)\}+n_{j2}(t_h)\log\{(1-\lambda^*)\bar{F}_1(t_h)\}\right]$$
$$+\sum_{h\in I_q}\sum_{j=1}^{n_+}V_{jh}\left[n_{j1}(t_h)\log\{\lambda^*F_2(t_h)\}+n_{j2}(t_h)\log\{\lambda^*\bar{F}_2(t_h)\}\right].$$

In the E-step of the rth iteration, for $h=1,\ldots,k$ and $j=1,\ldots,n_+$, we need to calculate

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})=E\left\{l_M^c(\lambda,p,F_1,F_2)|\mathcal{X},\boldsymbol{\theta}^{(r-1)}\right\},$$

where the expectation is with respect to the conditional distribution of $\mathcal{V}$ given $\mathcal{X}$ and substituting $\boldsymbol{\theta}^{(r-1)}$ for $\boldsymbol{\theta}$. With $n_{j1}(t_h)\sim(1-\lambda^*)\mathrm{Bin}\Big(1,F_1(t_h)\Big)+\lambda^*\mathrm{Bin}\Big(1,F_2(t_h)\Big)$, it can be checked that

$$E(V_{jh}|\mathcal{X},\boldsymbol{\theta}^{(r-1)})=\left\{a_h^{(r)}\right\}^{n_{j1}(t_h)}\left\{b_h^{(r)}\right\}^{n_{j2}(t_h)}.$$

Therefore,

$$\begin{aligned}
&Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)})\\
&=k_q\left[m_0\log p+m_+\log(1-p)+n_0\log\{p(1-\lambda)\}+n_+\log\{1-p(1-\lambda)\}\right]\\
&\quad+\sum_{h\in I_q}\sum_{i=1}^{m_+}\left\{m_{i1}(t_h)\log F_1(t_h)+m_{i2}(t_h)\log\bar{F}_1(t_h)\right\}\\
&\quad+\sum_{h\in I_q}\sum_{j=1}^{n_+}\left[(1-a_h^{(r)})n_{j1}(t_h)\log\{(1-\lambda^*)F_1(t_h)\}+(1-b_h^{(r)})n_{j2}(t_h)\log\{(1-\lambda^*)\bar{F}_1(t_h)\}\right]\\
&\quad+\sum_{h\in I_q}\sum_{j=1}^{n_+}\left[a_h^{(r)}n_{j1}(t_h)\log\{\lambda^*F_2(t_h)\}+b_h^{(r)}n_{j2}(t_h)\log\{\lambda^*\bar{F}_2(t_h)\}\right].
\end{aligned}$$

In the M-step, we update (λ,p,F_1,F_2) by

$$(\lambda^{(r)},p^{(r)},F_1^{(r)},F_2^{(r)})=\arg\max_{(\lambda,p,F_1,F_2)\in\Theta}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}).$$

After some algebra work, it can be shown that

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r-1)}) = Q_1(\lambda) + Q_2(p) + Q_3(F_1) + Q_4(F_2),$$

where

$$\begin{aligned}
Q_1(\lambda) =& k_q n_0 \log(1-\lambda) + \sum_{h\in I_q} \left\{ (1-a_h^{(r)}) n_1(t_h) + (1-b_h^{(r)}) n_2(t_h) \right\} \log(1-\lambda) \\
&+ \sum_{h\in I_q} \left\{ a_h^{(r)} n_1(t_h) + b_h^{(r)} n_2(t_h) \right\} \log\lambda, \\
Q_2(p) =& k_q \left[m_0 \log p + m_+ \log(1-p) + n_0 \log p \right] \\
&+ \sum_{h\in I_q} \left\{ (1-a_h^{(r)}) n_1(t_h) + (1-b_h^{(r)}) n_2(t_h) \right\} \log(1-p), \\
Q_3(F_1) =& \sum_{h\in I_q} \left[\left\{ m_1(t_h) + (1-a_h^{(r)}) n_1(t_h) \right\} \log F_1(t_h) + \left\{ m_2(t_h) + (1-b_h^{(r)}) n_2(t_h) \right\} \log \bar{F}_1(t_h) \right], \\
Q_4(F_2) =& \sum_{h\in I_q} \left[a_h^{(r)} n_1(t_h) \log F_2(t_h) + b_h^{(r)} n_2(t_h) \log \bar{F}_2(t_h) \right].
\end{aligned}$$

Hence, we can update λ via

$$\lambda^{(r)} = \arg\max_{\lambda} Q_1(\lambda) = \frac{1}{k_q n} \sum_{h\in I_q} \left\{ n_1(t_h) a_h^{(r)} + n_2(t_h) b_h^{(r)} \right\}$$

and update p via

$$p^{(r)} = \arg\max_{\lambda} Q_2(p) = \frac{m_0 + n_0}{m + n - n\lambda^{(r)}}.$$

To update F_1 and F_2, we have

$$F_1^{(r)} = \arg \max_{F_1 \text{ is a cdf}} Q_3(F_1), \quad F_2^{(r)} = \arg \max_{F_2 \text{ is a cdf}} Q_4(F_2).$$

Following Dykstra, Kochar, and Robertson (1995), we can obtain $F_1^{(r)}$ and $F_2^{(r)}$ via

$$F_1^{(r)} = \arg \min_{F \text{ is a cdf}} \sum_{h\in I_q} \left\{ m_+ + n_1(t_h) \left(1 - a_h^{(r)}\right) + n_2(t_h) \left(1 - b_h^{(r)}\right) \right\} \left\{ \tilde{F}_1^{(r)}(t_h) - F(t_h) \right\}^2,$$

$$F_2^{(r)} = \arg \min_{F \text{ is a cdf}} \sum_{h \in I_q} \left\{ n_1(t_h) a_h^{(r)} + n_2(t_h) b_h^{(r)} \right\} \left\{ \tilde{F}_2^{(r)}(t_h) - F(t_h) \right\}^2,$$

where

$$\tilde{F}_1^{(r)}(t_h) = \frac{m_1(t_h) + n_1(t_h)\{1 - a_h^{(r)}\}}{m_+ + n_1(t_h)\{1 - a_h^{(r)}\} + n_2(t_h)\{1 - b_h^{(r)}\}} \text{ and } \tilde{F}_2^{(r)}(t_h) = \frac{n_1(t_h) a_h^{(r)}}{n_1(t_h) a_h^{(r)} + n_2(t_h) b_h^{(r)}}.$$

Chapter 3

A powerful procedure to control the false discovery rate with directional information

3.1 Introduction

In many modern statistical applications, hundreds or thousands of hypotheses are tested simultaneously. For example, in microarray or RNA-seq studies aiming to find differentially expressed (DE) genes between diseased and healthy subjects, up to about twenty thousand hypotheses may be tested at the same time. To adjust for multiplicity, we aim to control the false discovery rate (FDR), which is the expected proportion of false discoveries among rejections (Benjamini and Hochberg, 1995).

Most existing procedures, such as those reviewed in Section 1.3.2, control the FDR based on p-values. In many practical applications, interesting findings have directions,

taking Example 1.2 for instance. In this example, we present a dataset consisting of gene expression levels from a prostate cancer study. The goal of the study is to find DE genes between diseased and healthy subjects. In order to compare the gene expression levels between the two subject groups, two-sample t-tests are used. The signs of the resulting t-statistics t_i's suggest whether the corresponding genes are potentially up- or down-regulated. However, only the absolute value of t_i is used when computing the two-sided p-values, and the directional information is lost in the p-value.

Thresholding on the p-values leads to symmetric rejection boundaries on the positive and negative sides of the original test statistics, such as the t-statistics in Example 1.2. By allowing asymmetric rejection boundaries, direction-adaptive procedures can improve power over traditional procedures. Examples include Sun and Cai (2007), Orr, Liu, and Nettleton (2014) and Zhao and Fung (2016), which are discussed in Section 1.3.2. However, none of them provide the FDR control in finite samples.

In many applications, conditional on the signs, the true null p-values still follow Uniform(0,1). To control the FDR, we can use some recently developed methods by treating the signs of the original test statistics as group labels. Li and Barber (2019) develop a structure-adaptive algorithm that controls the FDR in the grouped setting. If we treat the group label as a covariate, we can use the covariate-adaptive methods of Lei and Fithian (2018) and Ignatiadis and Huber (2018) to control the FDR in finite samples. However, the grouping strategy models the positive and negative statistics as two separate groups, which is not appropriate when there is only one true null distribution. In our motivating example, the true null t-statistics follow a central t distribution. Dividing the test statistics according to their signs artificially splits the null statistics into two groups. The above methods estimate the true null proportion separately in each group, and there is no guarantee that the two null distributions will be compatible. Other methods have been

developed using this grouping approach, for example, Orr, Liu, and Nettleton (2014) and Zhao and Fung (2016), but they do not provide FDR control in finite samples.

Based on the idea of knockoffs from Barber and Candès (2015) and Lei and Fithian (2018), we develop a powerful procedure to utilize the directional information in multiple testing problems. First, we introduce the signed p-value to preserve the directional information of the original test statistic. Next, knockoffs are constructed to mimic the behaviors of the signed p-values under true null hypotheses. Unordered pairs of the signed p-values and their knockoffs are given in the beginning instead of the original data, and the true identities of signed p-values will be gradually revealed in later steps. The number of knockoffs in the rejection region provides an empirical estimate of the number of false positives along the way, which allows us to control the FDR.

The rest of Chapter 3 is organized as follows. In Section 3.2, we propose a novel procedure called the signed-knockoff procedure and show that it controls the FDR in finite samples. In Section 3.3, through mixture modeling, we develop an algorithm to maximize the power of our procedure. We numerically evaluate the proposed signed-knockoff procedure through simulation studies in Section 3.4 and in real data applications in Section 3.5. For the convenience of presentation, technical details and additional simulation results are given in Section 3.6.

3.2 Signed-knockoff procedure

For the ease of discussion, we will assume that the test statistics and thus the p-values are all continuously distributed. Furthermore, we focus on the common setting of testing a simple null versus a composite alternative. For the ith gene in Example 1.2, the null hypothesis is $H_{0i} : \mu_{1i} = \mu_{2i}$, and the alternative hypothesis is $H_{1i} : \mu_{1i} \neq \mu_{2i}$, where μ_{1i}

and μ_{2i} are the mean expression levels for diseased and healthy subjects, respectively. The probability density function of the true null test statistic is typically symmetric about zero, and we call it the *symmetric null condition*. Examples include the central t distribution and the standard normal distribution. For such symmetric distributions, a true null test statistic has an equal probability to be positive or negative, and its sign is independent of the corresponding p-values.

3.2.1 Signed p-values

Suppose that t_i is the test statistic of the ith null hypothesis, and p_i is the corresponding p-value. To preserve the directional information in t_i, we could directly work on t_i but will need to pay attention to its null distribution, which may vary from application to application. Alternatively, we can combine $\text{sign}(t_i)$ and p_i to obtain a new statistic.

Definition (Signed p-value). *For the ith null hypothesis with the test statistic t_i and the p-value p_i, the signed p-value q_i is defined as*

$$q_i = \text{sign}(t_i)(1 - p_i).$$

We comment on some properties of the signed p-values. Under the symmetric null condition, it is straightforward to show that

$$q_i = 2F(t_i) - 1, \tag{3.1}$$

where F is the cdf of the test statistic under the null. It is clear that the signed p-value has a one-to-one relationship with the original test statistic. The signed p-value centers the cdf so that its sign matches the sign of the original test statistic and is scaled to range from -1 to 1. Furthermore, q_i follows $\text{Uniform}(-1, 1)$ if $i \in \mathcal{H}_0$. Finally, as the p-values

under alternative hypotheses are more likely to be close to 0, the signed p-values under alternative hypotheses tend to be close to -1 and 1. Thus, we expect the rejection region of signed p-values to take the form of $[-1, a) \bigcup (b, 1]$, where a is a negative number and b is a positive number. If we have defined the signed p-value using p_i instead of $(1 - p_i)$, not only the density of the signed p-values under alternative hypotheses would not be continuous, but also we would lose the simple expression in (3.1).

3.2.2 Proposed procedure

Similar in spirit to the mirror knockoff construction in Lei and Fithian (2018), we define the *knockoff* of the ith signed p-value q_i as

$$\tilde{q}_i = \text{sign}(q_i) - q_i.$$

As p-values are continuously distributed, no signed p-values should be exactly equal to 0, so we do not consider how to set the values of knockoffs for zeros. From the definition, q_i and $\tilde{q}_i$ share the same signs. If q_i is positive, then q_i and $\tilde{q}_i$ are symmetric about $1/2$. Similarly, if q_i is negative, then q_i and $\tilde{q}_i$ are symmetric about $-1/2$. Furthermore, if the ith null hypothesis is true, q_i and $\tilde{q}_i$ would both follow Uniform$(-1, 1)$. In summary, under the true null hypothesis, $\tilde{q}_i$ has the same distribution as q_i and hence is a knockoff of q_i. Note that the idea of knockoff originated in the feature selection problem in Barber and Candès (2015). Although their setting of knockoff is quite different from ours and the knockoffs in Lei and Fithian (2018), the general ideas are similar. All of the three methods set the knockoff as an imitation of the original statistic under true null hypotheses so that they are exchangeable.

Before getting into the details of our proposed procedure, we describe the general idea behind it. Our procedure is a stepwise procedure. We will start with an initial guess of

the rejection region, which should be reasonably large. In general, the less the number of rejections, the less the FDP. Thus, we make the rejection region R shrink towards the two endpoints -1 and 1 gradually at each step by accepting one null hypothesis at a time. At each step, we estimate the FDR as

$$\widehat{\text{FDR}} = \frac{1 + \#\{i : \ \tilde{q}_i \in R\}}{\#\{i : \ q_i \in R\} \vee 1}.$$

Because $\tilde{q}_i$ follows the same distribution as q_i under the true null hypothesis, the number of false discoveries $\#\{i : \ i \in \mathcal{H}_0 \ \& \ q_i \in R\}$ follows the same distribution as $\#\{i : \ i \in \mathcal{H}_0 \ \& \ \tilde{q}_i \in R\}$. Thus, the numerator of $\widehat{\text{FDR}}$, $1 + \#\{i : \ \tilde{q}_i \in R\}$, provides a slight over-estimate of the number of false discoveries. $\widehat{\text{FDR}}$ tends to decrease after repeated shrinkages. We will stop the shrinkage at the first time $\widehat{\text{FDR}} \leq \alpha$ or when all null hypotheses have been accepted, whichever comes the first.

We will first define some notation. We divide the signed p-values $\{q_i\}_{i=1}^n$ into two groups, $\{q_i^+\}_{i=1}^{n_+}$ and $\{q_i^-\}_{i=1}^{n_-}$, according to their signs. Denote $q_{(i)}^+$ as the positive signed p-value which is the ith closest to 1/2, and $q_{(j)}^-$ as the negative signed p-value which is the jth closest to $-1/2$. Let α be the nominal FDR level.

After k shrinkage steps in our procedure, let i_k and j_k be the number of accepted hypotheses on the positive side and negative side, respectively. Then we can define the rejection region after k shrinkage steps as $R_k = [-1, q_{(j_k)}^- \wedge \tilde{q}_{(j_k)}^-) \bigcup (q_{(i_k)}^+ \vee \tilde{q}_{(i_k)}^+, 1]$, where $a \wedge b = \min(a, b)$.

Let $\mathcal{I}_k = \{i : \ q_i \in \{q_{(1)}^+, ..., q_{(i_k)}^+, q_{(1)}^-, ..., q_{(j_k)}^-\}\}$ be the index set of accepted hypotheses after k shrinkage steps, and $\mathcal{J}_k = \{i : \ i = 1, ..., n\}/\mathcal{I}_k$ be the index set of unaccepted hypotheses after k steps. Let $b_i = I(|q_i| > 1/2)$ and $q_i^* = q_i \wedge \tilde{q}_i$. Define the σ-algebra $\mathcal{F}_k$ as

$$\mathcal{F}_k = \sigma\Big(\{q_i^*\}_{i=1}^n, \ \{b_i\}_{i \in \mathcal{I}_k}\Big).$$

Later in Step 2 of our procedure, we will require a decision to be $\mathcal{F}_k$-measurable. Let $\{q_i, \tilde{q}_i\}$ denote the unordered pair of q_i and its knockoff $\tilde{q}_i$. In $\mathcal{F}_k$, $\{q_i^*\}_{i=1}^n$ contains all the location information of $\{q_i, \tilde{q}_i\}$ pairs, and b_i is the indicator of which element of $\{q_i, \tilde{q}_i\}$ is q_i. Thus, $\mathcal{F}_k$ encapsulates the knowledge of all unordered pairs $\{q_i, \tilde{q}_i\}$'s of currently unaccepted null hypotheses plus the q_i values of already accepted null hypotheses after the kth step. In other words, for the currently unaccepted null hypotheses, the true identities of the signed p-values in the unordered pairs are *masked* from us. Once we decide to accept a null hypothesis, the true identity of the corresponding signed p-value will be *revealed* to us. Our proposed procedure is as follows.

Definition (Signed-knockoff procedure).

Step 1: Initialization.

Set the shrinkage step number $k = 0$. For convenience, we set $i_0 = j_0 = 1$ and let R_0 be the initial rejection region.

Step 2: Choice.

At the $(k+1)$th shrinkage step, we make the rejection region shrink to R_{k+1} by accepting one null hypothesis. More specifically, we choose one of the following:

① Accept a null hypothesis on the positive side: set $i_{k+1} = i_k + 1$ and $j_{k+1} = j_k$.

② Accept a null hypothesis on the negative side: set $i_{k+1} = i_k$ and $j_{k+1} = j_k + 1$.

As boundary conditions, we choose ① if $j_k = n_-$, and we choose ② if $i_k = n_+$.

If neither i_k nor j_k reaches its bound, we require the choice between ① and ② to be $\mathcal{F}_k$-measurable.

Increase k by 1.

Step 3: Stopping condition.

If

$$\widehat{\mathrm{FDR}}_k = \frac{1 + \#\{i : \ \tilde{q}_i \in R_k\}}{\#\{i : \ q_i \in R_k\} \vee 1} \leq \alpha,$$

or

$$i_k = n_+ \text{ and } j_k = n_-,$$

we stop and reject all null hypotheses whose corresponding signed p-values are in R_k. Otherwise, we return to Step 2.

3.2.3 False discovery rate control

We can see that the procedure above leaves out a key point, that is, how to decide whether to choose ① or ② in Step 2. In fact, it can be arbitrarily chosen as long as the choices are $\mathcal{F}_k$-measurable, that is, the choices are made based on the partially masked information in $\mathcal{F}_k$. We can show that regardless of the choices, the signed-knockoff procedure controls the FDR at the nominal level α under the *null independence condition* introduced in Section 1.3.2. The condition requires that the true null test statistics are independent of each other and independent of alternative test statistics.

Theorem 3.1. *Under the null independence condition, if the null p-values follow Uniform*$(0, 1)$ *and are independent of the signs of their corresponding test statistics, respectively, the signed-knockoff procedure controls the FDR at level* α*.*

The proof of Theorem 3.1 is similar in spirit to the proof of Theorem 2 in Barber and Candès (2015) and is presented in Section 3.6.1. The independence between the signs and

null p-values is naturally satisfied under the symmetric null condition and also holds when the p-values are generated from permutation tests.

Remark: The AdaPT method proposed by Lei and Fithian (2018) includes two parts: a framework which is guaranteed to control the FDR in finite samples and an implementation to maximize power with generic covariates. Because a signed p-value is defined by its corresponding sign and p-value, our procedure can be viewed as a special instance of the AdaPT theoretical framework with the sign as a covariate. More thorough technical arguments are given in Section 3.6.3. As a result, we can use their Theorem 1 to show that `SK` controls the FDR in finite samples with not only uniform but also mirror conservative null p-values, which include the permutation p-values. However, our modeling approach differs from the AdaPT method with the sign covariate, and we will illustrate their difference in the rest of Chapter 3.

3.3 Powerful choice

In Step 2 of our proposed procedure, the choices of which side to proceed on will not lead to the loss of FDR control, but they will affect power. In the ideal case where the data generating model is completely known, it is well established in the Bayesian decision-theoretic framework that the local FDR is the optimal ranking statistic to maximize power with a constraint on the FDR, see Müller et al. (2004), Sun and Cai (2007), and Lei and Fithian (2018). This result implies that we should choose the pair with larger local FDR, or equivalently, the pair whose corresponding null hypothesis is more likely to be true.

We propose the following parametric mixture model, whose parameters can be estimated in a computationally efficient manner. If all signed p-values are available to us, we

can consider the two-group model (Efron et al., 2001):

$$f(q) = \pi_0 f_0(q) + (1 - \pi_0) f_1(q),$$

where π_0 is the probability that the null hypothesis is true, and f_0, f_1, and f are the null, alternative, and overall densities of the signed p-values, respectively. Under the true null hypothesis, the signed p-value follows Uniform$(-1, 1)$, i.e., $f_0(q) = 1/2$. For the alternative density, we model it parametrically as a two-component mixture of transformed beta distributions. More specifically, let

$$f_1(q) = \lambda \frac{\alpha}{2} \left(\frac{q+1}{2} \right)^{\alpha-1} + (1 - \lambda) \frac{\beta}{2} \left(\frac{1-q}{2} \right)^{\beta-1},$$

where λ is the mixing proportion, and α and β are the shape parameters of the component beta distributions. The two components are regular beta distributions transformed from the range of $(0, 1)$ to the range of $(-1, 1)$. We set one of the shape parameters in each beta distribution to 1 so that the alternative negative or positive signed p-value density is strictly decreasing or increasing, respectively, over the range $(-1, 1)$.

