Estimating the local false discovery rate via a bootstrap solution to the reference class problem

Farnoosh A. Aghababazadeh,[1] Mayer Alvo,[1] and David R. Bickel[1,2,*]

February 15, 2016

[1] Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

[2] Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, Ontario, Canada.

∗ email: dbickel@uottawa.ca

Abstract

The local false discovery rate (LFDR) can be utilized as a statistical approach for simultaneously analyzing thousands of tests. We present a model for multiple hypothesis testing that incorporates a covariate into each test. Incorporating the covariates may improve the performance of testing procedures, because each covariate contains additional information based on the scientific context of the corresponding test. This method provides different LFDR estimates depending on a tuning parameter that determines a reference class of hypotheses from the covariate. We estimate the optimal

value of that parameter by choosing the one that minimizes the estimated LFDR resulting from bias and the variance in a bootstrap approach. Such an estimation method is called an *adaptive reference class* (ARC) method since the class of hypotheses depends on the data. We apply this method to brain data to detect dyslexic-non-dyslexic difference voxels.

We prove a result for the asymptotic performance of the ARC method under certain assumptions concerning the prior probability of each hypothesis test as a function of the covariate and the LFDR estimator. For finite numbers of hypotheses, we use simulation data to evaluate the performance of the estimator associated with the ARC method. The simulations assuming a large covariate effect indicate that the LFDR estimator has a smaller mean squared error under the ARC method than that under the method that uses the entire set of hypotheses without regard for the covariate.

Key words: Bias-variance trade off; Bootstrap approach; Local false discovery rate; Multiple Testing; Reference class; Tuning parameter.

1 Introduction

Modern scientific technologies, such as microarrays, genetic epidemiology, diffusion tensor imaging (DTI), genome-wide association (GWA) studies, and social science surveys, provide statisticians with hundreds or even thousands of tests to be considered simultaneously. The testing of many null hypotheses simultaneously may increase the number of Type I errors. Statistical approaches such as the family-wise error rate (FWER), the controlling false discovery rate (FDR) (Benjamini and Hochberg, 1995), and the tail-area FDR are often applied in cases where a large number of hypotheses are tested (Efron et al., 2001). Efron et al. (2001) first proposed the local false discovery rate (LFDR). The LFDR provides a measure of confidence in the truth of the null hypothesis that depends on its observed statistic (Efron and Tibshirani, 2002).

In many situations, the considered hypotheses are connected by a scientific context. An ignorance of this scientific context in data analysis can be misleading, as it may increase or decrease the number of false positives. For example, each test in a GWA study corresponds to a specific genetic marker, each test in a DTI brain scan corresponds to a specific brain location, and each test in a microarray corresponds to a specific gene expression level. Efron (2008) pointed out that using the entire set of hypotheses in multiple hypothesis testing approaches, such as FDR, may lead to incorrect inferences regarding the features.

1.1 Motivating Example

Schwartzman et al. (2005) used an advanced MRI technology, DTI, to measure water diffusion in the human brain by scanning the brain. DTI is used to map and characterize the three-dimensional diffusion of a water molecule randomly moving in brain tissue to pro-

vide information regarding the direction of diffusion. The measured diffusivity; that is, the diffusion coefficient, relates the diffusive flux to a concentration gradient (Sundgren et al., 2004), and has units of (mm^2/s). In this study, $n = 12$ children were tested, where $n_1 = 6$ were dyslexic and $n_2 = 6$ were not. Each child received DTI brain scans in $N = 15,443$ locations, with each represented by its own voxel's response. The aim is to determine the dyslexic-non-dyslexic difference at the i^{th} brain location, in relation to reading development in children aged $7-13$ (Deutsch et al., 2005). Each test corresponds to a specific voxel. We have two components (w_i, x_i) associated with each hypothesis for $i = 1, ..., N$, where w_i is an observed test statistic that compares the dyslexic children with those who are not, and x_i is the distance from the back of the brain to the front. The LFDR is used to estimate the proportion of dyslexic-non-dyslexic differences.

Consider N voxels in the brain, where for each voxel we compare the dyslexic patients with the non-dyslexic patients. Let H_{0i} denote the null hypothesis that there is no dyslexic-non-dyslexic difference for the i^{th} voxel. Under the i^{th} null hypothesis, assume that $Z_i \sim N(0, 1)$, where Z_i represents the *two-sample* test statistic. Let $W_i = Z_i^2$. The observed statistics $\boldsymbol{w} = (w_1, ..., w_N)^T$ are considered as a realization of $\boldsymbol{W} = (W_1, ..., W_N)^T$. Under the i^{th} null hypothesis it holds that $W_i \sim \chi^2_1$, while under the i^{th} alternative hypothesis we have $W_i \sim \chi^2_{1,\delta}$, where $\delta \in (0, \infty)$ is an unknown noncentrality parameter, according to the models employed in Bukszár et al. (2009) and Yang et al. (2013). Let A_i be an indicator variable for the event that the i^{th} alternative hypothesis H_{ai} is true. Assume the A_is are independent and identically distributed (i.i.d.) Bernoulli$\big(1 - \pi_0\big)$ variables, where π_0 is the prior probability that the i^{th} null hypothesis is true.