Recall that in Step 2 of the signed-knockoff procedure after k shrinkage steps, the decision needs to be made based on $\mathcal{F}_k$. If the ith hypothesis has not been accepted, instead of its signed p-value q_i, we only observe the masked unordered pair $\{q_i, \tilde{q}_i\}$, or equivalently, $q_i^* = q_i \wedge \tilde{q}_i$. The probability density function of q_i^* is the sum of densities of q_i and $\tilde{q}_i$:

$$g(q_i^*) = \pi_0 + (1 - \pi_0) f_1(q_i) + (1 - \pi_0) f_1(\tilde{q}_i).$$

The log-likelihood function based on $\mathcal{F}_k$ is

$$l^{(k)}(\pi_0, \lambda, \alpha, \beta) = \sum_{i \in \mathcal{I}_k} \log f(q_i) + \sum_{i \in \mathcal{J}_k} \log g(q_i^*)$$

$$= \sum_{i \in \mathcal{I}_k} \log \Big\{ \frac{\pi_0}{2} + (1-\pi_0) f_1(q_i; \lambda, \alpha, \beta) \Big\}$$
$$+ \sum_{i \in \mathcal{J}_k} \log \Big\{ \pi_0 + (1-\pi_0) f_1(q_i; \lambda, \alpha, \beta) + (1-\pi_0) f_1(\tilde{q}_i; \lambda, \alpha, \beta) \Big\}.$$

Maximizing the likelihood directly with respect to π_0, λ, α, and β can be numerically challenging. Instead, we develop an EM-algorithm by treating the true null statuses of all hypotheses and the true identities of q_i's within the masked pairs as missing information. The details of the EM-algorithm are described in Section 3.6.2, where the explicit solutions of parameters are derived. Through the EM-algorithm we obtain the maximum likelihood estimators $\hat{\pi}_0$, $\hat{\lambda}$, $\hat{\alpha}$, and $\hat{\beta}$. Finally, we compute the local FDR estimate for an unordered pair as

$$\text{Lfdr}(\{q, \tilde{q}\}) = \frac{\hat{\pi}_0}{g(q^*; \hat{\pi}_0, \hat{\lambda}, \hat{\alpha}, \hat{\beta})}.$$

After k shrinkage steps in Step 2 of the signed-knockoff procedure, we choose ① to proceed on the positive side if

$$\text{Lfdr}(\{q^+_{(i_k+1)}, \tilde{q}^+_{(i_k+1)}\}) \geq \text{Lfdr}(\{q^-_{(j_k+1)}, \tilde{q}^-_{(j_k+1)}\}),$$

i.e., if the next positive pair is more likely to come from the true null hypothesis than the next negative pair. We will choose ② otherwise.

In the above local FDR comparison, the exact value of $\hat{\pi}_0$ is inconsequential because it is the common numerator on both sides and can be canceled out. The crucial part in the comparison is the overall density estimate $g(q^*)$ in the denominator. These observations imply that our proposed parametric mixture model does not need to be a correctly specified model for the signed p-values. To have good power, we only need our mixture model provides a reasonable fit for the overall density so that we can make approximately correct choices.

In the implementation, running the EM-algorithm at each shrinkage step can be time-consuming. Similar to Lei and Fithian (2018), we rerun the EM-algorithm every $\lceil n/20 \rceil$ shrinkage steps to limit the total run times to be no more than 20.

3.4 Simulation study

In this section, we conduct simulation studies to compare the numerical performance of the proposed signed-knockoff procedure with other direction-adaptive procedures. The power advantage of direction-adaptive procedures over p-value-based methods has been demonstrated in the literature, see simulation results of Sun and Cai (2007) and Zhao and Fung (2016). Further evidence of this power advantage can also be found in Section 3.6.4.

3.4.1 Candidate procedures

We consider the following candidate procedures:

– `SK`, the proposed signed-knockoff procedure.

– `ORC`, the oracle signed-knockoff procedure. The only difference between `ORC` and `SK` is that the choice in Step 2 is based on the true local FDR of the unordered pairs instead of the estimates and the EM-algorithm is not needed.

– `AdaPT`, the adaptive p-value thresholding procedure proposed by Lei and Fithian (2018). The signs of test statistics are used as generic side information in the AdaPT implementation available in the `R` package `AdaptMT` on CRAN.

– `SABHA`, the structure-adaptive Benjamini–Hochberg algorithm proposed by Li and Barber (2019). The signs of test statistics are used to form two groups. For a given nominal

level of α, `SABHA` is shown to control the FDR at the level of $(1 + C)\alpha$, where C is a constant determined by the group sizes. To control the FDR at the same level as other competing methods, we use the adjusted nominal level of $\alpha/(1 + C)$. The performance of the unadjusted version of `SABHA` is shown in Section 3.6.4. The `R` code of `SABHA` is available at `http://www.stat.uchicago.edu/~rina/sabha.html`.

– `IHWc`, Independent Hypothesis Weighting with censoring proposed by Ignatiadis and Huber (2018). The signs of test statistics are used as the covariate. `IHWc` is available in the `R` package `IHW` on Bioconductor, and we use the `ihw` function with the parameter `censoring` set to `TRUE`.

The simulation results are based on 500 repeated runs, and we report the average FDR and power relative to `ORC`.

3.4.2 Independent normal statistics

The simulation setting here is similar to that in Sun and Cai (2007) and Zhao and Fung (2016). More specifically, we generate test statistics z_i independently and identically from the following normal mixture model:

$$z_i \sim (1 - p_1 - p_2)\mathrm{N}(0, 1) + p_1\mathrm{N}(\mu_1, 1) + p_2\mathrm{N}(\mu_2, 1),$$

where $\mu_1 < 0 < \mu_2$. The test statistics from the null distribution follow the standard normal distribution. Alternative test statistics follow the normal distribution with mean μ_1 or μ_2 and standard deviation 1. We emphasize that the generative model above differs from the uniform-beta mixture model we used to compute local FDR estimates in Section 3.3. That is, our estimation model is misspecified.

We study the following three cases of parameter combinations, which follow the simulation settings in Sun and Cai (2007):

(a) Let p_1 vary between 0 and 0.2 and set $p_2 = 0.2 - p_1$, $\mu_1 = -3$ and $\mu_2 = 3$.

(b) Let p_1 vary between 0 and 0.2 and set $p_2 = 0.2 - p_1$, $\mu_1 = -3$ and $\mu_2 = 6$.

(c) Let μ_2 vary between 2 and 6 and set $p_1 = 0.18$, $p_2 = 0.02$ and $\mu_1 = -3$.

Under each case, we test $n = 5000$ hypotheses simultaneously and set the nominal FDR level $\alpha = 0.1$.

The left column in Figure 3.1 plots the average realized FDR levels. Across all three cases, we can see that the FDR levels of `SK`, `ORC` and `AdaPT` are close to each other and to the target level. `SABHA` and `IHWc` are conservative.

The right column in Figure 3.1 plots the relative power. Across all three cases, the power of `SK` is close to `ORC` and generally more powerful than other procedures. We emphasize that `SK` achieves almost the power of `ORC` despite that we fit a misspecified parametric model based on limited information. `AdaPT` has the second-best overall power, but `SK` is at least as powerful as `AdaPT` in all settings. From the simulation results, we estimate that the standard deviation of the relative power of `SK` and `AdaPT` to be less than 0.002 in every case. Thus, `SK` shows significant power advantage over `AdaPT` in many parameter settings, such as in case (c) when $\mu_2 < 4$. In the vast majority of parameter settings, `AdaPT` is significantly more powerful than `SABHA` and `IHWc`. When $p_1 \leq 0.04$ in case (b), `AdaPT` has less power than `SABHA` and `IHWc`.

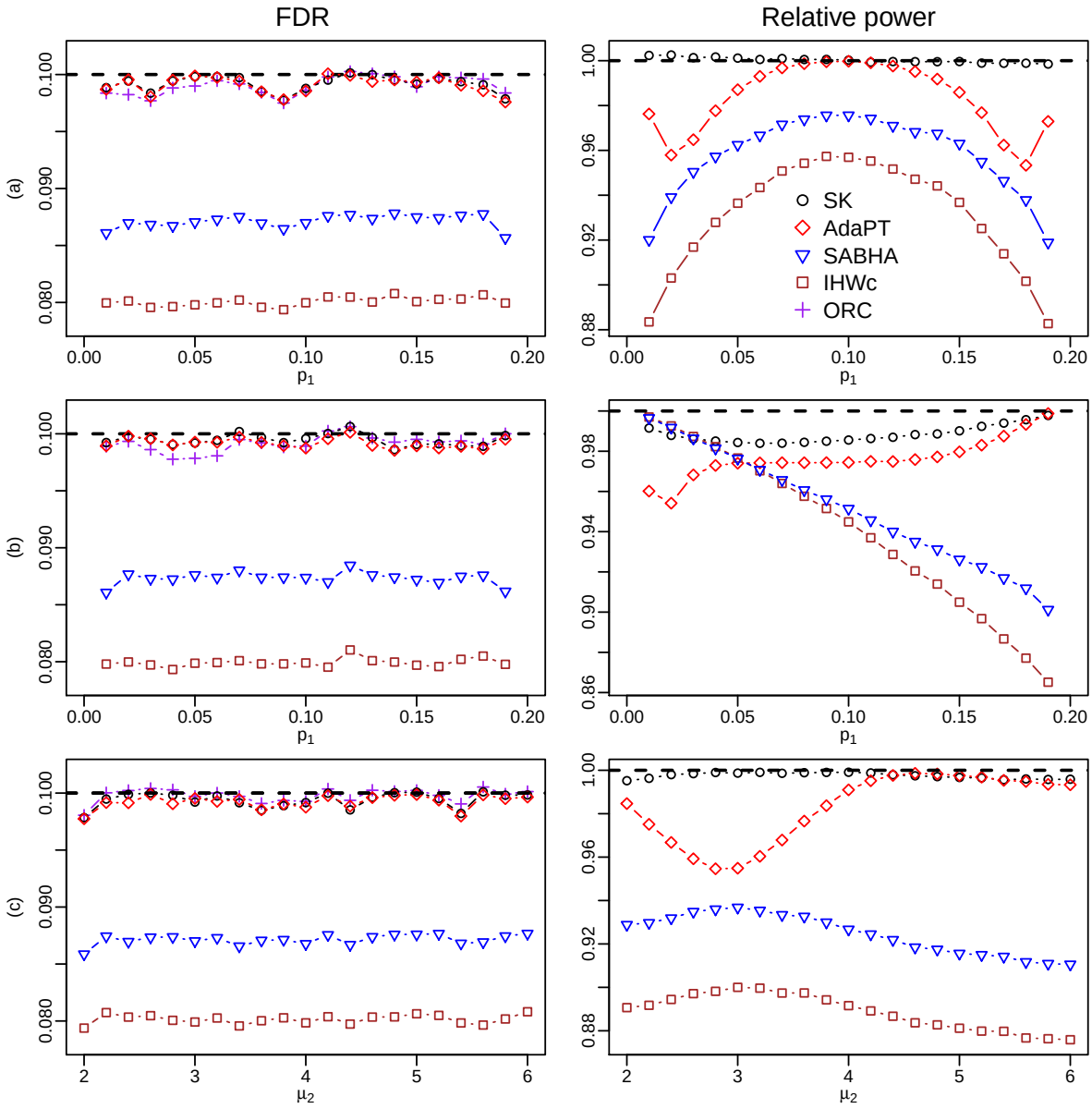

Figure 3.1: Realized FDR levels and relative power in the setting of independent normal statistics in Section 3.4.2. The left column shows the realized FDR levels, and the right column shows the relative power. Rows (a)–(c) show the results in cases (a)–(c), respectively.

3.4.3 Dependent t-statistics

In this subsection, we evaluate the performance of candidate procedures with a dependent structure that mimics the microarray gene expression data. The simulation setup is similar to the dependence simulation setting in Liang (2016), and the sample size in the t-test here is the same as the sample size in the first real data application in Section 3.5. More specifically, we randomly divide the null hypotheses into blocks of size $b = 20$. Within each block, the gene expression levels of 6 subjects are generated independently according to $MVN(\underline{0}_b, \Sigma_{b\times b})$, where $\Sigma_{b\times b} = (\sigma_{ij})_{b\times b} = (\rho^{|i-j|})_{b\times b}$ with $\rho = -.7$. That is, we consider an autoregressive order one (AR1) dependence structure within each block to mimic the mixing of positive and negative correlations among genes in the same biological pathway. Among the 6 subjects, 3 subjects are designated as the Treatment group, and the other 3 subjects form the Control group. The DE genes have a probability of p_1 to be down-regulated in the Treatment group subjects, and we add a treatment effect $\mu_1 < 0$ to their gene expression levels. Similarly, the DE genes have a probability of p_2 to be up-regulated in the Treatment group subjects, and we add a treatment effect $\mu_2 > 0$ to their gene expression levels. The test statistics $t_1, \ldots, t_n$ were computed by the regular two-sample t-test. Then we study three identical cases of parameter variations as in Section 3.4.2.

The average realized FDR and relative power levels are shown in the left column and the right column in Figure 3.2, respectively. Across all three cases, there is no evidence suggesting that any procedure exceeds the target FDR level. `SK` is at least as powerful as `AdaPT` in all settings and is significantly more powerful than `AdaPT` in many settings. When $p_1 \leq 0.08$ in case (b), `SABHA` and `IHWc` show a minor power advantage over `SK`. In other settings, `SK` is convincingly more powerful than `SABHA` and `IHWc`, sometimes almost twice or more as powerful.

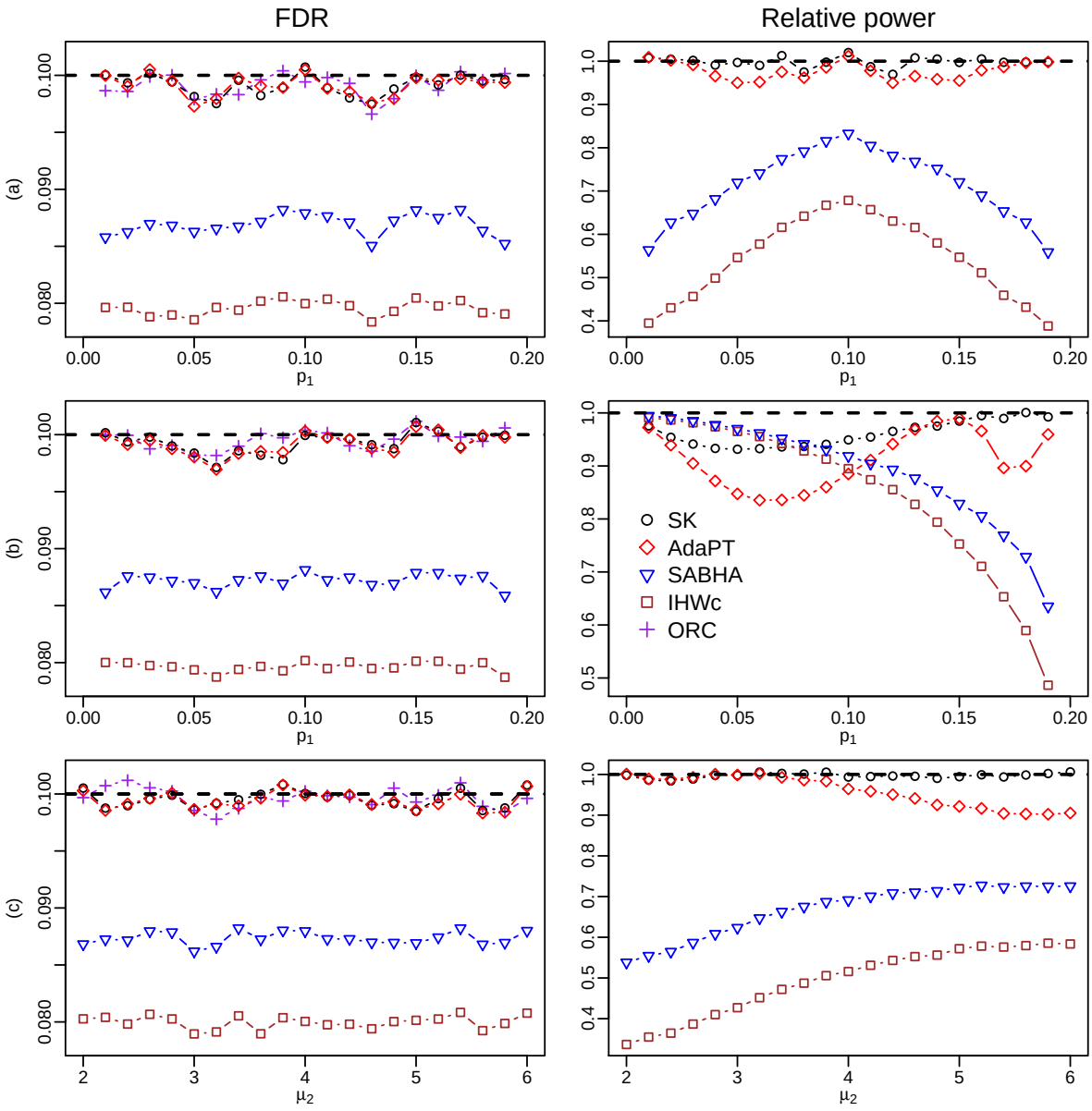

Figure 3.2: Realized FDR levels and relative power in the setting of dependent t-statistics in Section 3.4.3. The left column shows the realized FDR levels, and the right column shows the relative power. Rows (a)–(c) show the results in cases (a)–(c), respectively.

We briefly summarize the simulation results under both independence and dependence settings. The performance of SK is close to the oracle method ORC, and SK offers significant power improvement over other competing methods. We also considered simulation settings with weak signal strength, and the results are presented in the second part of Section 3.6.4.

3.5 Real data applications

We now illustrate the power advantage of our method through two real data applications. The test statistic distribution in the first application is strongly asymmetric while the distribution in the second application, which we describe in Example 1.2, is roughly symmetric about zero.

In addition to the methods compared in Section 3.4, we also considered a p-value-based procedure that does not utilize the directional information: RB, the right-boundary procedure proposed by Liang and Nettleton (2012), which is shown to control the FDR in finite samples (MacDonald, Liang, and Janssen, 2019). Simulation results suggest that RB is one of the most powerful p-value-based procedures. For space consideration, we choose not to present the results of SABHA and IHWc here because RB has more rejections than them in almost every setting, and on both of the positive side and the negative side. Thus, we compare SK, AdaPT, and RB in the applications. We let the nominal FDR level α vary from 0.01 to 0.2, and focus on the result when $\alpha = 0.1$.

3.5.1 Thale cress seedlings

Jang et al. (2014) conducted a microarray study to identify DE genes between two genotypes, the wild-type and the mutant type. The expression levels of $n = 22810$ genes are

compared between the two genotypes, using 3 samples of tissues from each genotype. We compute the regular two-sample t-statistics of all genes as $t_1, \ldots, t_n$. The corresponding p-values are computed as $p_i = 2\{1 - F(|t_i|)\}$, $i = 1, \ldots, n$, where F is the cdf of the t distribution with 4 degrees of freedom.

Figure 3.3(a) shows the histogram of test statistics as well as the rejection boundaries of all procedures considered when the target FDR level $\alpha = 0.1$. The t-statistic distribution is highly asymmetric. There are 646 t-statistics that are less than -5, while there are only 378 t-statistics that are greater than 5. We use the Kolmogorov–Smirnov test to test for any distributional difference between the positive and negative t-statistics, and the p-value is almost zero ($< 2.2\text{e}{-16}$). As a p-value-based method, `RB` cannot utilize the directional information and reports a symmetric rejection region about 0. On the other hand, the direction-adaptive procedures `SK` and `AdaPT` can use asymmetric rejection boundaries to gain power. As there seem to be more strong negative signals than the positive ones, `SK` and `AdaPT` shift the rejection boundaries towards the positive direction to allow more rejections on the negative side.

The difference between `SK` and `AdaPT` can be explained by their different modeling approaches. Recall that `SK` only needs to estimate one true null proportion. By using the signs of test statistics as a covariate, the model used in `AdaPT` estimates the true null proportion separately for negative and positive statistics. In the thale cress dataset, the `AdaPT` true null proportion estimate is larger on the positive side than the negative side, and `AdaPT`'s rejection boundaries are pushed further towards the positive side than `SK` as a result. Overall, `SK` has more rejections than `AdaPT` due to `SK`'s modeling advantage.

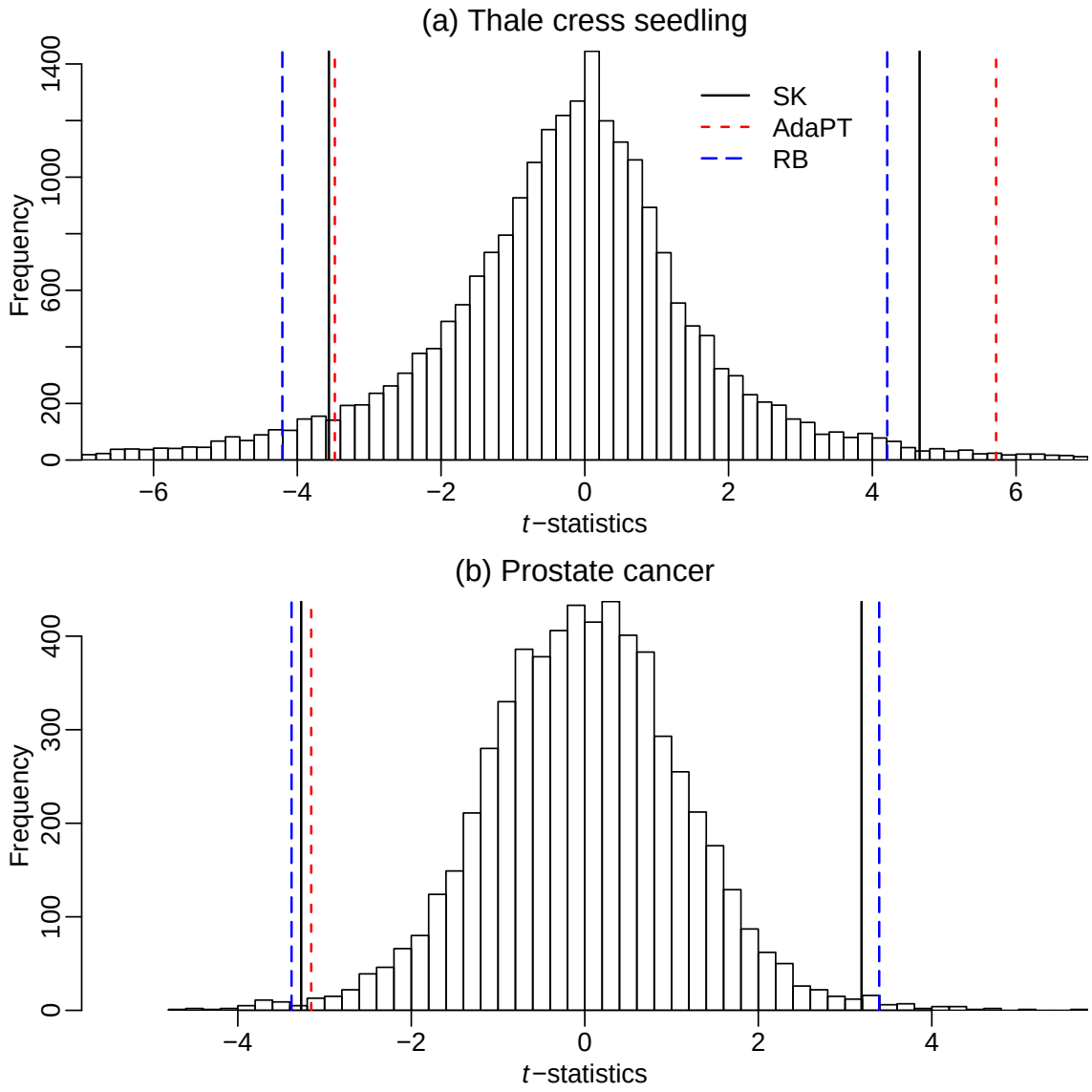

Figure 3.3: Histogram of t-statistics and boundaries of rejection regions given by different procedures when the nominal FDR level $\alpha = 0.1$. Panel (a) shows test statistics from the thale cress seedling data analysis, and Panel (b) shows test statistics from the prostate cancer data analysis. As `AdaPT` has no positive rejections in prostate cancer data analysis, there is no right rejection boundary of `AdaPT` in Panel (b).

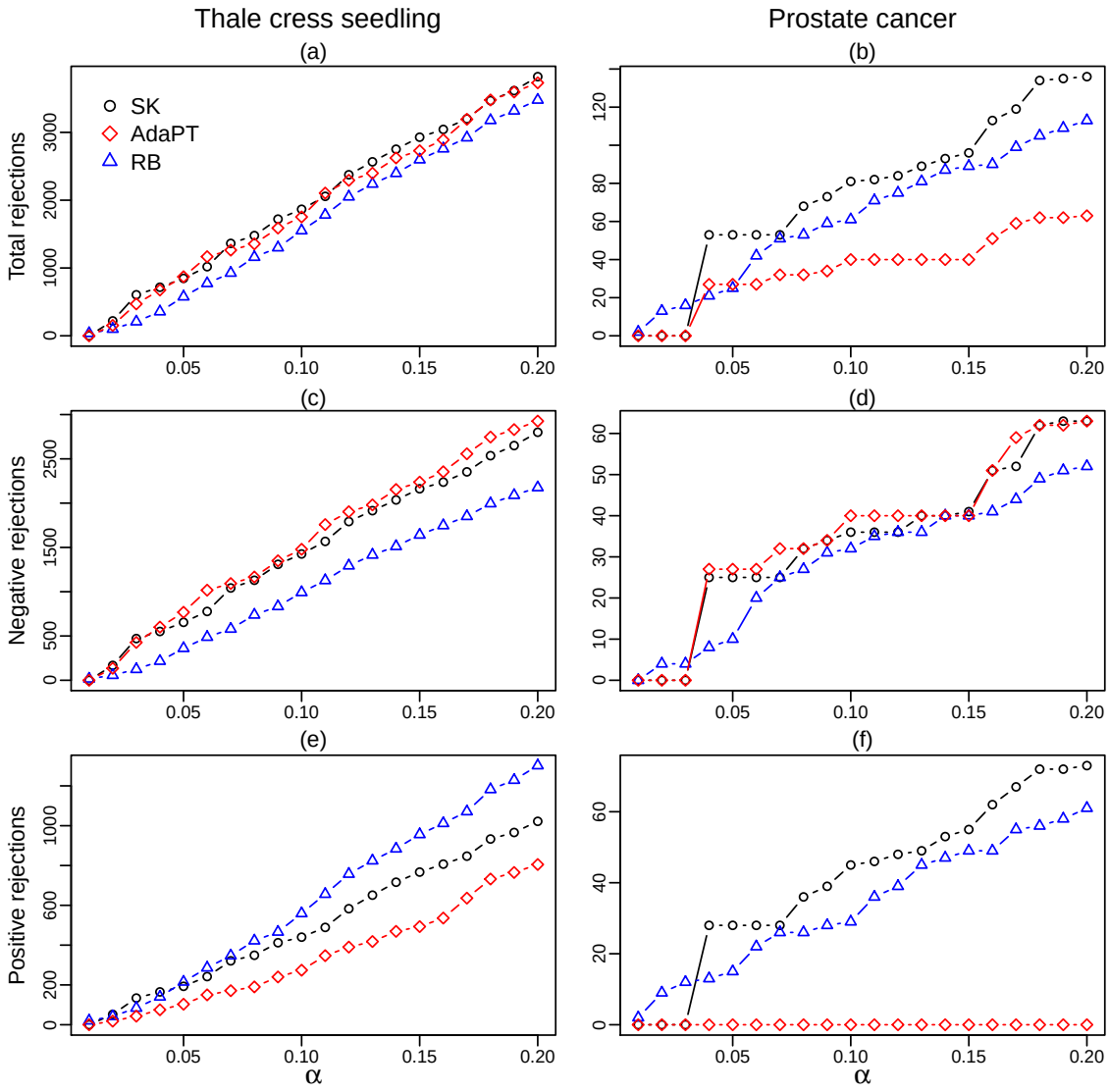

Figure 3.4: Plots of numbers of rejections versus the nominal FDR level α. Panels (a), (c) and (e) show numbers of rejections in thale cress seedling data analysis, and Panels (b), (d) and (f) show numbers of rejections in prostate cancer data analysis. The first row shows the total numbers of rejections, the second row shows the numbers of rejections on the negative side, and the last row shows the numbers of rejections on the positive side.

3.5.2 Prostate cancer

Now we apply the candidate methods to Example 1.2. Figure 3.3(b) shows the histogram of test statistics as well as the rejection boundaries of all procedures considered when the nominal FDR level $\alpha = 0.1$. There is no apparent distributional asymmetry in Figure 3.3(b), and the p-value from the Kolmogorov–Smirnov test is about 0.4. As a result, the rejection region from `SK` is reasonably symmetric about zero. However, `AdaPT` shows no positive rejection boundary because of the lack of positive rejections. Panels in the right column of Figure 3.4 show the relationship between the numbers of rejections and the nominal FDR levels. For reasonably large α, more specifically, when $\alpha > 0.03$, `SK` has the most overall rejections for all α values. Remarkably, `SK` can have equal or better power than the p-value-based procedure `RB` even when the directional information is not strong. It is interesting that `AdaPT` makes no positive rejections at all FDR levels, despite the evidence of the overall test statistic symmetry. Furthermore, `AdaPT` has fewer total rejections than `RB`, which is not direction-adaptive, at almost all target FDR levels. So we investigate the `AdaPT`'s fit to the data in detail.

Both `SK` and `AdaPT` use the knockoff framework and accept one hypothesis at a time from either the positive side or the negative side based on the local FDR comparison. However, their modeling approaches differ in that `SK` only has one true null proportion to estimate while for `AdaPT`, the model suggested in Lei and Fithian (2018) estimates the true null proportion separately for each side. As discussed at the end of Section 3.3, `SK` is robust to model misspecification because its true null proportion estimate is not needed in the local FDR comparison. On the other hand, the local FDR estimation method in `AdaPT` is sensitive to the correct model specification and reasonable true null or alternative proportion estimates. Using the prostate dataset as an example, the `AdaPT` estimate of

the alternative proportion on the positive side is about 1.5e−5, i.e., there is almost no positive signal. On the negative side, the alternative proportion is estimated to be about 1.9e−3. Although the absolute difference between the two alternative proportion estimates is small, the ratio between the two is about 124. It is straightforward to show that, when the true null proportion estimates are close to 1, the ratio between the alternative proportion estimates becomes a crucial factor for `AdaPT` to decide from which side or group to accept the next hypothesis. Thus, `AdaPT` highly prefers to accept null hypotheses on the positive side to the extent that it only rejects on the negative side. However, from the histogram in Figure 3.3(b), we see that the positive t-statistics has a slightly longer tail than that of the negative ones, indicating stronger effect sizes of up-regulated genes. The smallest p-value associated with a positive t-statistic is about 1.5e−7, and we would have rejected its corresponding null hypothesis at a stringent FWER of 0.001. Thus, `AdaPT`'s ad-hoc way of dealing with directional information can lead to not only power loss but also unreasonable rejection decisions.

3.6 Technical details and additional simulation results

This section is organized as follows. Section 3.6.1 includes the proof of Theorem 3.1. Section 3.6.2 describes the EM-algorithm, which is introduced in Section 3.3. Section 3.6.3 discusses the similarities and differences between our method and the AdaPT method (Lei and Fithian, 2018). Section 3.6.4 presents additional simulation results.

3.6.1 Proof of Theorem 3.1

Proof: Since the signed p-values q_i are determined by their corresponding test statistics, these values under the true null hypotheses are also mutually independent of each other. Furthermore, because the signed p-values are continuous, the probability of a tie or of a value equal to 0, $-1/2$, or $1/2$ is zero. Hence, we assume that there are no ties and none of the signed p-values are equal to 0, $-1/2$, or $1/2$. Recall that the signed-knockoff procedure terminates when

$$\widehat{\mathrm{FDR}}_k = \frac{1 + \#\{i:\ \tilde{q}_i \in R_k\}}{\#\{i:\ q_i \in R_k\} \vee 1} \leq \alpha \tag{3.2}$$

or all the null hypotheses are accepted. In the latter case, the FDP will be 0. It therefore suffices to prove that

$$\mathrm{E}\left(\mathrm{FDP} \middle| \widehat{\mathrm{FDR}}_k \leq \alpha\right) \leq \alpha.$$

Recall that $\mathcal{H}_0 = \{i:\ H_{0i} \text{ is true}\}$ is the true null index set. Let i^*, j^*, and k^* be the values of i_k, j_k, and k when the procedure terminates. Further, define

$$\begin{aligned}
v_k &= \#(\mathcal{H}_0 \bigcap \{i:\ q_i \in R_k\}),\\
\tilde{v}_k &= \#(\mathcal{H}_0 \bigcap \{i:\ \tilde{q}_i \in R_k\}),\\
\mathcal{G}_k &= \sigma\left(\mathcal{F}_k, \{I(i \in \mathcal{H}_0)\}_{i=1,\ldots,n}, \sum_{i \in \mathcal{H}_0} b_i, \sum_{i \notin \mathcal{H}_0} b_i\right).
\end{aligned}$$

In other words, after k shrinkage steps, v_k is the number of true null signed p-values in the rejection region, and $\tilde{v}_k$ is the number of true null knockoffs in the rejection region. $\mathcal{G}_k$ is a filtration containing the information in $\mathcal{F}_k$, the true null indicators, and the sums of the b_i's over the true null and alternative hypotheses, respectively. We need the following two lemmas:

Lemma 3.1. *k^* is a stopping time with respect to $\{\mathcal{G}_k\}_{k \geq 0}$.*

Lemma 3.2. *$\{v_k/(1+\tilde{v}_k)\}_{k\geq 0}$ is a supermartingale with respect to $\{\mathcal{G}_k\}_{k\geq 0}$.*

Lemma 3.1 implies that $\mathcal{G}_k$ is rich enough to determine whether the signed-knockoff procedure terminates after k shrinkage steps. Lemma 3.2 is crucial for our proof of Theorem 3.1, and its proof is similar in spirit to the proof of Lemma 3.1 in the Supplement to Barber and Candès (2015). The technical proofs of Lemmas 3.1 and 3.2 are presented in the end of the section.

Now we return to the proof of Theorem 3.1. Note that

$$\begin{aligned}\mathrm{FDR} &= \mathrm{E}\left(\frac{v_{k^*}}{\#\{i:\ q_i \in R_{k^*}\}\vee 1}\right)\\ &\leq \mathrm{E}\left(\frac{1+\#\{i:\ \tilde{q}_i \in R_{k^*}\}}{\#\{i:\ q_i \in R_{k^*}\}\vee 1}\times\frac{v_{k^*}}{1+\tilde{v}_{k^*}}\right),\end{aligned}$$

where in the last step, we use the fact that

$$\tilde{v}_{k^*} \leq \#\{i:\ \tilde{q}_i \in R_{k^*}\}.$$

According to the stopping condition (3.2), we have

$$\frac{1+\#\{i:\ \tilde{q}_i \in R_{k^*}\}}{\#\{i:\ q_i \in R_{k^*}\}\vee 1} \leq \alpha.$$

Hence,

$$\mathrm{FDR} \leq \alpha \mathrm{E}\left(\frac{v_{k^*}}{1+\tilde{v}_{k^*}}\right)$$

and it suffices to show that

$$\mathrm{E}\left(\frac{v_{k^*}}{1+\tilde{v}_{k^*}}\right) \leq 1. \tag{3.3}$$

In the proof of (3.3), a key step is to apply the optional stopping theorem. To do so, we need to verify that its conditions are satisfied:

(a) k^* is a stopping time with respect to $\{\mathcal{G}_k\}_{k\geq 0}$;

(b) k^* is bounded almost surely (a.s.);

(c) $\{v_k/(1+\tilde{v}_k)\}_{k\geq 0}$ is a supermartingale with respect to $\{\mathcal{G}_k\}_{k\geq 0}$.

Conditions (a) and (c) are proved in Lemmas 3.1 and 3.2. As for (b), it is obvious that $k^* \leq n$ and hence it is bounded. Then, according to the optional stopping theorem, we have

$$\mathrm{E}\left(\frac{v_{k^*}}{1+\tilde{v}_{k^*}}\right) \leq \mathrm{E}\left(\frac{v_0}{1+\tilde{v}_0}\right).$$

In the last step of the proof, we show that

$$\mathrm{E}\left(\frac{v_0}{1+\tilde{v}_0}\right) \leq 1. \tag{3.4}$$

Let n_0 be the number of elements in $\mathcal{H}_0 \bigcap \mathcal{J}_0$, where $\mathcal{J}_0$ is the index set of masked hypotheses at initialization. Then $v_0+\tilde{v}_0 = n_0$. Since the p-values under true null hypotheses follow Uniform$(0,1)$, we have $\mathrm{P}(b_i = 1) = 1/2$ for any $i \in \mathcal{H}_0$. Furthermore, $\{b_i\}_{i\in\mathcal{H}_0}$ are independent of each other, so $v_0 = \sum_{i\in\mathcal{H}_0\bigcap\mathcal{J}_0} b_i$ follows the binomial distribution Bin$(n_0, 1/2)$. Finally,

$$\begin{aligned}
\mathrm{E}\left(\frac{v_0}{1+\tilde{v}_0}\Big|n_0\right) &= \mathrm{E}\left(\frac{v_0}{1+n_0-v_0}\Big|n_0\right) \\
&= \sum_{i=0}^{n_0} \frac{i}{1+n_0-i}\binom{n_0}{i}\left(\frac{1}{2}\right)^{n_0} \\
&= \sum_{i=1}^{n_0} \binom{n_0}{i-1}\left(\frac{1}{2}\right)^{n_0} \\
&= 1-\left(\frac{1}{2}\right)^{n_0} \\
&\leq 1,
\end{aligned}$$

which implies (3.4). This completes the proof of Theorem 3.1.

Proof of Lemma 3.1

Proof: It suffices to show that $\widehat{\text{FDR}}_k \in \mathcal{G}_k$ for each $k \geq 1$. We show that both the denominator and numerator of $\widehat{\text{FDR}}_k$ belong to $\mathcal{G}_k$.

For the denominator of $\widehat{\text{FDR}}_k$,

$$\#\{i : \ q_i \in R_k\} = \sum_{i \in \mathcal{H}_0} b_i + \sum_{i \notin \mathcal{H}_0} b_i - \sum_{i \in \mathcal{I}_k} b_i.$$

Because $b_i \in \mathcal{F}_k$ for any $i \in \mathcal{I}_k$, we have $\#\{i : \ q_i \in R_k\} \in \mathcal{G}_k$ and therefore

$$\#\{i : \ q_i \in R_k\} \vee 1 \in \mathcal{G}_k.$$

Similarly, for the numerator,

$$\#\{i : \ \tilde{q}_i \in R_k\} = \#\mathcal{J}_k + \sum_{i \in \mathcal{I}_k} b_i - \sum_{i \in \mathcal{H}_0} b_i - \sum_{i \notin \mathcal{H}_0} b_i.$$

Because $\mathcal{I}_k \in \mathcal{F}_k$ and $\mathcal{J}_k$ is the complement set of $\mathcal{I}_k$, we have $\mathcal{J}_k \in \mathcal{F}_k$. Therefore, $\#\{i : \ \tilde{q}_i \in R_k\} \in \mathcal{G}_k$. Hence,

$$1 + \#\{i : \ \tilde{q}_i \in R_k\} \in \mathcal{G}_k.$$

Therefore, $\widehat{\text{FDR}}_k \in \mathcal{G}_k$. This completes the proof.

Proof of Lemma 3.2

Proof: For simplicity, define

$$M_k = \frac{v_k}{1 + \tilde{v}_k}.$$

We proceed in two steps. First, we prove that

$$M_k \in \mathcal{G}_k.$$

It suffices to show that $v_k \in \mathcal{G}_k$ and $\tilde{v}_k \in \mathcal{G}_k$.

Note that

$$\begin{aligned} v_k =& \#(\mathcal{H}_0 \bigcap \{i : \ q_i \in R_k\}) \\ =& \sum_{i \in \mathcal{H}_0} b_i - \sum_{i \in \mathcal{H}_0 \bigcap \mathcal{I}_k} b_i. \end{aligned}$$

By the definitions of $\mathcal{F}_k$ and $\mathcal{G}_k$, it follows that $v_k \in \mathcal{G}_k$. Similarly,

$$\tilde{v}_k = \#\mathcal{H}_0 - \sum_{i \in \mathcal{H}_0} b_i - \sum_{i \in \mathcal{H}_0 \bigcap \mathcal{I}_k} (1 - b_i),$$

which implies that $\tilde{v}_k \in \mathcal{G}_k$. Hence, $v_k/(1 + \tilde{v}_k) \in \mathcal{G}_k$.

In the second step, we prove that

$$\mathrm{E}\left(M_{k+1} | \mathcal{G}_k\right) \leq M_k.$$

Assume that q_{h_k} is the signed p-value of the revealed hypothesis in the kth shrinkage step. That is, $h_k \notin \mathcal{I}_k$ but $h_k \in \mathcal{I}_{k+1}$. Since the choice in Step 2 is $\mathcal{F}_k$-measurable, $h_k \in \mathcal{F}_k \subseteq \mathcal{G}_k$. Further, note that

$$v_{k+1} = v_k - I(h_k \in \mathcal{H}_0) I(q_{h_k} \in R_k)$$

and

$$\tilde{v}_{k+1} = \tilde{v}_k - I(h_k \in \mathcal{H}_0) I(\tilde{q}_{h_k} \in R_k).$$

If $h_k \notin \mathcal{H}_0$, v_k and $\tilde{v}_k$ will not change, so $M_{k+1} = M_k$.

If $h_k \in \mathcal{H}_0$, $q_{h_k} \in R_k$ is equivalent to $b_{h_k} = 1$, so

$$\mathrm{E}\left\{M_{k+1} | \mathcal{G}_k, h_k \in \mathcal{H}_0\right\} = \mathrm{P}\left(b_{h_k} = 1 | \mathcal{G}_k, h_k \in \mathcal{H}_0\right) \frac{v_k - 1}{1 + \tilde{v}_k} \tag{3.5}$$

$$+ \mathrm{P}\left(b_{h_k} = 0 | \mathcal{G}_k, h_k \in \mathcal{H}_0\right) \frac{v_k}{\tilde{v}_k \vee 1}.$$

To avoid a zero denominator, $\tilde{v}_k \vee 1$ is used in the second term instead of $\tilde{v}_k$.

We need to calculate $\mathrm{P}\left(b_{h_k} = 1 | \mathcal{G}_k, h_k \in \mathcal{H}_0\right)$. Because all true null signed p-values are independently and identically distributed, $\{b_i\}_{i \in \mathcal{H}_0}$ are also identically distributed, so $\mathrm{P}\left(b_i = 1 | \mathcal{G}_k\right)$ is the same for all $i \in \mathcal{H}_0 \bigcap \mathcal{J}_k$. Since $\sum_{i \in \mathcal{H}_0 \bigcap \mathcal{J}_k} b_i = v_k$ and $\#(\mathcal{H}_0 \bigcap \mathcal{J}_k) = v_k + \tilde{v}_k$, it is straightforward to show that

$$\mathrm{P}\left(b_{h_k} = 1 | \mathcal{G}_k, h_k \in \mathcal{H}_0\right) = \frac{v_k}{v_k + \tilde{v}_k}, \tag{3.6}$$

since $h_k \in \mathcal{H}_0 \bigcap \mathcal{J}_k$.

If $\tilde{v}_k \geq 1$, we substitute (3.6) into (3.5) and obtain

$$\mathrm{E}\left(M_{k+1} | \mathcal{G}_k, h_k \in \mathcal{H}_0, \tilde{v}_k \geq 1\right) = \frac{v_k}{v_k + \tilde{v}_k} \frac{v_k - 1}{1 + \tilde{v}_k} + \left(1 - \frac{v_k}{v_k + \tilde{v}_k}\right) \frac{v_k}{\tilde{v}_k} = \frac{v_k}{1 + \tilde{v}_k} = M_k. \tag{3.7}$$

If $\tilde{v}_k = 0$, $\mathrm{P}(b_{h_k} = 1 | \mathcal{G}_k, h_k \in \mathcal{H}_0, \tilde{v}_k = 0) = 1$, so

$$\mathrm{E}\left(M_{k+1} | \mathcal{G}_k, h_k \in \mathcal{H}_0\right) = \frac{v_k - 1}{1 + \tilde{v}_k} \leq \frac{v_k}{1 + \tilde{v}_k} = M_k. \tag{3.8}$$

Combining (3.7), (3.8), and the fact that $M_{k+1} = M_k$ if $h_k \notin \mathcal{H}_0$, we have

$$\mathrm{E}\left(M_{k+1} | \mathcal{G}_k\right) \leq M_k.$$

Finally, $v_k/(1 + \tilde{v}_k)$ is bounded between 0 and n. By combining the results above, we complete the proof.