The posterior probability that the i^{th} null hypothesis is true given $W_i = w_i$ is the LFDR (Efron et al., 2001), and is denoted as $\Psi(w_i)$, where

$$\Psi(w_i) = \mathrm{P}(A_i = 0 | W_i = w_i) = \frac{\pi_0 g_0(w_i)}{g\big(w_i; \pi_0, \delta\big)}, \tag{1}$$

and $g_0(w_i) \sim \chi^2_1$ denotes the null density of W_i. Denoting the mixture density of W_i by $g\big(w_i; \pi_0, \delta\big)$, we have

$$g\big(w_i; \pi_0, \delta\big) = \pi_0 g_0(w_i) + \big(1 - \pi_0\big) g_1(w_i; \delta), \tag{2}$$

and $g_1(w_i; \delta)$ represents the unknown alternative density of W_i. The model defined in (2) is called the *two-component* model. Under the two-component model with the alternative hypothesis, the observed test statistics are sampled from the same distribution $W_i \sim \chi^2_{1,\delta}$.

Under the i.i.d. assumption for the test statistics, the log-likelihood function now contains only two unknown parameters π_0 and δ, with

$$\ell(\pi_0, \delta) = \sum_{i=1}^{N} \log \big(\pi_0 g_0(w_i) + \big(1 - \pi_0\big) g_1(w_i; \delta)\big), \tag{3}$$

for which the maximum likelihood (ML) estimates $\widehat{\pi}_0(\boldsymbol{w})$ and $\widehat{\delta}(\boldsymbol{w})$ may be derived numerically. For the i^{th} hypothesis test, we have that

$$\widehat{\Psi}_i(\boldsymbol{w}) = \frac{\widehat{\pi}_0(\boldsymbol{w}) g_0(w_i)}{g\big(w_i; \widehat{\pi}_0(\boldsymbol{w}), \widehat{\delta}(\boldsymbol{w})\big)} \tag{4}$$

denotes an estimate of $\Psi(w_i)$ in (1).

For our considered brain data, the observed statistics vector $\boldsymbol{w}$ under the two-component model is applied to estimate the proportion of dyslexic-non-dyslexic differences. From (3), we estimate the values $\widehat{\pi}_0(\boldsymbol{w}) = 0.923$ and $\widehat{\delta}(\boldsymbol{w}) = 3.7$. Figure 1.1 illustrates $N = 15,443$

observed statistics for estimating the LFDR versus the brain location, $(x_i, \widehat{\Psi}_i(\boldsymbol{w}))$, where $\widehat{\Psi}_i(\boldsymbol{w})$ is computed from (4). A total of 119 dyslexic-non-dyslexic differences with LFDR estimates that are lower than 0.2 are identified.

From Figure 1.1 (a), consider the example of $x_i = 53$ and $w_i = 11.12$, which has an estimated LFDR of $\widehat{\Psi}_i(\boldsymbol{w}) = 0.19$ that is close to the threshold of 0.2. For that specific location, we define a reference class that contains the voxels in such a way that their locations are within a symmetric window around $x_i = 53$ with width 2Δ. Such a symmetric window is denoted by $\boldsymbol{w}_i^{\Delta}$, where

$$\boldsymbol{w}_i^{\Delta} = \{w_j : \ |x_j - x_i| \leq \Delta, \ j = 1, ..., N\}. \tag{5}$$

Different window widths yield different reference classes. Again, under the two-component model the reference class $\boldsymbol{w}_i^{\Delta}$ is used to estimate the LFDR $\Psi(w_i)$, with

$$\widehat{\Psi}_i(\boldsymbol{w}_i^{\Delta}) = \frac{\widehat{\pi}_0(\boldsymbol{w}_i^{\Delta}) g_0(w_i)}{g\big(w_i; \widehat{\pi}_0(\boldsymbol{w}_i^{\Delta}), \widehat{\delta}(\boldsymbol{w}_i^{\Delta})\big)}, \tag{6}$$

where $\widehat{\pi}_0(\boldsymbol{w}_i^{\Delta})$ and $\widehat{\delta}(\boldsymbol{w}_i^{\Delta})$ are ML estimates determined from (3) using only the w_js from $\boldsymbol{w}_i^{\Delta}$. Figure 1.1 illustrates the LFDR estimates versus the reference class width, $(2\Delta, \widehat{\Psi}_i(\boldsymbol{w}_i^{\Delta}))$. From Figure 1.1, it can be observed that changing the reference class width provides different LFDR estimates. This example raises the important question of how we can estimate the optimal reference class for estimating the LFDR.

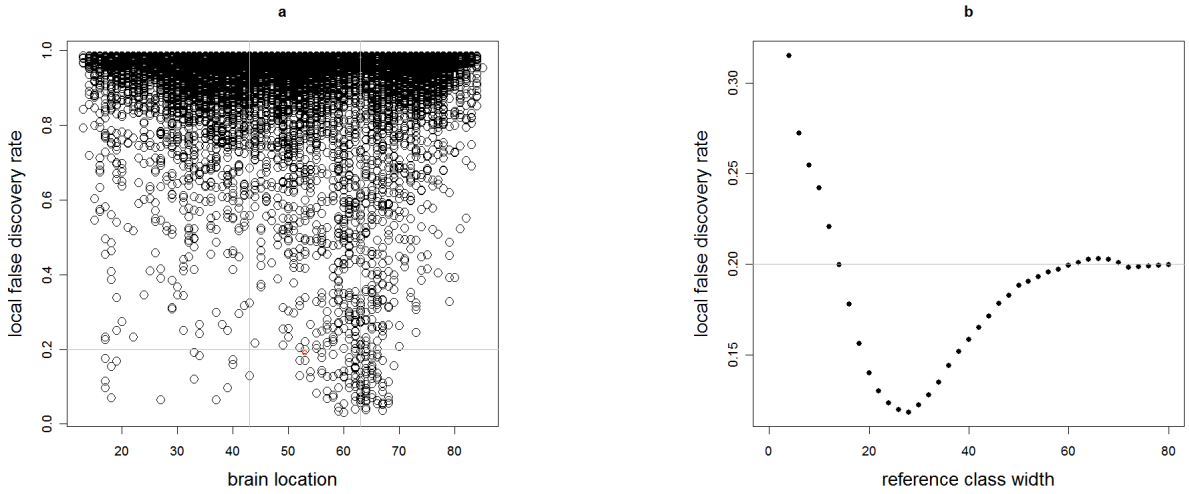

Figure 1.1: Brain data LFDR estimates under the two-component model versus the brain locations x_i (a) and reference class width (b) for a fixed location $x_i = 53$. The horizontal line represents the threshold of 0.2. The vertical lines in (a) indicate the symmetric around $x_i = 53$ with $\Delta = 10$.

Considering all N voxels in estimating LFDR in (3) for a specific brain location, instead of considering the different subsets of voxels, is called the *combined reference class* (CRC) method (see, e.g., (5)). A set of hypotheses or features for determining the posterior probability of a null hypothesis is called the *reference class*, and the problem of finding such a set is an example of the reference class problem (Bickel, 2013). For the considered brain data, the reference class problem consists of deciding which voxels should be used to determine the probability that a voxel does not represent a dyslexic-non-dyslexic difference.

1.2 Previous Research Regarding the Reference Class Problem

Efron (2008) pointed out that using the entire set of hypotheses as opposed to a different reference class may change the number of false positives. He proposed the separate-class model, in which the hypotheses were divided into distinct groups. For example, the brain

data can be divided into two distinct groups according to location in the brain, where we set thresholds on the front-half and rear-half of $x_i \leq 50$ and $x_i > 50$, respectively. Thus, we need to determine which reference class should be used to determine the posterior probability of no dyslexic-non-dyslexic difference occurring at the brain location $x_i = 53$. Should we use the entire set of voxels or the rear-half voxels?

In many cases, the hypotheses can be divided into groups based on the characteristics of the problem. In the above example, the locations for the brain data are incorporated as covariates. Efron's approach divides the covariates into a finite number of groups, where some information is lost. Many methods have been proposed for incorporating covariates into a variety of statistical techniques for handling multiple hypothesis testing. Benjamini and Hochberg (1997) used a p-value-weighting method and evaluated different procedures. Bickel (2004, §4.3) considered the effect of selecting to use the test statistics in estimators of the weighted and unweighted FDR, and found that smaller reference classes of null hypotheses yield lower estimated expected losses than larger reference classes. Some researchers have applied the idea of incorporating a group structure and weights to improve the statistical power. Such a group structure can be used in multiple hypothesis testing by assigning weights for the hypotheses or p-values in each group. Genovese et al. (2006) demonstrated that a p-value weighting procedure can be employed to control the FWER and FDR while increasing statistical power. Subsequently, Wasserman and Roeder (2006) introduced an optimal p-value weighting procedure for FWER control. Hu et al. (2010) proposed a weighting scheme based on a simple Bayesian framework. The proportion of null hypotheses that are true was employed within each group. Such an approach can control the FDR for both the independent hypotheses and p-values with certain dependence structures. The unknown proportion of true null hypotheses is estimated within each group. Each of the statistical

techniques reviewed above requires a finite number of groups specified before examination of the data.

By contrast, we adopt the general strategy of Zablocki et al. (2014), proposing methods for the presence of a covariate that generalizes the pre-data separation of hypotheses into groups. Zablocki et al. (2014) used a hierarchical Bayesian approach to incorporate a set of covariates, where the prior probability that the null hypothesis test is true and the alternative distribution of the test statistic are both modulated by covariates.

Instead of specifying the hyperprior distributions that are required for a hierarchical Bayesian approach, we use an empirical Bayes approach to estimate an optimal reference class for improving the LFDR estimate. We assume the prior probability to be a function of a covariate, as mentioned in Section 2.1. After introducing such a model, the LFDR corresponding to the model is estimated. In Section 2, we propose an adaptive reference class (ARC) method, where a bootstrap approach is used in order to estimate the optimal reference class. In Section 3, we prove under some assumptions that the ARC method has an asymptotically smaller mean squared error in estimating the LFDR than the CRC method. In Section 4.1, we conduct a simulation study to assess the performance of the LFDR estimator for each method. We present an application of the ARC method on a set of brain data in Section 4.2. Finally, we conclude the paper with a brief discussion in Section 5. The proof for the main theorem is included in the Appendix.