3.6.2 Details of the EM-algorithm

Suppose that k shrinkage steps of our procedure have been performed. Recall that $\mathcal{I}_k$ and $\mathcal{J}_k$ are the index sets of the revealed and masked hypotheses, respectively. Based on

the parametric mixture of beta distributions that we use to model the alternative signed p-values, we rewrite the log-likelihood as:

$$\begin{aligned} l(\pi_0, \alpha, \beta, \lambda) = & \sum_{i \in \mathcal{I}_k} \log \left\{ \frac{\pi_0}{2} + (1-\pi_0)\lambda \frac{\alpha}{2} \left(\frac{q_i + 1}{2} \right)^{\alpha - 1} + (1-\pi_0)(1-\lambda) \frac{\beta}{2} \left(\frac{1 - q_i}{2} \right)^{\beta - 1} \right\} \\ & + \sum_{i \in \mathcal{J}_k} \log \left\{ \pi_0 + (1-\pi_0)\lambda \frac{\alpha}{2} \left(\frac{q_i + 1}{2} \right)^{\alpha - 1} + (1-\pi_0)(1-\lambda) \frac{\beta}{2} \left(\frac{1 - q_i}{2} \right)^{\beta - 1} \right. \\ & \left. + (1-\pi_0)\lambda \frac{\alpha}{2} \left(\frac{\tilde{q}_i + 1}{2} \right)^{\alpha - 1} + (1-\pi_0)(1-\lambda) \frac{\beta}{2} \left(\frac{1 - \tilde{q}_i}{2} \right)^{\beta - 1} \right\}. \end{aligned}$$

Let $H_i = 1$ if the ith null hypothesis is false and $H_i = 0$ otherwise. If $H_i = 1$, let $\gamma_i = 1$ if q_i is from the negative signal component with the pdf $\alpha/2 \cdot ((q_i + 1)/2)^{\alpha - 1}$, and $\gamma_i = 0$ if q_i is from the positive signal component. Let $I_i = 1$ if $q_i > \tilde{q}_i$, and $I_i = 0$ otherwise. Note that some of the I_i's are observed. The log-likelihood function of the complete data is

$$\begin{aligned} l^c(\pi_0, \alpha, \beta, \lambda) = & \sum_{i=1}^{n} H_i \log(1 - \pi_0) + (1 - H_i) \log \pi_0 \\ & + \sum_{i=1}^{n} H_i \gamma_i \log \lambda + H_i (1 - \gamma_i) \log(1 - \lambda) \\ & + \sum_{i=1}^{n} H_i \gamma_i I_i \left\{ \log(\alpha/2) + (\alpha - 1) \log \left(\frac{\tilde{q}_i^* + 1}{2} \right) \right\} \\ & + \sum_{i=1}^{n} H_i \gamma_i (1 - I_i) \left\{ \log(\alpha/2) + (\alpha - 1) \log \left(\frac{q_i^* + 1}{2} \right) \right\} \\ & + \sum_{i=1}^{n} H_i (1 - \gamma_i) I_i \left\{ \log(\beta/2) + (\beta - 1) \log \left(\frac{1 - \tilde{q}_i^*}{2} \right) \right\} \\ & + \sum_{i=1}^{n} H_i (1 - \gamma_i)(1 - I_i) \left\{ \log(\beta/2) + (\beta - 1) \log \left(\frac{1 - q_i^*}{2} \right) \right\}. \end{aligned}$$

The EM-algorithm alternates between an E-step and an M-step. Let $\Theta^{(r)} = (\pi_0^{(r)}, \alpha^{(r)}, \beta^{(r)}, \lambda^{(r)})$ be the parameter estimates at the rth EM iteration.

Suppose we are at the $(r+1)$th iteration of the EM-algorithm. In the E-step, we need to compute the following quantities:

$$\begin{aligned}
&\mathrm{P}(H_i = 0|\Theta^{(r)});\\
&\mathrm{P}(\gamma_i = 1, H_i = 1|\Theta^{(r)});\\
&\mathrm{P}(I_i = 1, \gamma_i = 1, H_i = 1|\Theta^{(r)}), \text{ if } i \in \mathcal{J}_k;\\
&\mathrm{P}(I_i = 1, \gamma_i = 0, H_i = 1|\Theta^{(r)}), \text{ if } i \in \mathcal{J}_k.
\end{aligned}$$

For convenience, define

$$f_1^{(r)}(q) = \lambda^{(r)} \frac{\alpha^{(r)}}{2} \left(\frac{q+1}{2}\right)^{\alpha^{(r)}-1} + (1-\lambda^{(r)}) \frac{\beta^{(r)}}{2} \left(\frac{1-q}{2}\right)^{\beta^{(r)}-1}$$

as the alternative density function evaluated at q using $\Theta^{(r)}$.

If $i \in \mathcal{I}_k$, q_i is observed, so

$$\begin{aligned}
\mathrm{P}(H_i = 0|\Theta^{(r)}) &= \frac{\frac{1}{2}\pi_0^{(r)}}{\frac{1}{2}\pi_0^{(r)} + (1-\pi_0^{(r)}) f_1^{(r)}(q_i)};\\
\mathrm{P}(\gamma_i = 1, H_i = 1|\Theta^{(r)}) &= \frac{(1-\pi_0^{(r)})\lambda^{(r)} \frac{\alpha^{(r)}}{2} \left(\frac{q_i+1}{2}\right)^{\alpha^{(r)}-1}}{\frac{1}{2}\pi_0^{(r)} + (1-\pi_0^{(r)}) f_1^{(r)}(q_i)}.
\end{aligned}$$

If $i \in \mathcal{J}_k$, only q_i^* is observed, so

$$\begin{aligned}
\mathrm{P}(I_i = 1|H_i = 1, \gamma_i = 1, \Theta^{(r)}) &= \frac{(\tilde{q}_i^* + 1)^{\alpha^{(r)}-1}}{(\tilde{q}_i^* + 1)^{\alpha^{(r)}-1} + (q_i^* + 1)^{\alpha^{(r)}-1}};\\
\mathrm{P}(I_i = 1|H_i = 1, \gamma_i = 0, \Theta^{(r)}) &= \frac{(1-\tilde{q}_i^*)^{\beta^{(r)}-1}}{(1-\tilde{q}_i^*)^{\beta^{(r)}-1} + (1-q_i^*)^{\beta^{(r)}-1}};\\
\mathrm{P}(H_i = 0|\Theta^{(r)}) &= \frac{\pi_0^{(r)}}{\pi_0^{(r)} + (1-\pi_0^{(r)}) \left(f_1^{(r)}(q_i^*) + f_1^{(r)}(\tilde{q}_i^*)\right)};\\
\mathrm{P}(\gamma_i = 1|H_i = 1, \Theta^{(r)}) &= \frac{\lambda^{(r)} \frac{\alpha^{(r)}}{2} \left\{\left(\frac{q_i+1}{2}\right)^{\alpha^{(r)}-1} + \left(\frac{\tilde{q}_i+1}{2}\right)^{\alpha^{(r)}-1}\right\}}{f_1^{(r)}(q_i^*) + f_1^{(r)}(\tilde{q}_i^*)}.
\end{aligned}$$

The required quantities can be easily computed from the values above.

In the M-step, we maximize the following expectation:

$$
\begin{aligned}
&\mathrm{E}\left\{l^c(\pi_0,\alpha,\beta,\lambda)|\Theta^{(r)}\right\}\\
=&\sum_{i=1}^{n}\mathrm{P}(H_i=0|\Theta^{(r)})\log\pi_0+\left\{n-\sum_{i=1}^{n}\mathrm{P}(H_i=0|\Theta^{(r)})\right\}\log(1-\pi_0)\\
&+\sum_{i=1}^{n}\mathrm{P}(\gamma_i=1,H_i=1|\Theta^{(r)})\log\lambda+\sum_{i=1}^{n}\mathrm{P}(\gamma_i=0,H_i=1|\Theta^{(r)})\log(1-\lambda)\\
&+\sum_{i\in\mathcal{I}_k}\mathrm{P}(\gamma_i=1,H_i=1|\Theta^{(r)})\left\{\log(\alpha/2)+(\alpha-1)\log\left(\frac{q_i+1}{2}\right)\right\}\\
&+\sum_{i\in\mathcal{I}_k}\mathrm{P}(\gamma_i=0,H_i=1|\Theta^{(r)})\left\{\log(\beta/2)+(\beta-1)\log\left(\frac{1-q_i}{2}\right)\right\}\\
&+\sum_{i\in\mathcal{J}_k}\mathrm{P}(I_i=1,\gamma_i=1,H_i=1|\Theta^{(r)})\left\{\log(\alpha/2)+(\alpha-1)\log\left(\frac{\tilde{q}_i^*+1}{2}\right)\right\}\\
&+\sum_{i\in\mathcal{J}_k}\mathrm{P}(I_i=0,\gamma_i=1,H_i=1|\Theta^{(r)})\left\{\log(\alpha/2)+(\alpha-1)\log\left(\frac{q_i^*+1}{2}\right)\right\}\\
&+\sum_{i\in\mathcal{J}_k}\mathrm{P}(I_i=1,\gamma_i=0,H_i=1|\Theta^{(r)})\left\{\log(\beta/2)+(\beta-1)\log\left(\frac{1-\tilde{q}_i^*}{2}\right)\right\}\\
&+\sum_{i\in\mathcal{J}_k}\mathrm{P}(I_i=0,\gamma_i=0,H_i=1|\Theta^{(r)})\left\{\log(\beta/2)+(\beta-1)\log\left(\frac{1-q_i^*}{2}\right)\right\}.
\end{aligned}
$$

The explicit solutions are:

$$
\begin{aligned}
\pi_0^{(r+1)}=&\frac{1}{n}\sum_{i=1}^{n}\mathrm{P}(H_i=0|\Theta^{(r)});\\
\lambda^{(r+1)}=&\frac{\sum_{i=1}^{n}\mathrm{P}(\gamma_i=1,H_i=1|\Theta^{(r)})}{\sum_{i=1}^{n}\mathrm{P}(H_i=1|\Theta^{(r)})};\\
\alpha^{(r+1)}=&\sum_{i=1}^{n}\mathrm{P}(\gamma_i=1,H_i=1|\Theta^{(r)})\Bigg/\Bigg[\sum_{i\in\mathcal{I}_k}\mathrm{P}(\gamma_i=1,H_i=1|\Theta^{(r)})\log\left(\frac{q_i+1}{2}\right)\\
&+\sum_{i\in\mathcal{J}_k}\left\{\mathrm{P}(I_i=1,\gamma_i=1,H_i=1|\Theta^{(r)})\log\left(\frac{\tilde{q}_i^*+1}{2}\right)\right.
\end{aligned}
$$

$$+ \mathrm{P}(I_i = 0, \gamma_i = 1, H_i = 1 | \Theta^{(r)}) \log\left(\frac{q_i^* + 1}{2}\right) \Big\}\Bigg];$$

$$\beta^{(r+1)} = \sum_{i=1}^{n} \mathrm{P}(\gamma_i = 0, H_i = 1 | \Theta^{(r)}) \Bigg/ \Bigg[\sum_{i \in \mathcal{I}_k} \mathrm{P}(\gamma_i = 0, H_i = 1 | \Theta^{(r)}) \log\left(\frac{1 - q_i}{2}\right)$$
$$+ \sum_{i \in \mathcal{J}_k} \Big\{ \mathrm{P}(I_i = 1, \gamma_i = 0, H_i = 1 | \Theta^{(r)}) \log\left(\frac{1 - \tilde{q}_i^*}{2}\right)$$
$$+ \mathrm{P}(I_i = 0, \gamma_i = 0, H_i = 1 | \Theta^{(r)}) \log\left(\frac{1 - q_i^*}{2}\right) \Big\}\Bigg].$$

We repeat the EM iterations until

$$|l(\pi_0^{(r+1)}, \alpha^{(r+1)}, \beta^{(r+1)}, \lambda^{(r+1)}) - l(\pi_0^{(r)}, \alpha^{(r)}, \beta^{(r)}, \lambda^{(r)})| < \epsilon,$$

and in our implementation we set $\epsilon = 10^{-5}$.

We set the initial values of the parameters in the EM-algorithm as follows: if $k = 0$, we simply set $(\pi_0^{(0)}, \alpha^{(0)}, \beta^{(0)}, \lambda^{(0)}) = (0.5, 0.5, 0.5, 0.5)$; if $k \geq 1$, we set the initial values as the parameter estimates after $k - 1$ shrinkage steps.

3.6.3 Connection between `SK` and `AdaPT`

The `AdaPT` method of Lei and Fithian (2018) can be divided into two parts: a framework that is guaranteed to control the FDR in finite samples and an implementation. In this section, we first show that our method `SK` can be viewed as a special instance of the `AdaPT` framework. Then we describe the modeling difference between `SK` and `AdaPT` in implementation.

SK as a special case of the AdaPT framework

Although we developed SK directly based on the signed p-values, our work can also be viewed as a special instance/implementation of the AdaPT framework with the sign as a covariate.

As described in Section 3.2, our procedure is a stepwise procedure that accepts one null hypothesis at a time. Although the AdaPT framework is introduced as a procedure that updates a covariate-dependent rejection threshold, the discussion section of Lei and Fithian (2018) points out that it is a special case of a procedure that updates the rejection set, or equivalently, accepts hypotheses, in a stepwise fashion. The key to the AdaPT framework's FDR control is the limited information one can use to choose the next hypothesis to accept. For the ith null hypothesis, let t_i be the test statistic and p_i the p-value. At each step, AdaPT selects the next hypothesis to accept based on limited information on the p-values of the accepted hypotheses and the unordered pairs of $\{p_i, 1-p_i\}$ of the unaccepted hypotheses, in addition to the covariates. Recall that the signed p-value $q_i = \text{sign}(t_i)(1-p_i)$ and its knockoff $\tilde{q}_i = \text{sign}(q_i) - q_i$. Similarly to AdaPT, SK selects the next hypothesis to accept based on limited information on the signed p-values of the accepted hypotheses and the unordered pairs of $\{q_i, \tilde{q}_i\}$ of the unaccepted hypotheses. If we define the covariate $x_i = \text{sign}(t_i)$, our signed p-value can be computed from the covariate and p-value as $q_i = x_i(1-p_i)$, and the unordered pair $\{q_i, \tilde{q}_i\}$ can be computed from the covariate and the unordered pair $\{p_i, 1-p_i\}$ as $\{q_i, \tilde{q}_i\} = \{x_i(1-p_i), x_i p_i\}$. It is clear that SK fits into the AdaPT framework and controls the FDR as a result.

Modeling difference between `SK` and `AdaPT`

`SK` and `AdaPT` model the directional data differently. `SK` models the signed p-values as a three-component mixture, including one null distribution component and two alternative distribution components. On the other hand, the `AdaPT` implementation is geared toward generic covariates. If we use the sign information as a covariate in the `AdaPT` implementation, `AdaPT` treats the positive statistics and negative statistics as two separate groups and fits a null+alternative two-component mixture model in each group. This is because `AdaPT` models the true null probability and the parameter of the alternative distribution as separate regression functions of the covariate, and there are only two distinct covariate values (-1 and 1). Thus, `AdaPT` estimates the true null proportion separately for each group, and the two null distributions are unlikely to be compatible.

To illustrate this difference, we first generate a dataset containing the test statistics of the two-sided tests

$$H_0 : \mu = 0 \text{ vs } H_a : \mu \neq 0,$$

where the test statistics follow the normal distribution $\mathrm{N}(\mu, 1)$. For this comparison, we transform `AdaPT`'s fitted models on the p-values into fitted models on the signed p-values. The final fitted models of the signed p-values in the two methods are shown in Figure 3.5.

The models in Figure 3.5 are decomposed into the summation of several components. We can see that in our `SK` model, the density function of the signed p-values contains three components: the null, the positive signal (alternative), and the negative signal. All three components have continuous distributions from -1 to 1.

In contrast, `AdaPT` models the p-values with positive and negative test statistics separately. Within each of the two groups (sides), `AdaPT` has two components: one for the null and one for the alternative. Neither the null nor the alternative distribution is continuous

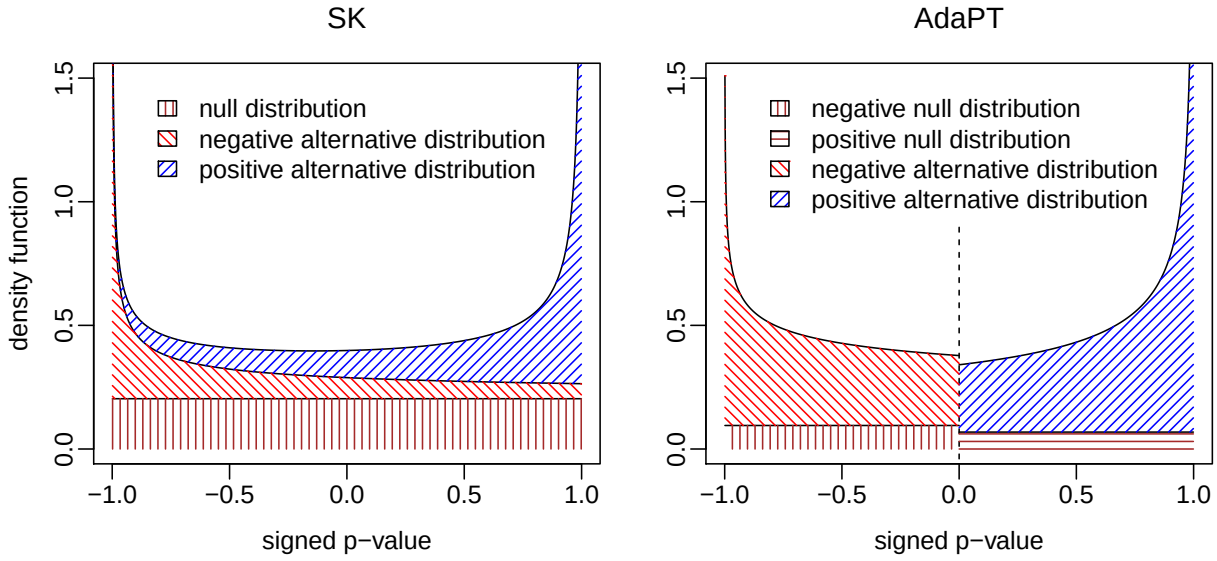

Figure 3.5: Decomposition of signed p-value models from `SK` and `AdaPT`. The left panel shows the fitted model for `SK`, and the right panel shows that for `AdaPT`. Both density functions are decomposed into summations of components in the mixture model in `SK` and `AdaPT`.

over $(-1, 1)$. In applications in which there is only one global null distribution, `AdaPT`'s model with the sign information as a generic covariate may lead to model inconsistency and loss of efficiency, as demonstrated in the simulation and real-data studies in Section 3.4 and Section 3.5.

3.6.4 Additional simulation results

Other multiple testing methods

As well as the procedures considered in Section 3.4, we compared our method `SK` with four other procedures in simulation studies:

- `WBH`, the weighted Benjamini–Hochberg procedure of Zhao and Fung (2016). `WBH` assigns a weight to each hypothesis according to the sign of its corresponding test statistic, and it controls the FDR asymptotically. Like in Zhao and Fung's simulation, we use the least-slope π_0-estimator from Benjamini and Hochberg (2000).

- `SABHA`*, the structure-adaptive Benjamini–Hochberg algorithm of Li and Barber (2019). In Section 3.4, we use an adjusted `SABHA` method so that the FDR is controlled at the nominal level. Here we use the original method proposed in Li and Barber (2019), which has been shown to control the FDR at a level above the nominal level.

- `AdaFDR`, a method that adaptively learns the optimal p-value threshold from the covariates (Zhang, Xia, and Zou, 2019). `AdaFDR` aims to maximize the number of discoveries while controlling the FDP, but it is not guaranteed to control the FDR in finite samples.

- `RB`, the right-boundary procedure of Liang and Nettleton (2012). `RB` has been shown to control the FDR in finite samples under the null independence condition (MacDonald, Liang, and Janssen, 2019). As a p-value-based method, `RB` cannot utilize the directional information, and it is used as a baseline to show the impact of this information.

We use the simulation settings given in Section 3.4. The results for the independent normal statistics setting and the dependent t-statistics setting are shown in Figures 3.6 and 3.7, respectively.

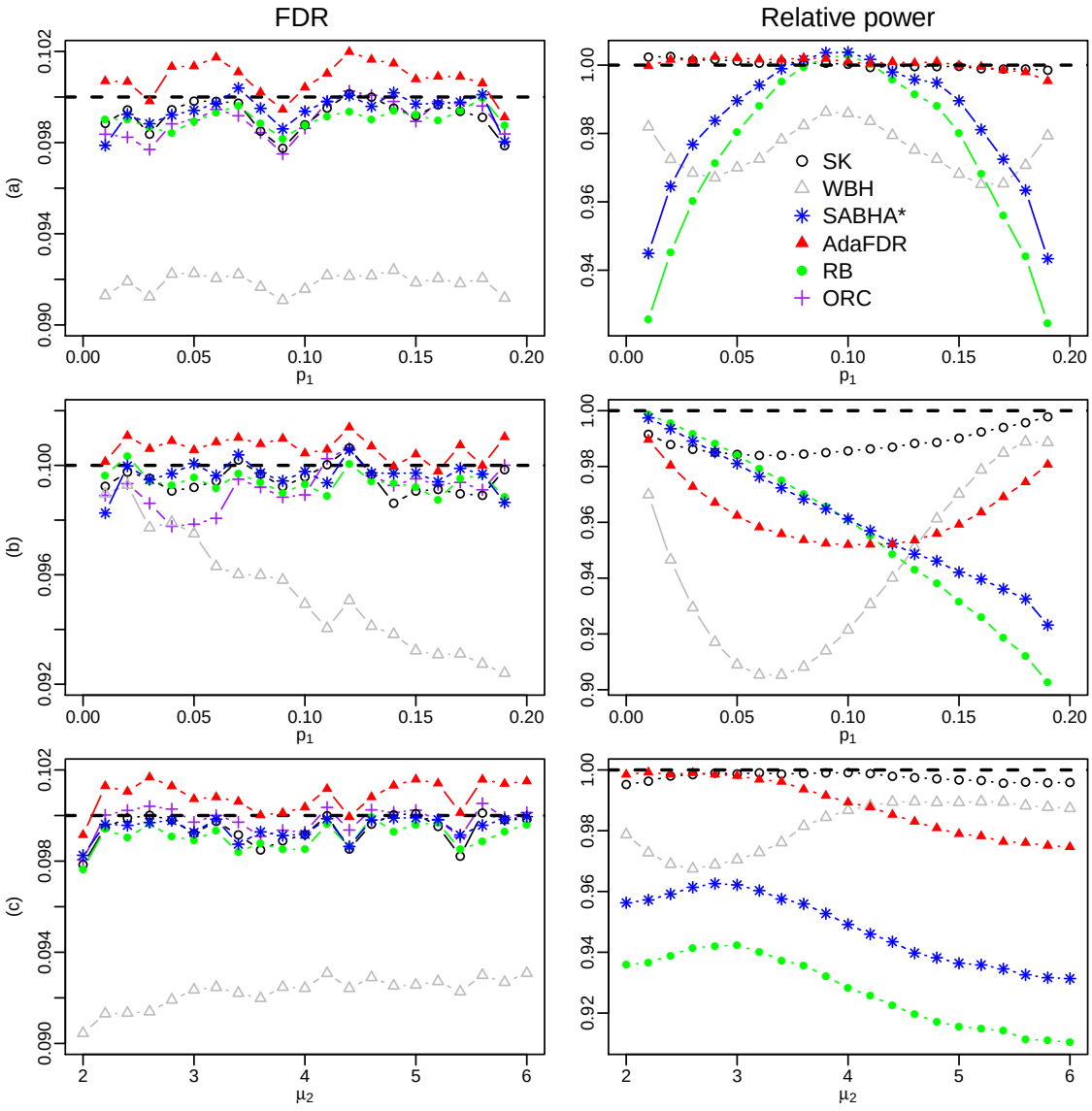

Figure 3.6: Realized FDR levels and relative power in the independent normal statistics setting. The left column shows the realized FDR levels, and the right column shows the realized relative power. Rows (a)–(c) show the results for cases (a)–(c), respectively.