2 Methodology

2.1 Proposed Model

The described model in Section 1 extends to the situation where a covariate related to the scientific context of each hypothesis test is incorporated. For our set of brain data, the covariate represents the location in the brain. Let $\boldsymbol{X} = (X_1, ..., X_N)^T$ be i.i.d. random variables with a probability distribution P_x. Any test statistics may be transformed to the standard normal statistic Z_i, for $i = 1, ..., N$. The observed statistics vector $\boldsymbol{z} = (z_1, ..., z_N)^T$ is considered to be a realization of $\boldsymbol{Z} = (Z_1, ..., Z_N)^T$. Let A_i be the event that the i^{th} alternative hypothesis H_{ai} is true. Assume that $A_i|X_i = x_i \sim \text{Bernoulli}\big(1 - \pi_0(x_i)\big)$, where $\pi_0(x_i)$ the prior probability that the i^{th} null hypothesis is true, is an unknown function of the given covariate $X_i = x_i$. We denote the posterior probability that the i^{th} null hypothesis is true given $Z_i = z_i$ and $X_i = x_i$ by

$$\Psi(z_i; x_i) \;\; = \mathrm{P}(A_i = 0 | Z_i = z_i, X_i = x_i) \;\; = \frac{\pi_0(x_i) f_0(z_i)}{f\big(z_i; \pi_0(x_i)\big)}, \tag{7}$$

where the mixture density of Z_i that is conditional on the covariate $X_i = x_i$ is given by

$$f\big(z_i; x_i\big) = \pi_0(x_i) f_0(z_i) + \big(1 - \pi_0(x_i)\big) f_1(z_i; x_i), \tag{8}$$

where $f_0(z_i)$ denotes the null density of Z_i and $f_1(z_i; x_i)$ is the alternative density of Z_i. The mixture density in (2) is a special form of (8), where instead of applying Z_i the statistic W_i from Section 1 is used. The quantities $\pi_0(x_i)$ and $f_1(z_i; x_i)$ are unknown. The ARC method is applied to estimate the LFDR in (7). Under the CRC method, the effects of the covariates

are ignored (see in Section 1), while under the proposed ARC method some assumptions are considered locally to estimate the LFDR $\Psi(z_i; x_i)$ defined in (7).

2.2 Adaptive Reference Class (ARC) Method

Under the ARC method, certain assumptions only hold locally, within a symmetric window for each covariate. Let a symmetric window of width 2Δ be centered at a given covariate $X_i = x_i$ in (5). Let Δ_0 denote the smallest considered value of the tuning parameter Δ. The reference class $\boldsymbol{z}_i^{\Delta}$ contains components z_j such that their covariates are within a distance Δ of x_i. Denoting the expected dimension of the reference class $\boldsymbol{z}_i^{\Delta}$ by d_i^{Δ}, we have

$$\mathrm{d}_i^{\Delta} = N\mathrm{P}\big(|X_j - x_i| \leq \Delta,\ j = 1, ..., N\big). \tag{9}$$

d_i^{Δ} increases as the number of null hypothesis tests N becomes larger, provided that the probability is positive. For each reference class $\boldsymbol{z}_i^{\Delta}$, we may apply any LFDR estimation approach such as the two-component model in Section 1.1. In contrast with the CRC method, instead of using the entire collection of observed statistics $\boldsymbol{z}$, only the reference class $\boldsymbol{z}_i^{\Delta}$ is used in obtaining the LFDR estimate $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$. The choice of tuning parameter Δ influences the LFDR estimates, and we will focus on choosing the one that results in the lowest error in estimating the LFDR, which is called the *optimal* tuning parameter.

2.2.1 Optimal Tuning Parameter

The optimal tuning parameter Δ specifies the symmetric window width of a given reference class, and will be determined by minimizing the errors resulting from bias and variance. In the following, we introduce some notational conventions.

Let the mean and variance of the estimator $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$ be defined as

$$\begin{aligned}\mu_{\Delta}(x_i) =& \mathrm{E}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big),\\ \sigma^2_{\Delta}(x_i) =& \mathrm{E}\big[\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta}) - \mu_{\Delta}(x_i)\big)^2|X_i = x_i\big],\end{aligned} \tag{10}$$

respectively. When $X_i = x_i$, the prediction bias for the estimator $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$ is denoted by $\mathcal{B}_{\Delta}(x_i)$, with

$$\mathcal{B}_{\Delta}(x_i) = \mathrm{E}\big[\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta}) - \Psi(z_i; x_i)\big)|X_i = x_i\big]. \tag{11}$$

Determining the optimal choice of Δ depends on the choice of a loss function for measuring errors in the estimation of the LFDR. The expectation of the quadratic loss function gives us a criterion for choosing the optimal tuning parameter Δ. The mean squared error for the estimator $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$ conditional on $X_i = x_i$ is defined as

$$\mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big) = \mathrm{E}\big[\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta}) - \Psi(z_i; x_i)\big)^2|X_i = x_i\big]. \tag{12}$$

It can be shown that the portion of the mean squared error that depends on Δ is given by

$$\mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big) = \sigma^2_{\Delta}(x_i) + \mathcal{B}^2_{\Delta}(x_i). \tag{13}$$

We shall employ the errors resulting from bias and variance in (13) to find the optimal Δ. Denoting the optimal Δ by $\Delta^{\star}_{0i}$, we have

$$\Delta^{\star}_{0i} = \arg \inf_{\Delta \geq \Delta_0} \mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big). \tag{14}$$

To estimate $\Delta_{0i}^{\star}$, it is necessary to estimate the variance and prediction bias of the LFDR estimator. The bootstrap approach is employed to estimate these quantities.

2.2.2 Bootstrap Estimation of the Optimal Tuning Parameter

We re-sample N pairs from $\{(z_1, x_1), ..., (z_N, x_N)\}$, until B bootstrap samples are obtained that contain the specific pair (z_i, x_i), where $z_i \in \boldsymbol{z}$ and $x_i \in \boldsymbol{x}$. Such samples are denoted by $(\boldsymbol{z}_1^{\star}, \boldsymbol{x}_1^{\star}), ..., (\boldsymbol{z}_B^{\star}, \boldsymbol{x}_B^{\star})$. The b^{th} bootstrap sample $(\boldsymbol{z}_b^{\star}, \boldsymbol{x}_b^{\star})$ contains pairs $(z_{bj}^{\star}, x_{bj}^{\star})$, for $j = 1, ..., N$ and $b = 1, ..., B$. From (5), the b^{th} bootstrap reference class is defined as

$$\boldsymbol{z}_{i,b}^{\Delta} = \{z_{bj}^{\star} : \ |x_{bj}^{\star} - x_i| \leq \Delta, \ j = 1, ..., N\}. \tag{15}$$

The estimate of $\Psi(z_i; x_i)$ based on the b^{th} bootstrap reference class is denoted by $\widehat{\Psi}_i(\boldsymbol{z}_{i,b}^{\Delta})$. The random variables $\widehat{\Psi}_i(\boldsymbol{z}_{i,1}^{\Delta}), ..., \widehat{\Psi}_i(\boldsymbol{z}_{i,B}^{\Delta})$ provide the estimators $\widehat{\mu}(\Delta, B)$ and $\widehat{\sigma}^2(\Delta, B)$ for estimating $\mu_{\Delta}(x_i)$ and $\sigma_{\Delta}^2(x_i)$, respectively, where

$$\widehat{\mu}(\Delta, B) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\Psi}_i(\boldsymbol{z}_{i,b}^{\Delta}) \ \text{ and } \ \widehat{\sigma}^2(\Delta, B) = \frac{1}{B-1}\sum_{b=1}^{B} \left(\widehat{\Psi}_i(\boldsymbol{z}_{i,b}^{\Delta}) - \widehat{\mu}(\Delta, B)\right)^2. \tag{16}$$

In order to estimate the prediction bias in (13), $\pi_0(x_i)$ must be estimated. We propose the use of a reference class $\boldsymbol{z}_{i,b}^{\Delta_0}$ that contains the observed statistics z_js whose covariates are within a distance Δ_0 of x_i. Thus, the estimator $\widehat{\mu}(\Delta_0, B)$ from (16) can be used to estimate $\pi_0(x_i)$. Denoting the bootstrap estimator of the prediction bias in Lemma 3 by $\widehat{\mathcal{B}}(\Delta, \Delta_0, B)$, we have that

$$\widehat{\mathcal{B}}(\Delta, \Delta_0, B) = \widehat{\mu}(\Delta, B) - \widehat{\mu}(\Delta_0, B). \tag{17}$$

The estimator of $\mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big)$ in (13) is denoted by $\widehat{\mathrm{err}}(\Delta, \Delta_0, B)$, and is computed by simply summing the bootstrap variance in (16) and the squared bootstrap prediction bias in (17). Let the optimal $\Delta_{0i}^{\star}$ be denoted by $\widehat{\Delta}_{0i}^{\star}$, which is given by

$$\widehat{\Delta}_{0i}^{\star} = \arg \inf_{\Delta \geq \Delta_0} \widehat{\mathrm{err}}(\Delta, \Delta_0, B). \tag{18}$$

After determining the optimal tuning parameter $\widehat{\Delta}_{0i}^{\star}$, the optimal reference class $\boldsymbol{z}_i^{\widehat{\Delta}_{0i}^{\star}}$ is determined from (5), which contains the z_js whose covariates are within a distance $\widehat{\Delta}_{0i}^{\star}$ of x_i. The optimal reference class $\boldsymbol{z}_i^{\widehat{\Delta}_{0i}^{\star}}$ is used to estimate the LFDR in (7). This LFDR estimate is denoted by $\widehat{\Psi}_i\big(\boldsymbol{z}_i^{\widehat{\Delta}_{0i}^{\star}}\big)$. The estimation methods detailed above yield two estimators. The estimator $\widehat{\Psi}_i(\boldsymbol{z})$ is related to the CRC method and $\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}_{0i}^{\star}})$ is computed using the ARC method.

3 Theoretical Results

In this section, we describe the theoretical properties of the LFDR estimator based on the ARC method. The connection between the true LFDR $\Psi(z_i; x_i)$ in (7) and the prior probability $\pi_0(x_i)$ is considered in the following lemma.