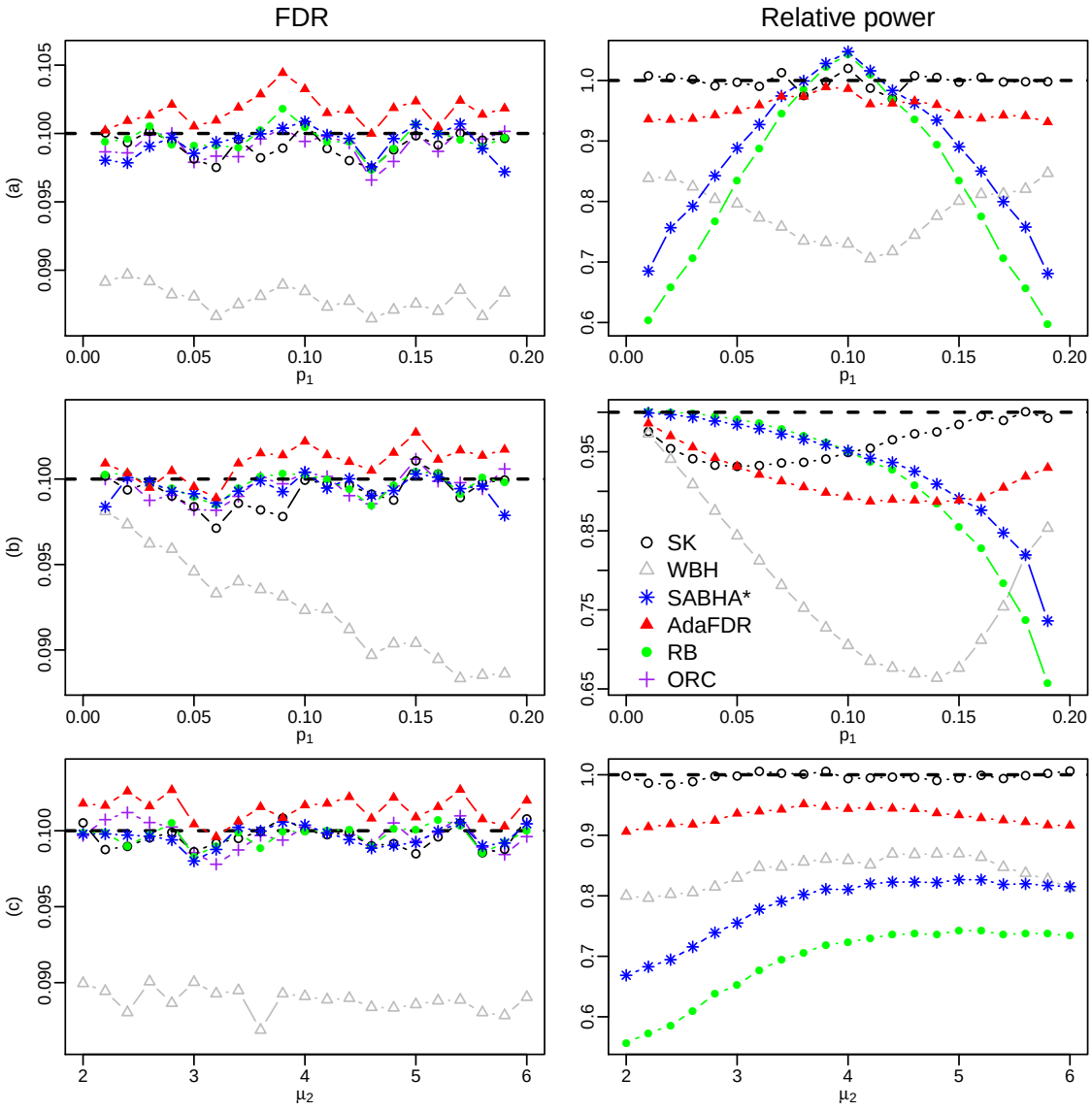

Figure 3.7: Realized FDR levels and relative power in the dependent t-statistics setting. The left column shows the realized FDR levels, and the right column shows the realized relative power. Rows (a)–(c) show the results for cases (a)–(c), respectively.

For the independent normal statistics setting in Section 3.4.2, Figure 3.6 shows that the realized FDR levels are all close to or below the nominal FDR level except that the realized FDR levels for `AdaFDR` are above the nominal level in most settings. For `AdaFDR`, the standard error of the FDP is less than 0.016 for every parameter setting in the independent normal statistics setting. Averaging over 500 replicates reduces the standard deviation to $0.016/\sqrt{500} = 0.0007$. Comparing the standard deviation with the gap between the realized FDR and the nominal level, we conclude that the 95% confidence interval of the realized FDR of `AdaFDR` fails to cover $\alpha = 0.1$ in some of the settings, for example, when $\mu_2 \geq 5.6$ in case (c). In terms of power, overall `SK` dominates the other procedures, and the power improvement is significant for most parameter combinations. Although `SK` has a power performance comparable to that of `AdaFDR` in case (a), its advantage is significant in the other two cases. In case (a) when p_1 is close to 0.1, there is little asymmetry in the alternative distribution, and the power of `SK` is comparable to that of `RB`. When p_1 is away from 0.1, `SK` shows a significant power improvement over `RB`. For the dependent t-statistics setting in Section 3.4.3, we come to a similar conclusion from Figure 3.7: every method except `AdaFDR` has good control of the FDR, and overall `SK` has the best power performance.

Weak signal

We also consider simulations with weaker signal strengths than those presented in Section 3.4. The signal strengths are reduced to 2/3 of the original values. For example, we change the effect size parameter $\mu = \pm 3$ to $\mu = \pm 2$ in case (a). Specifically, the parameter combinations are:

(a) p_1 varies between 0 and 0.2 and $p_2 = 0.2 - p_1$, $\mu_1 = -2$, $\mu_2 = 2$.

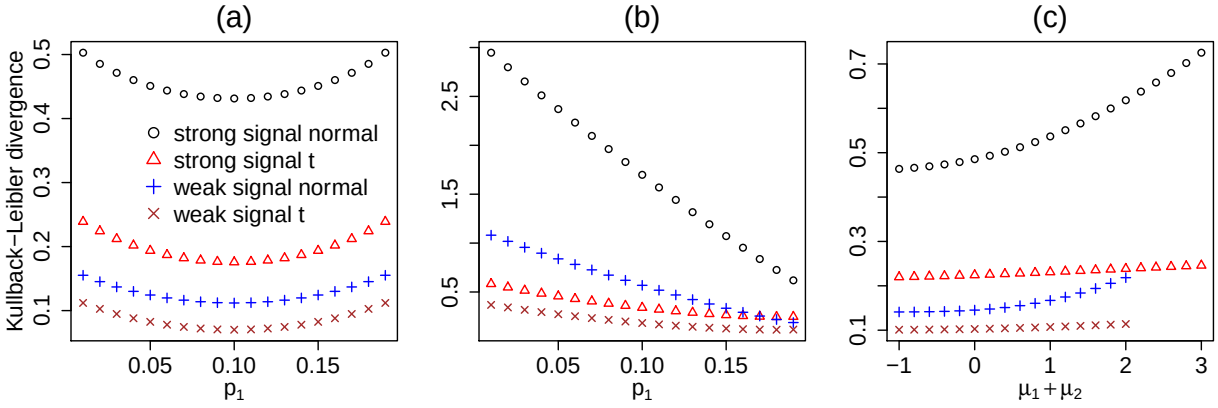

Figure 3.8: The Kullback–Leibler divergence between the null and alternative distributions in the strong signal cases of Section 3.4 and the weak signal cases of Section 3.6.4. In the legends, "normal" indicates the distributions used in the independent normal statistics setting, "t" indicates the distributions used in the dependent t-statistics settings, "strong signal" indicates the setting with strong signals, and "weak signal" indicates the setting with reduced signals.

(b) p_1 varies between 0 and 0.2 and $p_2 = 0.2 - p_1$, $\mu_1 = -2$, $\mu_2 = 4$.

(c) μ_2 varies between 1 and 4 and $p_1 = 0.18$, $p_2 = 0.02$, $\mu_1 = -2$.

To visualize the signal strength, in Figure 3.8 we plot the Kullback–Leibler (K–L) divergence between the null distribution and the alternative distribution in both strong and weak signal settings. Figure 3.8 shows that the K–L divergences between the different settings are well separated. The strong-signal normal setting has the strongest signal, and the signal strength from the weak-signal t setting is much weaker than that for the other settings.

The simulation results are shown in Figures 3.9 and 3.10. There is no strong evidence that any of the methods lose FDR control, except `AdaFDR` in the dependent t-statistics

setting. Overall, the power of SK is still close to that of ORC, and it has a power advantage over the other methods.

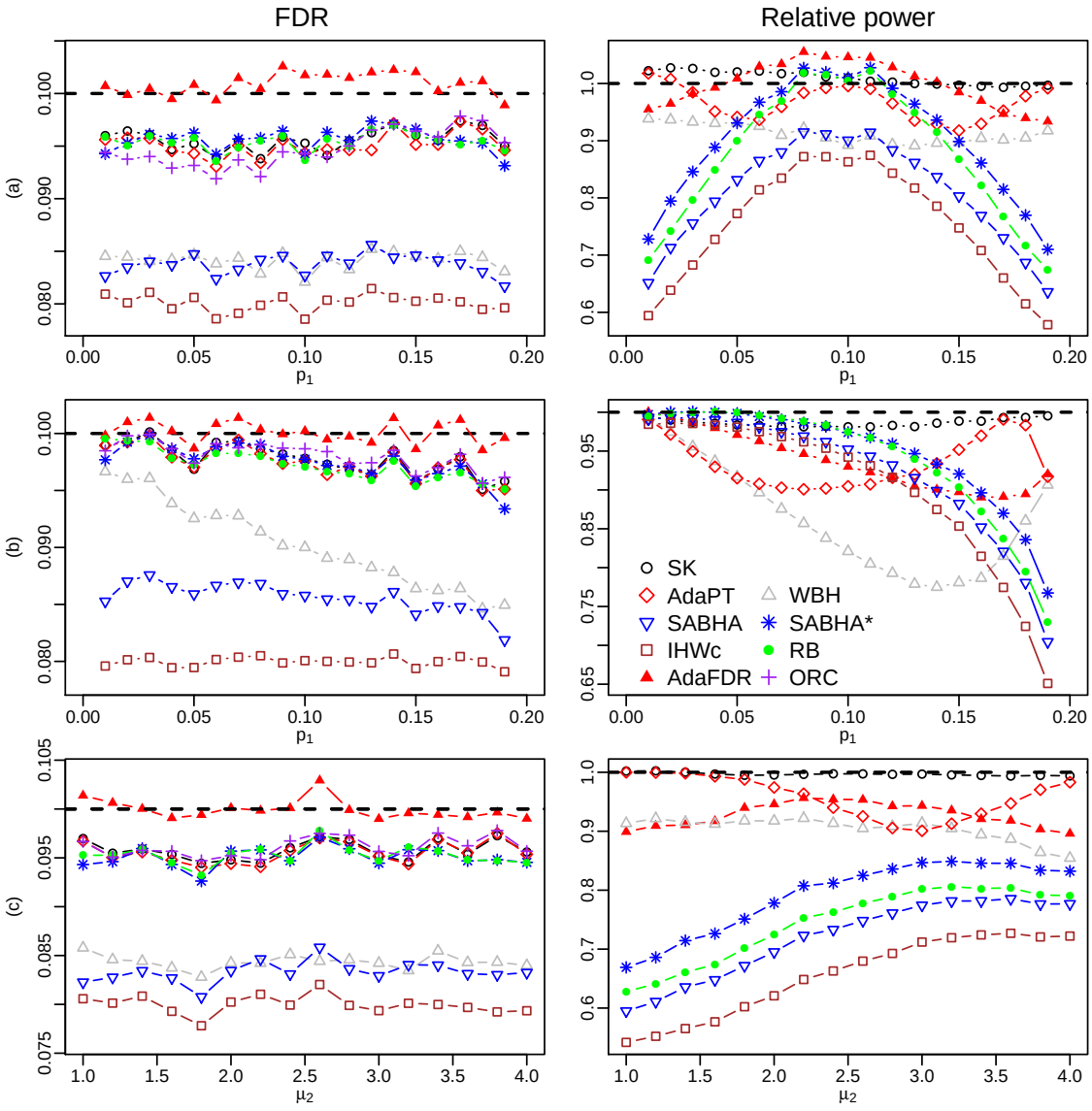

Figure 3.9: Realized FDR levels and relative power in the independent normal statistics setting with weak signals. The left column shows the realized FDR levels, and the right column shows the realized relative power. Rows (a)–(c) show the results for cases (a)–(c), respectively.

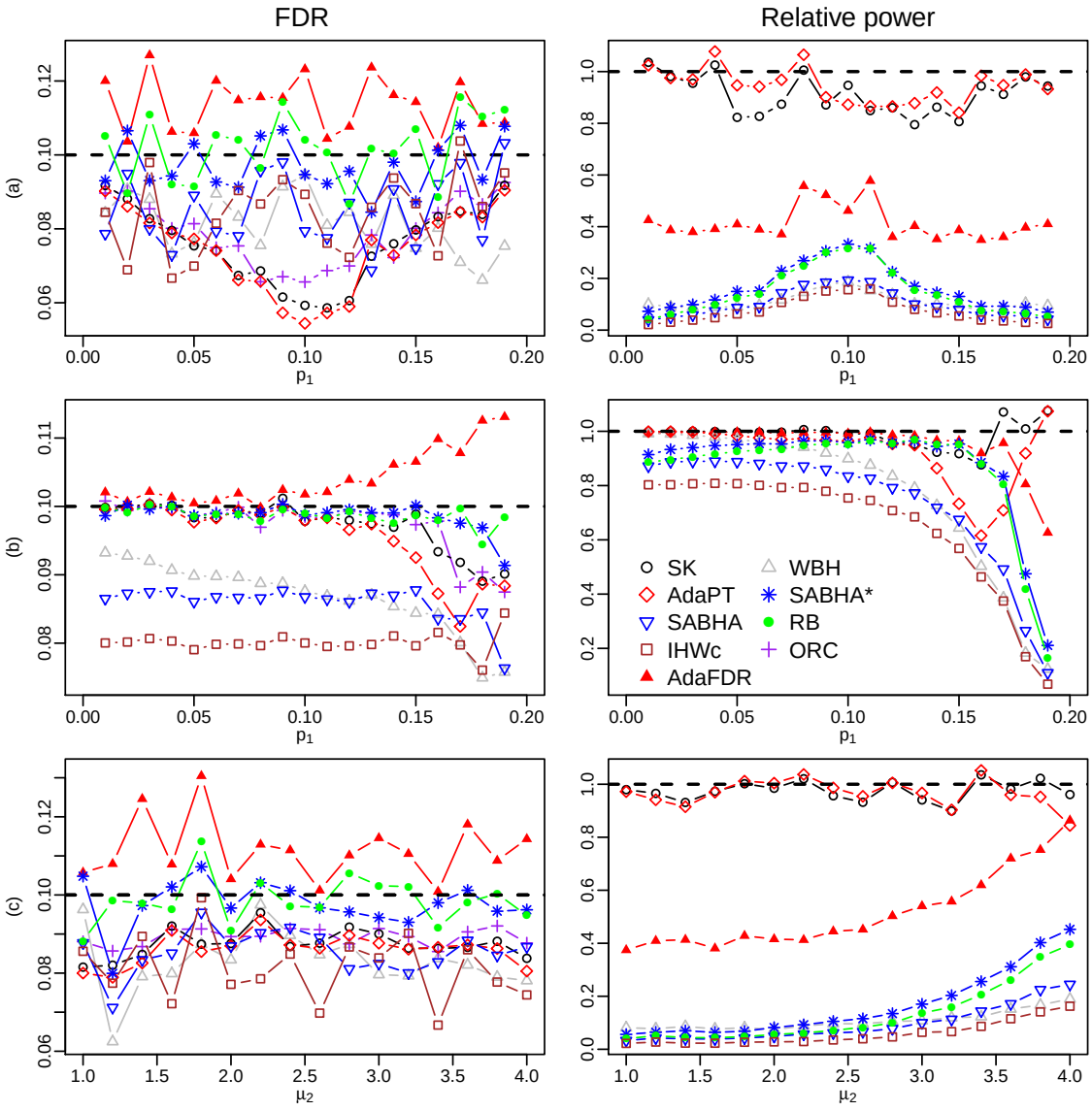

Figure 3.10: Realized FDR levels and relative power in the dependent t-statistics setting with weak signals. The left column shows the realized FDR levels, and the right column shows the realized relative power. Rows (a)–(c) show the results for cases (a)–(c), respectively.

Chapter 4

Control the false discovery rate with complex auxiliary information

4.1 Introduction

In many multiple testing applications, there exists auxiliary information available beyond a list of p-values, for example, information from related studies (Ferkingstad et al., 2008; Fortney et al., 2015). In some situations, the auxiliary information can have a complex structure. Take Example 1.3 for instance. In this example, the auxiliary information is a combination of the directional information and two numerical covariates, and one covariate is subject to missing.

As we have discussed in Section 1.3.2 and demonstrated in Chapter 3, the existing covariate-adaptive methods such as Ignatiadis and Huber (2018) and Lei and Fithian (2018), may not efficiently use the directional information. Furthermore, to the best of our knowledge, these methods did not consider the missing data in covariates. In this

chapter, we extend the signed-knockoff procedure, proposed in Chapter 3, to incorporate generic covariates and handle missing data. We call our new method as the `Codak` for covariate and direction adaptive knockoff procedure. We show that `Codak` controls the FDR in finite samples and has competitive power performance in numerical studies.

The rest of Chapter 4 is organized as follows. In Section 4.2, we define the `Codak` procedure and show its control of the FDR. In Section 4.3, we describe the implementation details of `Codak`. In Section 4.4, we compare `Codak` with existing methods in simulation studies. In Section 4.5, we apply our method to analyze Example 1.3. For convenience of presentation, technical details are presented in Section 4.6.

4.2 Covariate and direction adaptive knockoff procedure

Suppose there are n null hypotheses being tested at the same time: H_{0i}, $i = 1, \ldots, n$. For the ith null hypothesis H_{0i}, let t_i, p_i and $\mathbf{x}_i$ be the corresponding test statistic, p-value, and vector of auxiliary variables, respectively.

Recall that in Chapter 3, we define the signed p-value of the null hypothesis H_{0i} as

$$q_i = \text{sign}(t_i)(1 - p_i),$$

and the knockoff of a signed p-value q_i as

$$\tilde{q}_i = \text{sign}(q_i) - q_i.$$

We develop our method by extending the signed-knockoff procedure to incorporate generic covariates as auxiliary information. Similar to the signed-knockoff procedure, our

procedure is also an iterative procedure that reveals more and more information at each iteration. In the beginning, instead of signed p-values, we only observe unordered pairs of signed p-values and their corresponding knockoffs. As a result, the true identities of signed p-values are masked. Let $q_i^* = q_i \wedge \tilde{q}_i$ be the minimum of q_i and $\tilde{q}_i$, and knowing q_i^* is equivalent to knowing the unordered pair of $\{q_i, \tilde{q}_i\}$. An initial rejection set R_0 is decided based on the masked information, which can be expressed as the filtration:

$$\mathcal{F}_0 = \sigma\Big(\{q_i^*\}_{i=1}^n,\ \{\mathbf{x}_i\}_{i=1}^n\Big).$$

In general, the FDR of a rejection rule usually decreases when the rejection rule becomes more stringent. So we set a reasonably large R_0 and then accept null hypotheses one at a time by removing their indexes from the rejection set. We reveal the true identity of a signed p-value after its corresponding null hypothesis is accepted. Suppose that we have accepted k null hypotheses, let R_k be the current rejection set, and the information available to choose the next null hypothesis to accept is contained in the filtration:

$$\mathcal{F}_k = \sigma\Big(\mathcal{F}_0,\ \{b_i\}_{i \notin R_k}\Big),$$

where $b_i = I(|q_i| > 1/2)$. Combining b_i with q_i^* we can obtain the true identity of q_i. We estimate the FDR as:

$$\widehat{\mathrm{FDR}}_k = \frac{1 + \#\{i:\ i \in R_k \,\&\, |q_i| < 1/2\}}{\#\{i:\ i \in R_k \,\&\, |q_i| > 1/2\} \vee 1}.$$

Because the alternative signed p-values tend to be close to -1 and 1, we prefer to reject null hypotheses with large absolute signed p-values, and the count in the denominator of $\widehat{\mathrm{FDR}}_k$ is the number of rejections. Because signed p-values follow a uniform distribution, the numerator of $\widehat{\mathrm{FDR}}_k$ provides a slight over-estimate of the number of false rejections. We keep accepting null hypotheses until the FDR estimate is below the nominal FDR level α.