Lemma 1. *If the random indicator A_i that is conditional on $X_i = x_i$ is Bernoulli$\big(1-\pi_0(x_i)\big)$, then $E\big(\Psi(z_i; x_i)|X_i = x_i\big) = \pi_0(x_i)$.*

Proof. We have

$$\mathrm{E}(A_i|Z_i = z_i, X_i = x_i) = \mathrm{P}(A_i = 1|Z_i = z_i, X_i = x_i) = 1 - \Psi(z_i; x_i).$$

By taking the expectation with respect to the density of Z_i that is conditional on $X_i = x_i$,

we obtain

$$\begin{aligned}\mathrm{E}\big(\mathrm{E}(A_i|Z_i=z_i,X_i=x_i)|X_i=x_i\big) =& \mathrm{E}\big(1-\Psi(z_i;x_i)|X_i=x_i\big),\\ \mathrm{E}(A_i|X_i=x_i) =& 1-\mathrm{E}\big(\Psi(z_i;x_i)|X_i=x_i\big),\end{aligned}$$

and the result follows. □

The prediction bias for the estimator $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$ can be expanded as

$$\begin{aligned}\mathcal{B}_{\Delta}(x_i) &= \mathrm{E}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})-\Psi(z_i;x_i)\big)|X_i=x_i\big]\\ &= \mu_{\Delta}(x_i)-\pi_0(x_i).\end{aligned} \tag{19}$$

Then, it can be shown that the mean squared error of the estimator $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$ that is conditional on $X_i=x_i$ is expanded as

$$\mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i=x_i\big)=\sigma_{\Delta}^2(x_i)+\mathcal{B}_{\Delta}^2(x_i)+\mathrm{Var}\big(\Psi(z_i;x_i)|X_i=x_i\big). \tag{20}$$

Let $\widehat{\Psi}(\boldsymbol{z}_{i,1}^{\Delta}),...,\widehat{\Psi}(\boldsymbol{z}_{i,B}^{\Delta})$ be i.i.d. with mean $\mu_{\Delta}(x_i)$ and variance $\sigma_{\Delta}^2(x_i)$. It is known that the bootstrap sample mean and the bootstrap sample variance are weakly consistent estimators for the mean and variance of $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$, respectively, as B becomes large. In addition, because the estimator under the two-component model in (4) involves the ML estimators $\widehat{\pi}_0(\boldsymbol{w})$ and $\widehat{\delta}(\boldsymbol{w})$, it is a weakly consistent estimator of $\Psi(w_i)$.

Our main concern now is related to the estimator $\widehat{\mu}(\Delta_0,B)$, which depends on Δ_0. An appropriate choice of Δ_0 may give us a weakly consistent estimator for $\pi_0(x_i)$. For a given

x_0, we shall suppose that the unknown prior probability $\pi_0(X_i)$ is a step function of the covariate X_i, given by

$$\pi_0(X_i) = \begin{cases} \pi_{01} & \text{if } X_i \leq x_0, \\ \pi_{02} & \text{if } X_i > x_0, \end{cases} \tag{21}$$

where the prior probabilities π_{01} and π_{02} are both unknown with $\pi_{01} \leq \pi_{02}$. A function such as that in (21) will have a biologically meaningful interpretation. The function splits the N tests into two distinct groups, such that under each group the test statistics are i.i.d. Under the assigned values of x_0 and Δ_0, the observed vector of covariates $\boldsymbol{x}$ may be partitioned into three regions, as follows:

$$\begin{aligned} \mathcal{R}_1(x_0, \Delta_0) =& \{x_i : \ x_i \leq x_0 - \Delta_0, \ i = 1, ..., N\}, \\ \mathcal{R}_2(x_0, \Delta_0) =& \{x_i : \ x_0 - \Delta_0 < x_i < x_0 + \Delta_0, \ i = 1, ..., N\}, \\ \mathcal{R}_3(x_0, \Delta_0) =& \{x_i : \ x_i \geq x_0 + \Delta_0, \ i = 1, ..., N\}. \end{aligned} \tag{22}$$

We prove the following results for the region $\mathcal{R}_1(x_0, \Delta_0)$, and the same results can be demonstrated to hold for the region $\mathcal{R}_3(x_0, \Delta_0)$. The following lemma, which is proven in the Appendix, indicates that in the region $\mathcal{R}_1(x_0, \Delta_0)$ the bootstrap estimator $\widehat{\mu}(\Delta_0, B)$ is a weakly consistent estimator for π_{01} as N becomes large.

Lemma 2. *For $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, the bootstrap estimator $\widehat{\mu}(\Delta_0, B)$ is a weakly consistent estimator of π_{01}. That is,*

$$\lim_{N \to \infty} \lim_{B \to \infty} P\big(|\widehat{\mu}(\Delta_0, B) - \pi_{01}| > \epsilon | X_i = x_i\big) = 0 \ \textit{for any } \epsilon > 0.$$

The next lemma shows that the bootstrap estimator $\widehat{\mathcal{B}}(\Delta, \Delta_0, B)$ is a weakly consistent estimator of the prediction bias in (11).