More formally, our procedure is defined as follows:

Definition (`Codak`).

Step 1: Set $k = 0$. Set the $\mathcal{F}_0$-measurable initial rejection set $R_0 \subseteq \{1, 2, \ldots, n\}$.

Step 2: Choose an index $i \in R_k$ and remove it from R_k. Set $R_{k+1} = R_k/\{i\}$. The index choice must be $\mathcal{F}_k$-measurable. Increase k by 1.

Step 3: If $\widehat{\text{FDR}}_k \leq \alpha$ or $R_k = \varnothing$, stop and reject any H_{i0} with $i \in R_k$ and $|q_i| > 1/2$. Otherwise, return to Step 2.

Remark 4.1. *In our implementation, we set the initial rejection set as* $R_0 = \{i : \ |q_i| < 0.45$ *or* $|q_i| > 0.55\}$ *in Step 1. Consequently, all null hypotheses with p-values larger than 0.45 will be accepted under all circumstances.*

Remark 4.2. *As we discussed in Section 3.6.3, the directional information can be viewed as a group label. However, as we demonstrated, if the distribution of null test statistics is smooth at* 0*, treating the directional information as a group label will lead to an inappropriate model and result in power loss in rejection results. This explains why we use the signed p-values to utilize the directional information instead of including the directional information in* $\mathbf{x}_i$*.*

`Codak` controls the FDR for mirror-conservative true null p-values (Lei and Fithian, 2018). A p-value p_i is mirror-conservative if and only if, under the null,

$$\text{P}(a_1 \leq p_i \leq a_2) \leq \text{P}(1 - a_2 \leq p_i \leq 1 - a_1), \text{ for any } 0 \leq a_1 \leq a_2 \leq 0.5.$$

We also require that true null test statistics satisfy the *null independence condition* introduced in Section 1.3.2. The condition requires that the true null test statistics are

independent of each other and independent of alternative test statistics. We show that `Codak` controls the FDR in finite samples.

Theorem 4.1. *Under the null independence condition, if the true null p-values are independent of the signs of their corresponding test statistics, and they follow Uniform*$(0, 1)$ *or are mirror-conservative,* ***`Codak`*** *controls the FDR at level* α*.*

It can be shown that `Codak` is a special instance of the AdaPT framework (Lei and Fithian, 2018), and our proof can be established by appealing to their Theorem 1. The details are in Section 4.6.1.

4.3 Implementation

In Step 2 of our `Codak` procedure, we only require the choice of next null hypothesis to accept to be $\mathcal{F}_k$-measurable but leave the details unspecified. As we have discussed in Section 3.3, the local FDR is the optimal ranking statistic for the signed-knockoff procedure. This conclusion also holds for the `Codak` procedure. The local FDR is defined as the posterior probability of the null hypothesis being true conditional on its test statistic and auxiliary information. To maintain the FDR control, the challenge is to estimate the local FDR with the limited information in $\mathcal{F}_k$.

4.3.1 Two-group model

Suppose that all the signed p-values are observable, consider the following two-group model. For $i = 1, \ldots, n$,

$$\begin{cases} H_i|\mathbf{x}_i \sim \text{Bernoulli}\left\{\pi_0(\mathbf{x}_i)\right\}, \\ q_i|H_i \sim H_i f_0 + (1 - H_i) f_{1\mathbf{x}_i}, \end{cases}$$

where H_i is the true null indicator, f_0 is the pdf of Uniform$(-1, 1)$, and $f_{1\mathbf{x}_i} = f_1(\cdot|\mathbf{x}_i)$ is the alternative pdf conditioning on $\mathbf{x}_i$. Both the probability of $H_i = 1$, $\pi_0(\mathbf{x}_i)$, and $f_{1\mathbf{x}_i}$ can depend on the covariates $\mathbf{x}_i$ and need to be estimated.

Under the two-group model, the overall density function f conditional on $\mathbf{x}_i$ is

$$f(q_i|\mathbf{x}_i) = \pi_0(\mathbf{x}_i)/2 + \left\{1 - \pi_0(\mathbf{x}_i)\right\} f_1(q_i|\mathbf{x}_i).$$

The local FDR can be expressed as

$$\text{Lfdr}_i = \pi_0(\mathbf{x}_i) / \left\{2 f(q_i|\mathbf{x}_i)\right\}.$$

Motivated by the choice of f_1 in Section 3.3, we propose the following parametric model of $f_{1\mathbf{x}_i}$:

$$f_1(q_i|\mathbf{x}_i) = \lambda(\mathbf{x}_i) \frac{\alpha(\mathbf{x}_i)}{2} \left(\frac{q_i + 1}{2}\right)^{\alpha(\mathbf{x}_i) - 1} + \left\{1 - \lambda(\mathbf{x}_i)\right\} \frac{\beta(\mathbf{x}_i)}{2} \left(\frac{1 - q_i}{2}\right)^{\beta(\mathbf{x}_i) - 1}. \tag{4.1}$$

We see that $f_{1\mathbf{x}_i}$ is a mixture of two components with transformed beta distributions, where $\lambda(\mathbf{x}_i)$ is the mixing proportion. We denote the two components as the α component and the β component, respectively, since $\alpha(\mathbf{x}_i)$ and $\beta(\mathbf{x}_i)$ are the shape parameter functions in the two components. By constraining $\alpha(\mathbf{x}_i)$ and $\beta(\mathbf{x}_i)$ between 0 and 1, the α component is monotonically decreasing on $(-1, 1)$, and the β component is increasing.

As both of $\pi_0(\mathbf{x}_i)$ and $\lambda(\mathbf{x}_i)$ represent proportions, we model them through the logistic regression:

$$\log\left\{\frac{\pi_0(\mathbf{x}_i)}{1-\pi_0(\mathbf{x}_i)}\right\} = \boldsymbol{\theta}_\pi^{\mathrm{T}}\boldsymbol{\Phi}_\pi(\mathbf{x}_i), \tag{4.2}$$

$$\log\left\{\frac{\lambda(\mathbf{x}_i)}{1-\lambda(\mathbf{x}_i)}\right\} = \boldsymbol{\theta}_\lambda^{\mathrm{T}}\boldsymbol{\Phi}_\lambda(\mathbf{x}_i), \tag{4.3}$$

where $\boldsymbol{\theta}_\pi$ and $\boldsymbol{\theta}_\lambda$ are two parameter vectors, and $\boldsymbol{\Phi}_\pi(\mathbf{x}_i) = \big(\Phi_{\pi 1}(\mathbf{x}_i), \ldots, \Phi_{\pi m_\pi}(\mathbf{x}_i)\big)^{\mathrm{T}}$ and $\boldsymbol{\Phi}_\lambda(\mathbf{x}_i) = \big(\Phi_{\lambda 1}(\mathbf{x}_i), \ldots, \Phi_{\lambda m_\lambda}(\mathbf{x}_i)\big)^{\mathrm{T}}$ are two featurizations of $\mathbf{x}_i$.

For computational consideration, we model $\alpha(\mathbf{x}_i)$ and $\beta(\mathbf{x}_i)$ as

$$-\log\{\alpha(\mathbf{x}_i)\} = \boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i), \tag{4.4}$$

$$-\log\{\beta(\mathbf{x}_i)\} = \boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i), \tag{4.5}$$

where $\boldsymbol{\theta}_\alpha$ and $\boldsymbol{\theta}_\beta$ are two parameter vectors, and $\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)$ and $\boldsymbol{\Phi}_\beta(\mathbf{x}_i)$ are two featurizations of $\mathbf{x}_i$. The appropriate featurization depends on the specific application. The most basic featurization directly use the covariates in $\mathbf{x}_i$. For continuous covariates, we recommend using thin-plate spline basis functions, though other basis expansions such as tensor product can also be used (Wood, 2017).

4.3.2 $\mathcal{F}_k$-measurable EM-algorithm

In Step 2 of `Codak` after k hypotheses have been accepted, we are allowed to estimate the local FDR only based on the limited information in $\mathcal{F}_k$, which are $\{q_i^*\}_{i=1}^n$, $\{\mathbf{x}_i\}_{i=1}^n$ and $\{b_i = I(|q_i| > 1/2)\}_{i\notin R_k}$.

For each $i \in R_k$, we only observe q_i^*, and it can be shown that the density function of q_i^* is

$$g(q_i^*|\mathbf{x}_i) = f(q_i^*|\mathbf{x}_i) + f(\tilde{q}_i^*|\mathbf{x}_i) = \pi_0(\mathbf{x}_i) + \{1-\pi_0(\mathbf{x}_i)\}\, f_1(q_i^*|\mathbf{x}_i) + \{1-\pi_0(\mathbf{x}_i)\}\, f_1(\tilde{q}_i^*|\mathbf{x}_i).$$

On the other hand, for each $i \notin R_k$, we observe q_i through q_i^* and b_i.

As a result, the log-likelihood function of $\left\{\{q_i^*\}_{i=1}^n, \{b_i\}_{i\notin R_k}\right\}$ is

$$\begin{aligned} l^{(k)}\big(\Theta|\mathcal{F}_k\big) &= \sum_{i\notin R_k} \log f(q_i|\mathbf{x}_i) + \sum_{i\in R_k} \log g(q_i^*|\mathbf{x}_i) \\ &= \sum_{i\notin R_k} \log\left[\pi_0(\mathbf{x}_i)/2 + \{1-\pi_0(\mathbf{x}_i)\}\, f_1(q_i|\mathbf{x}_i)\right] \\ &\quad + \sum_{i\in R_k} \log\left[\pi_0(\mathbf{x}_i) + \{1-\pi_0(\mathbf{x}_i)\}\, f_1(q_i^*|\mathbf{x}_i) + \{1-\pi_0(\mathbf{x}_i)\}\, f_1(\tilde{q}_i^*|\mathbf{x}_i)\right], \end{aligned}$$

where $\Theta = (\boldsymbol{\theta}_\pi, \boldsymbol{\theta}_\lambda, \boldsymbol{\theta}_\alpha, \boldsymbol{\theta}_\beta)$ is the combination of all four parameter vectors.

Maximizing the log-likelihood function $l^{(k)}$ directly with respect to Θ is computationally challenging. As an alternative, we develop an EM-algorithm to estimate Θ, and its details can be found in Section 4.6.2.

4.3.3 Estimating the local FDR

After we estimate Θ, for any $i \in R_k$, the local FDR of H_{0i} is

$$\widehat{\text{Lfdr}}_i = \frac{\hat{\pi}_0(\mathbf{x}_i)}{\hat{\pi}_0(\mathbf{x}_i) + \{1-\hat{\pi}_0(\mathbf{x}_i)\}\, \hat{f}_1(q_i^*;\mathbf{x}_i) + \{1-\hat{\pi}_0(\mathbf{x}_i)\}\, \hat{f}_1(\tilde{q}_i^*;\mathbf{x}_i)}.$$

Inherently, the two-group model is not identifiable, and $\widehat{\text{Lfdr}}_i$ is not unique. Similarly in spirit to the conservative identifying assumption in Lei and Fithian (2018), we add the minimum values of the estimated two alternative components to the numerator. Specifically, the minimum values of the two density functions are

$$\min_{q\in[-1,1]} \frac{\alpha(\mathbf{x}_i)}{2}\left(\frac{q+1}{2}\right)^{\alpha(\mathbf{x}_i)-1} = \frac{\alpha(\mathbf{x}_i)}{2},$$

and

$$\min_{q\in[-1,1]} \frac{\beta(\mathbf{x}_i)}{2}\left(\frac{1-q}{2}\right)^{\beta(\mathbf{x}_i)-1} = \frac{\beta(\mathbf{x}_i)}{2}.$$

In our implementation, we use the following local FDR estimate:

$$\widehat{\text{Lfdr}}_i = \frac{\hat{\pi}_0(\mathbf{x}_i) + \{1 - \hat{\pi}_0(\mathbf{x}_i)\} \left[\hat{\lambda}(\mathbf{x}_i)\hat{\alpha}(\mathbf{x}_i) + \{1 - \hat{\lambda}(\mathbf{x}_i)\}\hat{\beta}(\mathbf{x}_i)\right]}{\hat{\pi}_0(\mathbf{x}_i) + \{1 - \hat{\pi}_0(\mathbf{x}_i)\}\, \hat{f}_1(q_i^*; \mathbf{x}_i) + \{1 - \hat{\pi}_0(\mathbf{x}_i)\}\, \hat{f}_1(\tilde{q}_i^*; \mathbf{x}_i)}.$$

4.3.4 Missing data in covariates

In Example 1.3, some values of one covariate are missing. In such case where the missing pattern is simple, we can perform separate model fittings for the null hypotheses grouped by the missing pattern then accept the null hypothesis with the highest local FDR value across all groups. In this example, we group the hypotheses according to whether the corresponding microarray test statistic is missing. In the group where the microarray test statistic is missing, only the complete covariate, the gene coding region length, is utilized in the regression model. This strategy can be further extended to situations where null hypotheses are divided into multiple groups and accompanied by different sets of covariates in different groups.

Instead of the above grouping strategy, we could also impute the missing covariate values, for example, using mean or regression imputations. The imputation strategy should work well when the missing pattern is complex and the grouping strategy could lead to many groups with small group sizes.

4.4 Simulation study

In this section, we compare `Codak` with existing methods which can utilize auxiliary information through simulation experiments.

4.4.1 Simulation settings

The simulation settings are adopted from those in Lei and Fithian (2018) and Liang (2019). We assume that each null hypothesis H_i is accompanied by a bivariate covariate $\mathbf{x}_i = (x_{i1}, x_{i2})$. We generate the test statistics t_i from the distribution $f(t_i|\mathbf{x}_i) = \pi_0(\mathbf{x}_i) f_0(t_i) + \{1 - \pi_0(\mathbf{x}_i)\} f_1(t_i|\mathbf{x}_i)$. By default, we assume that f_0, the distribution of t_i under true null hypotheses, is the pdf of the standard normal distribution. The other parameters are specified in the following settings:

(a) We let $\mathbf{x}_i$ form an equally-spaced 50×50 grid in the area $[-100, 100] \times [-100, 100]$. Set $\pi_0(\mathbf{x}_i) = I(x_{i1}^2 + x_{i2}^2 \geq 900)$. Under alternative hypotheses, $t_i \sim \mathrm{N}(2, 1)$. The nominal FDR level α varies between 0.01 and 0.15.

(b) We generate x_{i1} and x_{i2} from Uniform$(0, 1)$ independently. Set $\pi_0(\mathbf{x}_i) = 0.6 + 0.3(1 - \omega)(x_{i1}^2 + x_{i2}^2 - 2/3) - 0.375\omega x_{i1} x_{i2}$, and the parameter ω varies between 0 and 1. The f_1 is a mixture of a negative component g_1 and a positive component g_2:

$$f_1(t_i|\mathbf{x}_i) = \pi_1(\mathbf{x}_i) g_1(t_i) + \{1 - \pi_1(\mathbf{x}_i)\} g_2(t_i),$$

and $\pi_1(\mathbf{x}_i) = (1 + x_{i1} x_{i2})/2$. The two components g_1 and g_2 are the pdfs of $\mathrm{N}(-1.5, 1)$ and $\mathrm{N}(3, 1)$, respectively.

(c) We start with same setup as in the setting (b) but set $\pi_1(\mathbf{x}_i) = (2 + x_{i1} + x_{i2})/4$. Furthermore, let f_0 be the pdf of the t-distribution with 8 degrees of freedom, and let g_1 and g_2 be the pdfs of non-central t-distributions $t_{8,-\sqrt{5/2}}$ and $t_{8,2\sqrt{5/2}}$, respectively.

In the setting (a), the number of null hypotheses $n = 2500$. In the settings (b) and (c), $n = 4000$, and the nominal FDR level α is set to 0.1.

To mimic Example 1.3, we also investigate the settings (a)–(c) with x_{i2} missing randomly with a probability 0.3. We use "missing" to indicate these settings to distinguish them from the original "full" settings.

4.4.2 Candidate procedures

We focus on the following procedures that can control the FDR in finite samples:

– `Codak`, the proposed covariate and direction adaptive knockoff procedure.

– `AdaPT`, the adaptive p-value thresholding procedure of Lei and Fithian (2018) applied to the covariate $\mathbf{x}_i$.

– `AdaPTsign`, a direction-adaptive implementation of `AdaPT`. We group the hypotheses according to the signs of the test statistics and fit separate models in the two groups. This strategy is similar to our grouping strategy to handle missing covariates in Section 4.3.4.

Both `AdaPT` and `AdaPTsign` are based on the `R` package `AdaptMT` available on CRAN, and we compute the two-sided p-values from the test statistics. More implementation details can be found in Section 4.6.3.

For settings where some values of x_2 are missing, we apply the candidate procedures with the grouping strategy detailed in Section 4.3.4.

Under each simulation setting, the simulation experiments are repeated 100 times, and we report the average realized FDR and power.

4.4.3 Simulation results

The results in the settings (a)–(c), as well as the corresponding settings with missing x_{i2}, are shown in Figure 4.1.

We see that all procedures control the FDR properly in each setting. In terms of power, it is not surprising that the power of every procedure with both covariates available is larger than the power with missing x_{i2} in every setting. `AdaPTsign` shows a better power than `AdaPT` in almost every parameter setting because `AdaPTsign` utilizes the directional information. Furthermore, `Codak` is more powerful than both `AdaPT` and `AdaPTsign` in every setting regardless of whether x_2 is missing or not.

4.5 Real data application

Here we analyze the RNA-seq dataset in Example 1.3. Recall that this dataset contains gene expression levels of 18151 genes in pairs of healthy and diseased samples and was first analyzed in Jabbari et al. (2012). There are two additional covariates available: the length of the gene coding region (x_1) and the test statistic from a related microarray study (x_2). Due to the technological difference between the two platforms, only 16490 of the 18151 RNA-seq genes were measured in the microarray study of Gudjonsson et al. (2010).

Liang (2019) develop a nonparametric empirical Bayes procedure to estimate the local FDR of each hypothesis conditioning on its covariates, but it is unclear whether the FDR can be controlled. We apply our procedure, `AdaPT` and `AdaPTsign` to the RNA-seq dataset with the two covariates. This dataset is available from the `R` package `calm` on Bioconductor. Similar to Liang (2019), we standardize both x_1 and x_2 to between 0 and 1 by dividing their ranks by the total number of genes. To handle the missing values in x_2, we employ

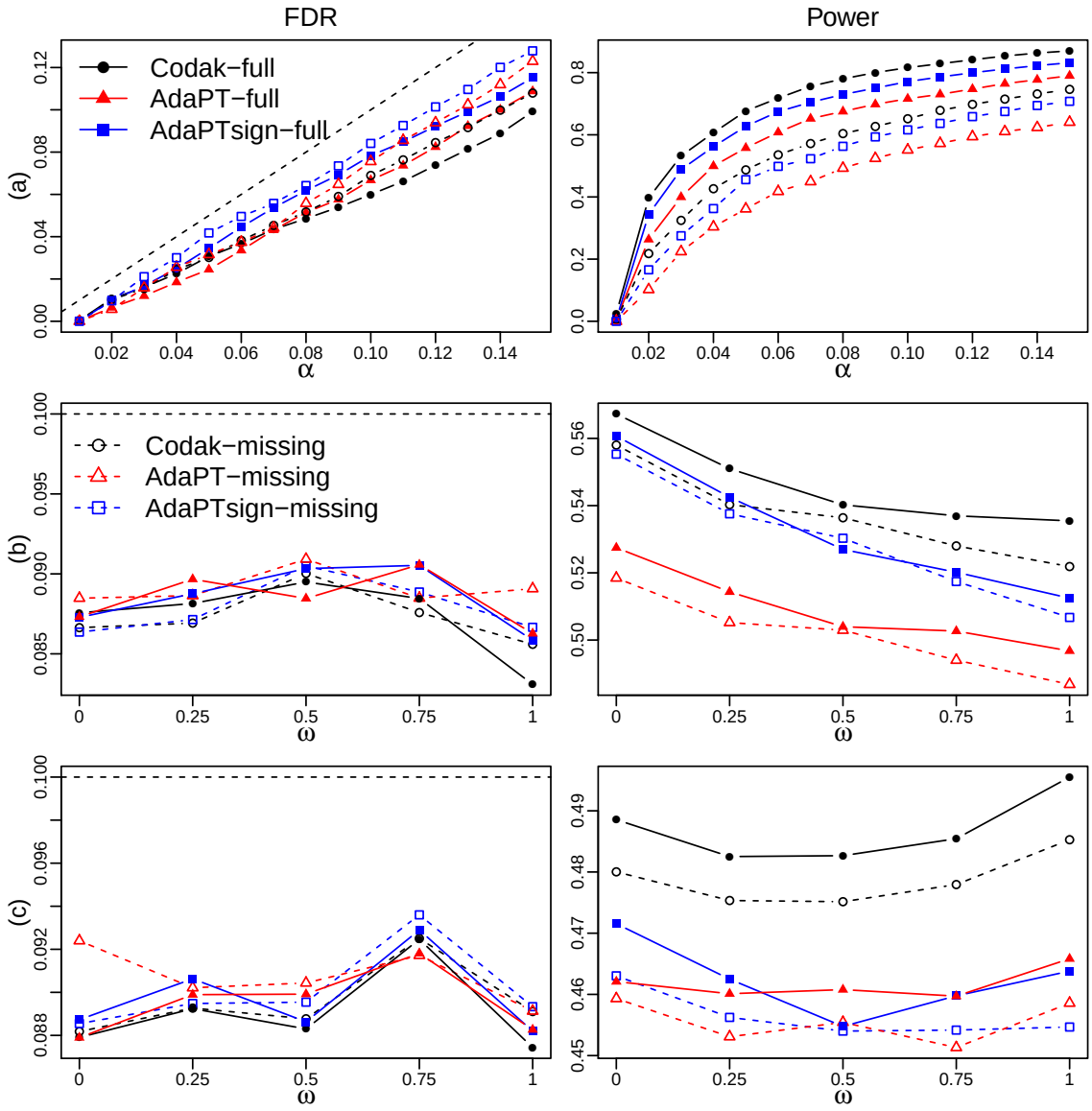

Figure 4.1: Realized FDR levels and power in settings (a), (b) and (c) in Section 4.4. The left column shows the realized FDR levels, and the dash lines indicate the target FDR levels. The right column shows power. Row labels indicate the simulation settings. Solid symbols: both covariates available; hollow symbols: missing x_2.

the grouping strategy for all three procedures.