Lemma 3. *If $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, then the bootstrap estimator $\widehat{\mathcal{B}}(\Delta, \Delta_0, B)$ is a weakly consistent estimator of the prediction bias $\mathcal{B}_\Delta(x_i)$. That is,*

$$\lim_{N\to\infty} \lim_{B\to\infty} P\big(|\widehat{\mathcal{B}}(\Delta, \Delta_0, B) - \mathcal{B}_\Delta(x_i)| > \epsilon | X_i = x_i\big) = 0 \ \textit{for any}\, \epsilon > 0.$$

Proof. By Markov's inequality, we have for any $\epsilon > 0$ that

$$\begin{aligned}
\mathrm{P}\big(|\widehat{\mathcal{B}}(\Delta, \Delta_0, B) - \mathcal{B}_\Delta(x_i)| > \epsilon | X_i = x_i\big) \leq & \frac{\mathrm{E}\big[|\widehat{\mathcal{B}}(\Delta, \Delta_0, B) - \mathcal{B}_\Delta(x_i)|\big| X_i = x_i\big]}{\epsilon} \\
\leq & \frac{\mathrm{E}\big[|\widehat{\mu}(\Delta, B) - \mu_\Delta(x_i)|\big| X_i = x_i\big]}{\epsilon} \\
& + \frac{\mathrm{E}\big[|\widehat{\mu}(\Delta_0, B) - \pi_{01}|\big| X_i = x_i\big]}{\epsilon}.
\end{aligned}$$

Because $\widehat{\mu}(\Delta, B)$ is an unbiased estimator of $\mu_\Delta(x_i)$ whose variance is asymptotically zero in (24), the result follows from Lemma 2 and the fact that

$$\lim_{B\to\infty} \mathrm{E}\big[|\widehat{\mu}(\Delta, B) - \mu_\Delta(x_i)|\big| X_i = x_i\big] \leq \lim_{B\to\infty} \mathrm{E}\big[\big(\widehat{\mu}(\Delta, B) - \mu_\Delta(x_i)\big)^2 | X_i = x_i\big]^{\frac{1}{2}} = 0.$$

□

In the next lemma, we consider the consistency of the bootstrap estimator $\widehat{\Delta}^\star_{0i}$.

Lemma 4. *For $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, the bootstrap estimator $\widehat{\Delta}^\star_{0i}$ is a weakly consistent estimator of $\Delta^\star_{0i}$. That is,*

$$\lim_{N\to\infty}\lim_{B\to\infty} P\big(|\widehat{\Delta}^{\star}_{0i} - \Delta^{\star}_{0i}| > \epsilon | X_i = x_i\big) = 0 \;\; \textit{for any}\, \epsilon > 0.$$

Proof. The bootstrap sample variance is a weakly consistent estimator of the variance of $\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})$ and it follows from Lemma 3, that

$$\lim_{N\to\infty}\lim_{B\to\infty} \mathrm{P}\big(|\widehat{\mathrm{err}}\big(\Delta, \Delta_0, B\big) - \mathrm{err}\big(\widehat{\Psi}(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big)| > \epsilon | X_i = x_i\big) = 0.$$

Therefore, the result follows from the continuous mapping Theorem and the fact that

$$\lim_{N\to\infty}\lim_{B\to\infty} \mathrm{P}\big(|\arg\inf_{\Delta\geq\Delta_0} \widehat{\mathrm{err}}\big(\Delta, \Delta_0, B\big) - \arg\inf_{\Delta\geq\Delta_0} \mathrm{err}\big(\widehat{\Psi}(\boldsymbol{z}_i^{\Delta})|X_i = x_i\big)| > \epsilon | X_i = x_i\big) = 0.$$

□

The main result states that the estimator of the ARC method has an asymptotically smaller mean squared error than the estimator of the CRC method for the regions $\mathcal{R}_1(x_0, \Delta_0)$ and $\mathcal{R}_3(x_0, \Delta_0)$.

Theorem 1. *Let $\widehat{\Psi}_i(\boldsymbol{z})$ be a weakly consistent estimator of $\Psi(z_i)$ when N becomes large. If $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, then*

$$\lim_{N\to\infty}\lim_{B\to\infty} \Big[MSE\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^{\star}_{0i}})|\mathcal{R}_1(x_0, \Delta_0)\big) - MSE\big(\widehat{\Psi}_i(\boldsymbol{z})|\mathcal{R}_1(x_0, \Delta_0)\big)\Big] \leq 0.$$

Proof. See the Appendix. □

4 Results

4.1 Simulated Data Analysis

The aim of the following simulation analysis is to compare the performances of the CRC and ARC methods in estimating the LFDR in (7). In this section, each test statistic is assigned a prior probability that is a function of the covariate.

Let $\pi_0(x_i)$ denote the true prior probability that the i^{th} null hypothesis, representing no dyslexic-non-dyslexic difference voxel, is true. We assume that the proportion of dyslexic-non-dyslexic difference voxels tends to be very small. We present several simulation studies, each with different value of $x_0 \in [0.05, 0.40]$. The data sets were simulated as follows. In each simulation study, we randomly generated $1,000$ data sets, each corresponding to an artificial case-control study. For each data set, we simultaneously generate both covariates and observed Wald χ^2 test statistics, which are denoted by x_i and w_i, respectively. Each observed covariate x_i is generated randomly from the uniform distribution between 0 and 1.