A naive strategy of handling missing covariate values is to ignore the covariates with any missing values, which will be x_2 in this case. For the comparison purpose, we also apply the three procedures to the RNA-seq dataset with only x_1 as the covariate. We denote this covariate setting as the x_1-only setting to distinguish it from the original $x_1\&x_2$ setting, where both x_1 and x_2 are utilized.

Figure 4.2 shows the numbers of rejections under different nominal FDR levels. In the $x_1\&x_2$ setting, `AdaPT` has the least numbers of rejections by a large margin compared with the other two procedures, indicating the importance of the directional information in Example 1.3. `Codak` outperforms `AdaPTsign` when $\alpha \leq 0.1$, but in general the difference between the two is small. In the x_1-only setting, the differences among the three procedures are small, and their powers are much lower than those in the $x_1\&x_2$ setting. Intuitively, the test statistics from a closely related study can be highly informative. To avoid handling the missing data, it is convenient to simply ignore the covariates with missing values. However, it is clear such a strategy could lead to a significant power loss, and a good strategy to handle the missing data will be highly useful in practice.

4.6 Technical details

4.6.1 Proof of Theorem 4.1

Let t_i and p_i be the test statistic and p-value for the ith null hypothesis, respectively. Recall the signed p-value is defined as $q_i = \text{sign}(t_i)(1 - p_i)$, which is a function of $\text{sign}(t_i)$ and p_i. Following the argument in Section 3.6.3, `Codak` can be reconstructed as a special instance of the `AdaPT` framework in Lei and Fithian (2018) by treating $\text{sign}(t_i)$ as an

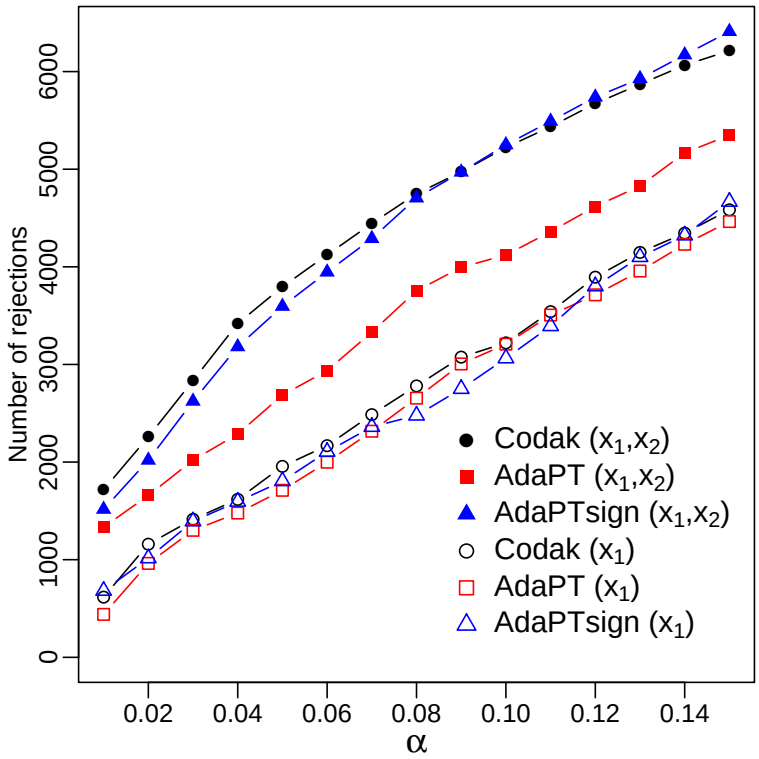

Figure 4.2: Numbers of rejections versus the nominal FDR levels α. Solid symbols and (x_1, x_2): the x_1&x_2 settings; hollow symbols and (x_1): the x_1-only settings.

additional covariate. Specifically, `Codak` has a parallel structure of the algorithm 3 of `AdaPT`. Furthermore, it is straightforward to show that $\mathcal{F}_k = \sigma\Big(\{q_i^*\}_{i=1}^n, \{\mathbf{x}_i\}_{i=1}^n, \{b_i\}_{i\notin R_k}\Big) \subseteq \sigma\Big(\{\mathbf{x}_i, \text{sign}(t_i), \tilde{p}_{k,i}\}_{i=1}^n\Big)$, where $\tilde{p}_{k,i} = p_i$ if $i \notin R_k$ and $\tilde{p}_{k,i} = \{p_i, 1-p_i\}$ if $i \in R_k$. The latter σ-field contains all the information `AdaPT` need to choose the next hypothesis to accept and is a superset of $\mathcal{F}_k$, the information can be used in `Codak`. Finally, we can use Theorem 1 in Lei and Fithian (2018) to prove our theorem.

4.6.2 Details of EM-algorithm

Recall that R_k is the rejection set after we accept k null hypotheses. Similar to Section 3.6.2, let $H_i = 1$ if the ith null hypothesis is false and $H_i = 0$ otherwise. If $H_i = 0$, i.e., q_i is from the alternative hypothesis, let $\gamma_i = 1$ if q_i is from the α component, and $\gamma_i = 0$ if q_i is from the β component. Recall that $b_i = I(|q_i| > 1/2)$. This information is useful to obtain the true values of each q_i.

Denote

$$q_{i,1} = q_i^* I(q_i^* < 0) + \tilde{q}_i^* I(q_i^* > 0),$$
$$q_{i,0} = q_i^* I(q_i^* > 0) + \tilde{q}_i^* I(q_i^* < 0).$$

With the model assumptions in (4.1)–(4.5), the log-likelihood function of the complete data is

$$\begin{aligned}
&l^c(\Theta; q_i, H_i, \gamma_i, b_i, \mathbf{x}_i) \\
=&\sum_{i=1}^n H_i \boldsymbol{\theta}_\pi^{\mathrm{T}} \boldsymbol{\Phi}_\pi(\mathbf{x}_i) - \sum_{i=1}^n \log\left[1 + \exp\left\{\boldsymbol{\theta}_\pi^{\mathrm{T}} \boldsymbol{\Phi}_\pi(\mathbf{x}_i)\right\}\right] \\
&+ \sum_{i=1}^n (1-H_i)\gamma_i \boldsymbol{\theta}_\lambda^{\mathrm{T}} \boldsymbol{\Phi}_\lambda(\mathbf{x}_i) - \sum_{i=1}^n (1-H_i) \log\left[1 + \exp\left\{\boldsymbol{\theta}_\lambda^{\mathrm{T}} \boldsymbol{\Phi}_\lambda(\mathbf{x}_i)\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i\notin R_k}(1-H_i)\gamma_i\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_i}{2}\right)\right\}\\
&+\sum_{i\in R_k}(1-H_i)\gamma_i b_i\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_{i,1}}{2}\right)\right\}\\
&+\sum_{i\in R_k}(1-H_i)\gamma_i(1-b_i)\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_{i,0}}{2}\right)\right\}\\
&+\sum_{i\notin R_k}(1-H_i)(1-\gamma_i)\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_i}{2}\right)\right\}\\
&+\sum_{i\in R_k}(1-H_i)(1-\gamma_i)b_i\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_{i,1}}{2}\right)\right\}\\
&+\sum_{i\in R_k}(1-H_i)(1-\gamma_i)(1-b_i)\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_{i,0}}{2}\right)\right\}.
\end{aligned}
$$

The core of the EM-algorithm is the EM-iteration, which contains an E-step and an M-step. We use $\Theta^{(r-1)}=(\boldsymbol{\theta}_\pi^{(r-1)},\boldsymbol{\theta}_\lambda^{(r-1)},\boldsymbol{\theta}_\alpha^{(r-1)},\boldsymbol{\theta}_\beta^{(r-1)})$ denote the value of Θ after $r-1$ iterations. When $r=1$, $\Theta^{(0)}$ denotes the initial value of Θ.

E-step

In the E-step of the rth iteration, we need to calculate the following conditional probabilities:

$$
\begin{aligned}
w_{1,i}^{(r)}=&\mathrm{P}(H_i=1|\mathcal{F}_k,\Theta^{(r-1)}),\text{ for any } i;\\
w_{2,i}^{(r)}=&\mathrm{P}(H_i=0,\gamma_i=1|\mathcal{F}_k,\Theta^{(r-1)}),\text{ for any } i;\\
w_{3,i}^{(r)}=&\mathrm{P}(H_i=0,\gamma_i=1,b_i=1|\mathcal{F}_k,\Theta^{(r-1)}),\text{ for } i\in R_k;\\
w_{4,i}^{(r)}=&\mathrm{P}(H_i=0,\gamma_i=0,b_i=1|\mathcal{F}_k,\Theta^{(r-1)}),\text{ for } i\in R_k.
\end{aligned}
$$

For convenience, let

$$\pi_{0,i}^{(r)} = \frac{\exp\left\{\boldsymbol{\theta}_{\pi}^{(r-1)\mathrm{T}}\boldsymbol{\Phi}_{\pi}(\mathbf{x}_i)\right\}}{1+\exp\left\{\boldsymbol{\theta}_{\pi}^{(r-1)\mathrm{T}}\boldsymbol{\Phi}_{\pi}(\mathbf{x}_i)\right\}};$$

$$\lambda_{i}^{(r)} = \frac{\exp\left\{\boldsymbol{\theta}_{\lambda}^{(r-1)\mathrm{T}}\boldsymbol{\Phi}_{\lambda}(\mathbf{x}_i)\right\}}{1+\exp\left\{\boldsymbol{\theta}_{\lambda}^{(r-1)\mathrm{T}}\boldsymbol{\Phi}_{\lambda}(\mathbf{x}_i)\right\}};$$

$$\alpha_{i}^{(r)} = \exp\left\{-\boldsymbol{\theta}_{\alpha}^{(r-1)\mathrm{T}}\boldsymbol{\Phi}_{\alpha}(\mathbf{x}_i)\right\};$$

$$\beta_{i}^{(r)} = \exp\left\{-\boldsymbol{\theta}_{\beta}^{(r-1)\mathrm{T}}\boldsymbol{\Phi}_{\beta}(\mathbf{x}_i)\right\},$$

and

$$\begin{aligned} f_i^{(r)} =& 2\pi_{0,i}^{(r)} + (1-\pi_{0,i}^{(r)})\lambda_i^{(r)}\alpha_i^{(r)}\left\{\left(\frac{1+q_i^*}{2}\right)^{\alpha_i^{(r)}-1} + \left(\frac{1+\tilde{q}_i^*}{2}\right)^{\alpha_i^{(r)}-1}\right\} \\ &+ (1-\pi_{0,i}^{(r)})(1-\lambda_i^{(r)})\beta_i^{(r)}\left\{\left(\frac{1-q_i^*}{2}\right)^{\beta_i^{(r)}-1} + \left(\frac{1-\tilde{q}_i^*}{2}\right)^{\beta_i^{(r)}-1}\right\}. \end{aligned}$$

For $i \notin R_k$, q_i is $\mathcal{F}_k$-measurable, so $w_{1,i}^{(r)}$ and $w_{2,i}^{(r)}$ can be calculated with the formulas below:

$$w_{1,i}^{(r)} = \frac{\pi_{0,i}^{(r)}}{\pi_{0,i}^{(r)} + (1-\pi_{0,i}^{(r)})\lambda_i^{(r)}\alpha_i^{(r)}\left(1/2+q_i/2\right)^{\alpha_i^{(r)}-1} + (1-\pi_{0,i}^{(r)})(1-\lambda_i^{(r)})\beta_i^{(r)}\left(1/2-q_i/2\right)^{\beta_i^{(r)}-1}};$$

$$w_{2,i}^{(r)} = \frac{(1-\pi_{0,i}^{(r)})\lambda_i^{(r)}\alpha_i^{(r)}\left(1/2+q_i/2\right)^{\alpha_i^{(r)}-1}}{\pi_{0,i}^{(r)} + (1-\pi_{0,i}^{(r)})\lambda_i^{(r)}\alpha_i^{(r)}\left(1/2+q_i/2\right)^{\alpha_i^{(r)}-1} + (1-\pi_{0,i}^{(r)})(1-\lambda_i^{(r)})\beta_i^{(r)}\left(1/2-q_i/2\right)^{\beta_i^{(r)}-1}}.$$

For $i \in R_k$, only q_i^* is available in $\mathcal{F}_k$. As the density function of q_i^* can be directly derived from that of q_i: $g(q_i^*) = f(q_i^*) + f(\tilde{q}_i^*)$, $w_{1,i}^{(r)}$, $w_{2,i}^{(r)}$, $w_{3,i}^{(r)}$, and $w_{4,i}^{(r)}$ can be calculated

as follows:

$$
\begin{aligned}
w_{1,i}^{(r)} =& \frac{2\pi_{0,i}^{(r)}}{f_i^{(r)}};\\
w_{2,i}^{(r)} =& \frac{(1-\pi_{0,i}^{(r)})\lambda_i^{(r)}\alpha_i^{(r)}\left\{(1/2+q_i^*/2)^{\alpha_i^{(r)}-1}+(1/2+\tilde{q}_i^*/2)^{\alpha_i^{(r)}-1}\right\}}{f_i^{(r)}};\\
w_{3,i}^{(r)} =& \frac{(1-\pi_{0,i}^{(r)})\lambda_i^{(r)}\alpha_i^{(r)}\,(1/2+q_{i,1}/2)^{\alpha_i^{(r)}-1}}{f_i^{(r)}};\\
w_{4,i}^{(r)} =& \frac{(1-\pi_{0,i}^{(r)})(1-\lambda_i^{(r)})\beta_i^{(r)}\,(1/2-q_{i,1}/2)^{\beta_i^{(r)}-1}}{f_i^{(r)}}.
\end{aligned}
$$

M-step

In the M-step of the rth iteration, we need to find the value of Θ that maximizes the following expectation:

$$
\begin{aligned}
&\mathrm{E}\left\{l^c(\Theta; H_i, \gamma_i, b_i, \mathbf{x}_i)|\mathcal{F}_k, \Theta^{(r-1)}\right\}\\
=&\sum_{i=1}^{n} w_{1,i}^{(r)}\boldsymbol{\theta}_\pi^{\mathrm{T}}\boldsymbol{\Phi}_\pi(\mathbf{x}_i) - \sum_{i=1}^{n}\log\left[1+\exp\left\{\boldsymbol{\theta}_\pi^{\mathrm{T}}\boldsymbol{\Phi}_\pi(\mathbf{x}_i)\right\}\right]\\
&+\sum_{i=1}^{n} w_{2,i}^{(r)}\boldsymbol{\theta}_\lambda^{\mathrm{T}}\boldsymbol{\Phi}_\lambda(\mathbf{x}_i) - \sum_{i=1}^{n}(1-w_{1,i}^{(r)})\log\left[1+\exp\left\{\boldsymbol{\theta}_\lambda^{\mathrm{T}}\boldsymbol{\Phi}_\lambda(\mathbf{x}_i)\right\}\right]\\
&+\sum_{i\notin R_k} w_{2,i}^{(r)}\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_i}{2}\right)\right\}\\
&+\sum_{i\in R_k} w_{3,i}^{(r)}\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_{i,1}}{2}\right)\right\}\\
&+\sum_{i\in R_k} (w_{2,i}^{(r)}-w_{3,i}^{(r)})\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_{i,0}}{2}\right)\right\}\\
&+\sum_{i\notin R_k} (1-w_{1,i}^{(r)}-w_{2,i}^{(r)})\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_i}{2}\right)\right\}
\end{aligned}
$$

$$+\sum_{i\in R_k} w_{4,i}^{(r)}\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_{i,1}}{2}\right)\right\}$$
$$+\sum_{i\in R_k}(1-w_{1,i}^{(r)}-w_{2,i}^{(r)}-w_{4,i}^{(r)})\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_{i,0}}{2}\right)\right\}.$$

After some algebra work, it can be shown that

$$\mathrm{E}\left\{l^c(\Theta;H_i,\gamma_i,b_i,\mathbf{x}_i)|\mathcal{F}_k,\Theta^{(r-1)}\right\}=f_\pi(\boldsymbol{\theta}_\pi)+f_\lambda(\boldsymbol{\theta}_\lambda)+f_\alpha(\boldsymbol{\theta}_\alpha)+f_\beta(\boldsymbol{\theta}_\beta),$$

where

$$f_\pi(\boldsymbol{\theta}_\pi)=\sum_{i=1}^n w_{1,i}^{(r)}\boldsymbol{\theta}_\pi^{\mathrm{T}}\boldsymbol{\Phi}_\pi(\mathbf{x}_i)-\sum_{i=1}^n\log\left[1+\exp\left\{\boldsymbol{\theta}_\pi^{\mathrm{T}}\boldsymbol{\Phi}_\pi(\mathbf{x}_i)\right\}\right]; \tag{4.6}$$

$$f_\lambda(\boldsymbol{\theta}_\lambda)=\sum_{i=1}^n w_{2,i}^{(r)}\boldsymbol{\theta}_\lambda^{\mathrm{T}}\boldsymbol{\Phi}_\lambda(\mathbf{x}_i)-\sum_{i=1}^n(1-w_{1,i}^{(r)})\log\left[1+\exp\left\{\boldsymbol{\theta}_\lambda^{\mathrm{T}}\boldsymbol{\Phi}_\lambda(\mathbf{x}_i)\right\}\right]; \tag{4.7}$$

$$\begin{aligned}f_\alpha(\boldsymbol{\theta}_\alpha)=&\sum_{i\notin R_k} w_{2,i}^{(r)}\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_i}{2}\right)\right\}\\&+\sum_{i\in R_k} w_{3,i}^{(r)}\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_{i,1}}{2}\right)\right\}\\&+\sum_{i\in R_k}(w_{2,i}^{(r)}-w_{3,i}^{(r)})\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\alpha^{\mathrm{T}}\boldsymbol{\Phi}_\alpha(\mathbf{x}_i)\right\}\log\left(\frac{1+q_{i,0}}{2}\right)\right\};\end{aligned} \tag{4.8}$$

$$\begin{aligned}f_\beta(\boldsymbol{\theta}_\beta)=&\sum_{i\notin R_k}(1-w_{1,i}^{(r)}-w_{2,i}^{(r)})\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_i}{2}\right)\right\}\\&+\sum_{i\in R_k} w_{4,i}^{(r)}\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_{i,1}}{2}\right)\right\}\\&+\sum_{i\in R_k}(1-w_{1,i}^{(r)}-w_{2,i}^{(r)}-w_{4,i}^{(r)})\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)+\exp\left\{-\boldsymbol{\theta}_\beta^{\mathrm{T}}\boldsymbol{\Phi}_\beta(\mathbf{x}_i)\right\}\log\left(\frac{1-q_{i,0}}{2}\right)\right\}.\end{aligned} \tag{4.9}$$

Hence, we update Θ via

$$\boldsymbol{\theta}_\pi^{(r)}=\arg\max_{\boldsymbol{\theta}_\pi} f_\pi(\boldsymbol{\theta}_\pi);$$

$$\boldsymbol{\theta}_{\lambda}^{(r)} = \arg\max_{\boldsymbol{\theta}_{\lambda}} f_{\lambda}(\boldsymbol{\theta}_{\lambda});$$
$$\boldsymbol{\theta}_{\alpha}^{(r)} = \arg\max_{\boldsymbol{\theta}_{\alpha}} f_{\alpha}(\boldsymbol{\theta}_{\alpha});$$
$$\boldsymbol{\theta}_{\beta}^{(r)} = \arg\max_{\boldsymbol{\theta}_{\beta}} f_{\beta}(\boldsymbol{\theta}_{\beta}).$$

The above four optimization problems can be easily solved through the `glm` function in `R`. This is because the functions in (4.6)–(4.9) can be viewed as either the log-likelihood function of $2n$ Binomial random variables or the log-likelihood function of $2n$ Gamma random variables.

We consider (4.6) first. Assume that random variables Y_{1i} follow the binomial distribution with covariates $\mathbf{x}_i^*$:

$$Y_{1i} \sim \text{Bin}\big(1, \pi_0(\mathbf{x}_i^*)\big).$$

The $2n$ observations for $(Y_1, \mathbf{x}^*)$ are

$$Y_{1i} = \begin{cases} 1, \, 1 \le i \le n \\ 0, \, n+1 \le i \le 2n \end{cases} \quad \text{and} \quad \mathbf{x}_i^* = \begin{cases} \mathbf{x}_i, \, 1 \le i \le n \\ \mathbf{x}_{i-n}, \, n+1 \le i \le 2n \end{cases}.$$

The weights for $2n$ observations are $\mathbf{w}_{\pi}^{(r)} = (w_{\pi,1}^{(r)}, \ldots, w_{\pi,2n}^{(r)})$ with

$$w_{\pi,i}^{(r)} = \begin{cases} w_{1,i}^{(r)}, \, 1 \le i \le n \\ 1 - w_{1,i-n}^{(r)}, \, n+1 \le i \le 2n \end{cases}.$$

The log-likelihood function of $\{(Y_{1i}, \mathbf{x}_i^*, w_{\pi,i}^{(r)})\}_{i=1}^{2n}$ is the same as (4.6).

For (4.7), we consider random variables Y_{2i} with

$$Y_{2i} \sim \text{Bin}\big(1, \lambda(\mathbf{x}_i^*)\big).$$

The $2n$ observations for Y_2 are

$$Y_{2i} = \begin{cases} 1,\, 1 \le i \le n \\ 0,\, n+1 \le i \le 2n \end{cases}$$

and the corresponding weights $\mathbf{w}_\lambda^{(r)} = (w_{\lambda,1}^{(r)}, \ldots, w_{\lambda,2n}^{(r)})$ are

$$w_{\lambda,i}^{(r)} = \begin{cases} w_{2,i}^{(r)},\, 1 \le i \le n \\ 1 - w_{1,i-n}^{(r)} - w_{2,i-n}^{(r)},\, n+1 \le i \le 2n \end{cases}.$$

The log-likelihood function of $\{(Y_{2i}, \mathbf{x}_i^*, w_{\lambda,i}^{(r)})\}_{i=1}^{2n}$ is the same as (4.7).

For (4.8), we consider random variables Y_{3i} with

$$Y_{3i} \sim \text{Gamma}\big(1, \alpha(\mathbf{x}_i^*)\big),$$

where the first parameter of the Gamma distribution indicates the shape and the second parameter indicates the rate. The $2n$ observations for Y_3 are

$$Y_{3i} = \begin{cases} -\log\left\{(1+q_i)/2\right\},\, 1 \le i \le n \,\&\, i \notin R_k \\ -\log\left\{(1+q_{i,1})/2\right\},\, 1 \le i \le n \,\&\, i \in R_k \\ -\log\left\{(1+\tilde{q}_{i-n})/2\right\},\, n+1 \le i \le 2n \,\&\, i-n \notin R_k \\ -\log\left\{(1+q_{i-n,0})/2\right\},\, n+1 \le i \le 2n \,\&\, i-n \in R_k \end{cases}$$

and the corresponding weights $\mathbf{w}_\alpha^{(r)} = (w_{\alpha,1}^{(r)}, \ldots, w_{\alpha,2n}^{(r)})$ are

$$w_{\alpha,i}^{(r)} = \begin{cases} w_{2,i}^{(r)},\, 1 \le i \le n \,\&\, i \notin R_k \\ w_{3,i}^{(r)},\, 1 \le i \le n \,\&\, i \in R_k \\ 0,\, n+1 \le i \le 2n \,\&\, i-n \notin R_k \\ w_{2,i-n}^{(r)} - w_{3,i-n}^{(r)},\, n+1 \le i \le 2n \,\&\, i-n \in R_k \end{cases}.$$

The log-likelihood function of $\{(Y_{3i}, \mathbf{x}_i^*, w_{\alpha,i}^{(r)})\}_{i=1}^{2n}$ is the same as (4.8).

For (4.9), we consider random variables Y_{4i} with

$$Y_{4i} \sim \text{Gamma}\big(1, \beta(\mathbf{x}_i)\big).$$

The $2n$ observations for Y_4 are

$$Y_{4i} = \begin{cases} -\log\left\{(1-q_i)/2\right\},\ 1 \le i \le n \,\&\, i \notin R_k \\ -\log\left\{(1-q_{i,1})/2\right\},\ 1 \le i \le n \,\&\, i \in R_k \\ -\log\left\{(1-\tilde{q}_{i-n})/2\right\},\ n+1 \le i \le 2n \,\&\, i-n \notin R_k \\ -\log\left\{(1-q_{i-n,0})/2\right\},\ n+1 \le i \le 2n \,\&\, i-n \in R_k \end{cases}$$

and the corresponding weights $\mathbf{w}_\beta^{(r)} = (w_{\beta,1}^{(r)}, \ldots, w_{\beta,2n}^{(r)})$ are

$$w_{\beta,i}^{(r)} = \begin{cases} 1 - w_{1,i}^{(r)} - w_{2,i}^{(r)},\ 1 \le i \le n \,\&\, i \notin R_k \\ w_{4,i}^{(r)},\ 1 \le i \le n \,\&\, i \in R_k \\ 0,\ n+1 \le i \le 2n \,\&\, i-n \notin R_k \\ 1 - w_{1,i-n}^{(r)} - w_{2,i-n}^{(r)} - w_{4,i-n}^{(r)},\ n+1 \le i \le 2n \,\&\, i-n \in R_k \end{cases}.$$

The log-likelihood function of $\{(Y_{4i}, \mathbf{x}_i^*, w_{\beta,i}^{(r)})\}_{i=1}^{2n}$ is the same as (4.9).

With the above discussion, $(\boldsymbol{\theta}_\pi^{(r)}, \boldsymbol{\theta}_\lambda^{(r)}, \boldsymbol{\theta}_\alpha^{(r)}, \boldsymbol{\theta}_\beta^{(r)})$ can be obtained with the following pesudocode:

$\boldsymbol{\theta}_\pi^{(r)} \leftarrow \text{glm}(Y_1 \sim \boldsymbol{\Phi}_\pi(x), \text{data} = \mathbf{x}^*, \text{family} = \text{binomial}, \text{weights} = \mathbf{w}_\pi^{(r)})$;

$\boldsymbol{\theta}_\lambda^{(r)} \leftarrow \text{glm}(Y_2 \sim \boldsymbol{\Phi}_\lambda(x), \text{data} = \mathbf{x}^*, \text{family} = \text{binomial}, \text{weights} = \mathbf{w}_\lambda^{(r)})$;

$\boldsymbol{\theta}_\alpha^{(r)} \leftarrow \text{glm}(Y_3 \sim \boldsymbol{\Phi}_\alpha(x), \text{data} = \mathbf{x}^*, \text{family} = \text{Gamma}(\text{link}=\text{“log”}), \text{weights} = \mathbf{w}_\alpha^{(r)})$;

$\boldsymbol{\theta}_\beta^{(r)} \leftarrow \text{glm}(Y_4 \sim \boldsymbol{\Phi}_\beta(x), \text{data} = \mathbf{x}^*, \text{family} = \text{Gamma}(\text{link}=\text{“log”}), \text{weights} = \mathbf{w}_\beta^{(r)})$.

Initialization

We first discuss the initial value $\Theta^{(0)} = (\boldsymbol{\theta}_\pi^{(0)}, \boldsymbol{\theta}_\lambda^{(0)}, \boldsymbol{\theta}_\alpha^{(0)}, \boldsymbol{\theta}_\beta^{(0)})$ of Θ when $k = 0$.

To initialize $\boldsymbol{\theta}_\pi$ and $\boldsymbol{\theta}_\lambda$, we define

$$(H_{0i}, H_{1i}, H_{2i}) = \begin{cases} (1,0,0), \text{ if } i \notin R_0 \\ (0,1,0), \text{ if } i \in R_0 \text{ and } q_i < 0 \\ (0,0,1), \text{ if } i \in R_0 \text{ and } q_i > 0 \end{cases}$$

and model (H_{0i}, H_{1i}, H_{2i}) as

$$(H_{0i}, H_{1i}, H_{2i}) \sim \text{Multi}\Big(1; \pi_0(\mathbf{x}_i), \{1-\pi_0(\mathbf{x}_i)\}\,\lambda(\mathbf{x}_i), \{1-\pi_0(\mathbf{x}_i)\}\,\{1-\lambda(\mathbf{x}_i)\}\Big).$$

We further assign a weight 10 to each i not in R_0 and a weight 1 to each i in R_0. Then $(\boldsymbol{\theta}_\pi^{(0)}, \boldsymbol{\theta}_\lambda^{(0)})$ is chosen to be the maximum likelihood estimate of $(\boldsymbol{\theta}_\pi, \boldsymbol{\theta}_\lambda)$ based on $\{(H_{0i}, H_{1i}, H_{2i})\}_{i=1}^n$ and their corresponding weights.

To initialize $\boldsymbol{\theta}_\alpha$, we define

$$w_{\alpha,i}^{(0)} = \begin{cases} I(q_i < -1/2),\ 1 \le i \le n \,\&\, i \notin R_0 \\ I(q_i^* < 0),\ 1 \le i \le n \,\&\, i \in R_0 \\ I(\tilde{q}_{i-n} < -1/2),\ n+1 \le i \le 2n \,\&\, i \notin R_0 \\ 0,\ n+1 \le i \le 2n \,\&\, i - n \in R_0 \end{cases}$$

and choose $\boldsymbol{\theta}_\alpha^{(0)}$ to be the maximum likelihood estimate of $\boldsymbol{\theta}_\alpha$ based on $\{(Y_{3i}, \mathbf{x}_i^*, w_{\alpha,i}^{(0)})\}_{i=1}^{2n}$.

To initialize $\boldsymbol{\theta}_\beta$, we define

$$w_{\beta,i}^{(0)} = \begin{cases} I(q_i > 1/2),\ 1 \le i \le n \,\&\, i \notin R_0 \\ I(q_i^* > 0),\ 1 \le i \le n \,\&\, i \in R_0 \\ I(\tilde{q}_{i-n} > 1/2),\ n+1 \le i \le 2n \,\&\, i-n \notin R_0 \\ 0,\ n+1 \le i \le 2n \,\&\, i-n \in R_0 \end{cases}$$

and choose $\boldsymbol{\theta}_\beta^{(0)}$ to be the maximum likelihood estimate of $\boldsymbol{\theta}_\beta$ based on $\{(Y_{4i}, \mathbf{x}_i^*, w_{\beta,i}^{(0)})\}_{i=1}^{2n}$.

When $k \ge 1$, we set the initial value of Θ as its estimate after accepting $k-1$ hypotheses.

4.6.3 Implementation of `AdaPT` and `AdaPTsign` in Section 4.4

In this section, we outline how to implement `AdaPT` and `AdaPTsign` by using `AdaPT_gam` function in the `R` package `AdaptMT`, which is available on CRAN.

We first consider the case that there is no missing value in x_2. The implementation of `AdaPT` is quite straightforward. This can be achieved by using the following code.

```
adaptMT::adapt_gam(x, p, pi_formulas = "s(x1, x2)", mu_formulas = "s(x1,
                                    x2)", alphas = 0.05)
```

In the above code, we use `x` to denote the data frame with two numerical covariates `x1` and `x2`, and use `p` to denote the p-values.

The implementation of `AdaPTsign` in the first case can also be accomplished by using the code as follows.

```
adaptMT::adapt_gam(x, p, pi_formulas = "s(x1, x2, by = sign) + sign",
     mu_formulas = "s(x1, x2, by = sign) + sign", alphas = 0.05)
```

In the above code, the new component `sign` is a factor vector indicating the signs of test statistics. The `R` code separates the hypotheses according to the values of `sign` and fit two different models in the two groups.

We next consider the case that there are some missing values in x_2. In this case, the implementation of `AdaPT` can be achieved using the following `R` code.

```
adaptMT::adapt_gam(x, p, pi_formulas = "s(x1, x2, by = 1-x3) + s(x1, by =
  x3)", mu_formulas = "s(x1, x2, by = 1-x3) + s(x1, by = x3)", alphas =
                                0.05)
```

The new component `x3` is a numerical vector. It takes value 1 if x_{i2} is missing and 0 otherwise.

The implementation of `AdaPTsign` in the second case can be accomplished by dividing the p-values into four groups.

```
regformula = "s(x1, x2, by = (1 - x3) * signy) + s(x1, x2, by = (1 - x3) *
  (1 - signy)) + s(x1, by = x3 * signy) + s(x1, by = x3 * (1 - signy))"

   adaptMT::adapt_gam(x, p, pi_formulas = regformula, mu_formulas =
                        regformula, alphas = 0.05)
```

In the above code, `signy` is a numerical vector which takes value 1 if the test statistic is positive and 0 otherwise.

Chapter 5

Summary and future work

5.1 Summary

We first provide a summary of our contributions in this thesis.

In Chapter 2, we propose a maximum multinomial likelihood method to estimate the unknown parameters in a compound mixture model (2.2). We establish the asymptotic properties of the proposed estimator and develop an EM-algorithm to numerically calculate the estimate. The proposed method is implemented in `R` package `mmle`. Numerical studies show that our method is more efficient than the binomial estimator, and exhibits clear advantage over parametric and semiparametric methods when the model assumptions on F_1 and F_2 are not satisfied. In the end, we illustrate our method using a malaria dataset.

In Chapter 3, we propose a novel direction-adaptive procedure for multiple hypothesis testing problems, which controls the FDR in finite samples. We also develop an EM-algorithm to estimate the local FDRs to improve the power performance of our procedure. The proposed method is implemented in `R` package `SK`. Simulation results demonstrate our

power advantage over existing methods. We also apply our method to real data applications, including a thale cress seedling dataset and a prostate cancer dataset.

In Chapter 4, we extend the method in Chapter 3 to incorporate other types of auxiliary information, including group structures and generic covariates, while maintaining its finite-sample FDR control property. Our method can work in situations where null hypotheses are divided into groups in which null hypotheses are accompanied by different sets of covariates. We have implemented our method in `R` package `Codak` and demonstrated our power advantage over existing covariate-adaptive methods through simulation studies and a real data application of psoriasis vulgaris disease.

5.2 Future work

The proposed methods in this thesis are expected to be promising for other research problems. In this section, we outline several interesting topics worthy of further studies.

The method discussed in Chapter 2 can be generalized to the case where covariates exist. Suppose that we have covariates (such as age and sex) $\mathbf{Z}_{11}, \ldots, \mathbf{Z}_{1m}$ associated with the parasite levels $X_1, \ldots, X_m$ in the dry season and $\mathbf{Z}_{21}, \ldots, \mathbf{Z}_{2n}$ associated with the parasite levels $Y_1, \ldots, Y_n$ in the wet season. We may incorporate these covariates by studying a semiparametric model: $\log(X) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_1 + \epsilon_X$ and $\log(Y) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_2 + \epsilon_Y$, where $X > 0$ and $Y > 0$ are the positive parasite levels for nonmalaria and malaria individuals, and ϵ_X and ϵ_Y have cdfs G_1 and G_2, respectively. The maximum multinomial likelihood method can be used to estimate $(\lambda, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, G_1, G_2)$. More specifically, for $i = 1, \ldots, m$,

$$\mathbf{M}_i(t) \sim \mathrm{Multi}\Big(1; p, (1-p)G_1(t - \mathbf{Z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1), (1-p)\bar{G}_1(t - \mathbf{Z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)\Big),$$

and for $j = 1, \ldots, n$

$$\begin{aligned}\mathbf{N}_j(t) \;\sim\; \text{Multi}\Big(1; (1-\lambda)p, (1-\lambda)(1-p)G_1(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_1) + \lambda G_2(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_2),\\ (1-\lambda)(1-p)\bar{G}_1(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_1) + \lambda \bar{G}_2(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_2)\Big),\end{aligned}$$

where $\mathbf{M}_i(t)$ and $\mathbf{N}_j(t)$ are defined in (2.5) and (2.6), respectively. For any given $t > 0$, we define

$$\begin{aligned}&\tilde{l}_M(\lambda, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, G_1, G_2; t)\\ &= \sum_{i=1}^{m} \left[m_{i0} \log p + m_{i1}(t) \log\{(1-p)G_1(t - \mathbf{Z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)\} + m_{i2}(t) \log\{(1-p)\bar{G}_1(t - \mathbf{Z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)\}\right]\\ &\quad + \sum_{j=1}^{n} \Big[n_{j0} \log\{(1-\lambda)p\} + n_{j1}(t) \log\{(1-\lambda)(1-p)G_1(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_1) + \lambda G_2(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_2)\}\\ &\qquad + n_{j2}(t) \log\{(1-\lambda)(1-p)\bar{G}_1(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_1) + \lambda \bar{G}_2(t - \mathbf{Z}_{2j}^{\mathrm{T}}\boldsymbol{\beta}_2)\}\Big].\end{aligned}$$

Taking summation over all $t_h \in I_q$, we obtain the multinomial likelihood

$$l_M(\lambda, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, G_1, G_2) = \sum_{h \in I_q} \tilde{l}_M(\lambda, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, G_1, G_2; t_h).$$

Maximizing $l_M(\lambda, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, G_1, G_2)$ gives the estimator of $(\lambda, p, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, G_1, G_2)$. We plan to investigate the asymptotic properties of the resultant estimator in future research.

In Chapter 3, we show that the proposed signed-knockoff procedure can control the FDR in finite samples under the *null independence condition*. Furthermore, we show through simulation studies in Section 3.4.3 that our procedure is reasonably robust in a setting that mimics the dependence pattern in gene expression data. It would be interesting to study if the signed-knockoff procedure can control the FDR under such dependence pattern theoretically. Also, we plan to investigate other dependence settings, under which our method can be shown to control the FDR. We leave these theoretical and numerical studies for future research.

In Chapter 3, our proposed method assigns equal weights to the signals on the positive side and the negative side. In practice, there may be reasons to emphasize more on the positive signals or the negative ones. For example, Basu et al. (2018) proposed a method that controls a weighted version of the FDR. By assigning different weights to false rejections on different sides, their method can give a rejection rule which is stricter on one side than on the other side. Motivated by Basu et al. (2018), we are interested in incorporating prior knowledge about directions into our methods. This will leave as future research.

In Section 4.5, we group the null hypotheses according to whether one covariate is missing or not and then model separately in the two groups. Alternatively, we could impute the missing values, and we plan to investigate the implementation details in our future study.

Bibliography

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* 37, 3099–3132.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* 26, 641–647.

Balabdaoui, F. (2017). Revisiting the Hodges–Lehmann estimator in a location mixture model: is asymptotic normality good enough? *Electronic Journal of Statistics* 11, 4563–4595.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 43, 2055–2085.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association* 113, 1172–1183.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009). An EM–like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics* 18, 505–526.

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* 57, 289–300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 25, 60–83.

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

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature* 526, 207–211.

Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis* 51, 5429–5443.

Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics* 34, 1204–1232.

Chauveau, D., Hunter, D. R., and Levine, M. (2015). Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys* 9, 1–31.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39, 1–38.

Dykstra, R., Kochar, S., and Robertson, T. (1995). Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association* 90, 1034–1040.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96, 1151–1160.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics* 9, 139–158.

Efron, B. (2010). *Large–Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press.

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics* 2, 714–735.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B., and Kim, S. K. (2015). Genome–wide scan informed by age–related disease identifies loci for exceptional human longevity. *PLOS Genetics* 11, 1–23.

Gudjonsson, J. E., Ding, J., Johnston, A., Tejasvi, T., Guzman, A. M., Nair, R. P., Voorhees, J. J., Abecasis, G. R., and Elder, J. T. (2010). Assessment of the psoriatic transcriptome in a large sample: additional regulated genes and comparisons with in vitro models. *Journal of Investigative Dermatology* 130, 1829–1840.

Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics* 31, 201–224.

Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics* 35, 224–251.

Ignatiadis, N. and Huber, W. (2018). Covariate–powered weighted multiple testing with false discovery rate control. *arXiv preprint arXiv:1701.05179*.

Jabbari, A., Suárez-Fariñas, M., Dewell, S., and Krueger, J. G. (2012). Transcriptional profiling of psoriasis using RNA–seq reveals previously unidentified differentially expressed genes. *Journal of Investigative Dermatology* 132, 246–249.

Jang, Y. H., Park, H.-Y., Lee, K. C., Thu, M. P., Kim, S.-K., Suh, M. C., Kang, H., and Kim, J.-K. (2014). A homolog of splicing factor SF1 is essential for development and is involved in the alternative splicing of pre–mRNA in Arabidopsis thaliana. *The Plant Journal* 78, 591–603.

Kitua, A. Y., Smith, T., Alonso, P. L., Masanja, H., Urassa, H., Menendez, C., Kimario, J., and Tanner, M. (1996). Plasmodium falciparum malaria in the first year of life in an area of intense and perennial transmission. *Tropical Medicine & International Health* 1, 475–484.

Lee, K., Bhattacharya, B. B., Qin, J., and Small, D. S. (2016). Nonparametric inference for distributional treatment effects in instrumental variable models. *arXiv preprint arXiv:1605.03868*.

Lei, L. and Fithian, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B* 80, 649–679.

Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* 98, 403–416.

Li, A. and Barber, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *Journal of the American Statistical Association* 112, 837–849.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure–adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B* 81, 45–74.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B* 74, 163–182.

Liang, K. (2016). False discovery rate estimation for large–scale homogeneous discrete p-values. *Biometrics* 72, 639–648.

Liang, K. (2019). Empirical Bayes analysis of RNA sequencing experiments with auxiliary information. *The Annals of Applied Statistics* 13, 2452–2482.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80, 221–239.

MacDonald, P., Liang, K., and Janssen, A. (2019). Dynamic adaptive procedures that control the false discovery rate. *Electronic Journal of Statistics* 13, 3009–3024.

Maroufy, V. (2016). *Applications of geometry in optimization and statistical estimation*. PhD thesis. University of Waterloo.

Maroufy, V. and Marriott, P. (2017). Mixture models: building a parameter space. *Statistics and Computing* 27, 591–597.

Marriott, P. (2002). On the local geometry of mixture models. *Biometrika* 89, 77–93.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* 99, 990–1001.

Orr, M., Liu, P., and Nettleton, D. (2014). An improved method for computing q-values when the distribution of effect sizes is asymmetric. *Bioinformatics* 30, 3044–3053.

Owen, A. B. (2001). *Empirical Likelihood*. London: Chapman and Hall.

Pommeret, D. and Vandekerkhove, P. (2018). Semiparametric false discovery rate model Gaussianity test. Available at `https://hal.archives-ouvertes.fr/hal-01868272`.

Qin, J. (2017). *Biased Sampling, Over–identified Parameter Problems and Beyond*. Singapore: Springer.

Qin, J. and Leung, D. H. Y. (2005). A semiparametric two–component "compound" mixture model and its application to estimating malaria attributable fractions. *Biometrics* 61, 456–464.

Qin, J., Garcia, T. P., Ma, Y., Tang, M.-X., Marder, K., and Wang, Y. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. *The Annals of Applied Statistics* 8, 1182–1208.

Serfling, R. J. (2000). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209.

Smith, T., Schellenberg, J. A., and Hayes, R. (1994). Attributable fraction estimates and case definitions for malaria in endemic. *Statistics in Medicine* 13, 2345–2358.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B* 66, 187–205.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association* 102, 901–912.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Vounatsou, P., Smith, T., and Smith, A. F. M. (1998). Bayesian analysis of two–component mixture distributions applied to estimating malaria attributable fractions. *Journal of the Royal Statistical Society: Series C* 47, 575–587.

Wood, S. N. (2017). *Generalized Additive Models: an Introduction with R*. Boca Raton: CRC Press.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* 11, 95–103.

Xiang, S., Yao, W., and Yang, G. (2019). An overview of semiparametric extensions of finite mixture models. *Statistical Science* 34, 391–404.

Yu, T., Li, P., and Qin, J. (2019). Maximum smoothed likelihood component density estimation in mixture models with known mixing proportions. *Electronic Journal of Statistics* 13, 4035–4078.

Zhang, M. J., Xia, F., and Zou, J. (2019). Fast and covariate–adaptive method amplifies detection power in large–scale multiple hypothesis testing. *Nature Communications* 10, 3433.

Zhao, H. and Fung, W. K. (2016). A powerful FDR control procedure for multiple hypotheses. *Computational Statistics & Data Analysis* 98, 60–70.

Zheng, C. and Wu, Y. (2019). Nonparametric estimation of multivariate mixtures. *Journal of the American Statistical Association*, 1–16.