In each simulation study, the true prior probability $\pi_0(x_i)$ is determined according the given value of x_0 as the function of the observed covariates that is defined in (21). From (21), let $\bar{\pi}_0 = \mathrm{E}(\pi_0(X_i))$ for $i = 1, ..., N$, where $\bar{\pi}_0 \in [0.60, 0.95]$. To generate the observed statistics, we generate each $A_i \sim$ Bernoulli$(1 - \pi_0(x_i))$ independently. To generate the observed χ^2 test statistics, if $A_i = 1$ the observed statistics are sampled from $\chi^2_{1,\delta}$ with a noncentrality parameter δ. For each given value of x_0, a different value of $\delta \in [1, 5]$ is assigned. The Wald χ^2 test statistics when $A_i = 0$ are sampled from χ^2_1.

Each data set has N pairs (w_i, x_i). The total number of pairs N is equal to $10,000$. In each simulation study, a pair (w_i, x_i) is randomly selected from each data set to estimate $\Psi(w_i; x_i)$. For a given covariate x_i, the estimators of the LFDR are computed using the two

methods. Under the ARC method, Δ_0 has to be specified in advance to determine $\widehat{\Delta}_{0i}^{\star}$. We consider the range $\Delta_0 \in (0, x_0)$, and set $B = 1,000$. Thus, under the two-component model described in Section 1.1, the estimators $\widehat{\Psi}_i(\boldsymbol{w})$ and $\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}})$ are computed. The conditional MSE approximations to measure the performances of the estimators are given by

$$\begin{aligned}
\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w})|\mathcal{R}_r(x_0,\Delta_0)\big) =& \frac{1}{\#\{x_i \in \mathcal{R}_r(x_0,\Delta_0)\}} \sum_{x_i \in \mathcal{R}_r(x_0,\Delta_0)} \big(\widehat{\Psi}_i(\boldsymbol{w}) - \Psi(w_i;x_i)\big)^2, \\
\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}})|\mathcal{R}_r(x_0,\Delta_0)\big) =& \frac{1}{\#\{x_i \in \mathcal{R}_r(x_0,\Delta_0)\}} \sum_{x_i \in \mathcal{R}_r(x_0,\Delta_0)} \big(\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}}) - \Psi(w_i;x_i)\big)^2,
\end{aligned}$$

for $r = 1, 2, 3$. The marginal MSE approximations are computed as follows:

$$\begin{aligned}
\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w})\big) =& \frac{1}{1000} \sum_{i=1}^{1000} \big(\widehat{\Psi}_i(\boldsymbol{w}) - \Psi(w_i;x_i)\big)^2, \\
\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}})\big) =& \frac{1}{1000} \sum_{i=1}^{1000} \big(\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}}) - \Psi(w_i;x_i)\big)^2.
\end{aligned}$$

The relative MSE of the two estimators is a convenient measure for comparing the MSEs. The conditional and marginal relative MSEs are denoted by

$$\mathrm{ReMSE}_{\mathrm{cond}} = \frac{\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}})|\mathcal{R}_r(x_0,\Delta_0)\big)}{\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w})|\mathcal{R}_r(x_0,\Delta_0)\big)} \text{ and } \mathrm{ReMSE}_{\mathrm{marg}} = \frac{\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w}_i^{\widehat{\Delta}_{0i}^{\star}})\big)}{\widehat{\mathrm{MSE}}\big(\widehat{\Psi}_i(\boldsymbol{w})\big)}, \tag{23}$$

respectively. From Figure 4.1, we observe that the performance of the ARC method depends on the Δ_0 values and the region of covariates. When $\bar{\pi}_0 \in [0.60, 0.95]$, increasing the value of Δ_0 results in a smaller MSE approximation for the ARC method in the regions $\mathcal{R}_1(x_0,\Delta_0)$ and $\mathcal{R}_3(x_0,\Delta_0)$. Figure 4.1 shows that the ARC method has a smaller marginal

MSE approximation than the CRC method. From Table 4.1, if we consider the true prior probabilities for all measured voxels to be independent of the covariates, that is, $\pi_0(x_i) = \pi_0$ for $i = 1, ..., N$, then the CRC method has a smaller MSE than the ARC method. In such cases, the CRC method should be used instead of the ARC method to analyze the data.

π_0	0.60	0.80	0.90	0.95
$\text{ReMSE}_{\text{marg}}$	0.9030	0.8156	1.3729	2.0908

Table 4.1: MSE of the ARC method relative to the CRC method when there is no covariate effect. The true prior probabilities are constant, $\pi_0(x_i) = \pi_0$. The $\log_2$ value is given for the marginal relative MSE. Under the ARC method, Δ_0 is 0.01.

4.2 Brain Data Analysis

The brain related data introduced in Section 1.1, with a total of $N = 15,443$ voxels, is employed in the following statistical analysis. Under the CRC method, all observed statistics $\boldsymbol{w}$ are considered to estimate the LFDR where 119 dyslexic-non-dyslexic difference voxels are identified. The brain location is incorporated as a covariate. Under the ARC method, the optimal reference class is determined for each location x_i, which depends on a choice of Δ_0. Figure 4.2 presents the LFDR estimates under the CRC method versus the ARC method when $\Delta_0 = 20$. We observe from Figure 4.2, that changing Δ_0 has a direct effect on the number of dyslexic-non-dyslexic difference voxels. Under the ARC method, increasing the value of Δ_0 brings the proportion of dyslexic-non-dyslexic difference voxels closer to the corresponding proportion under the CRC method.

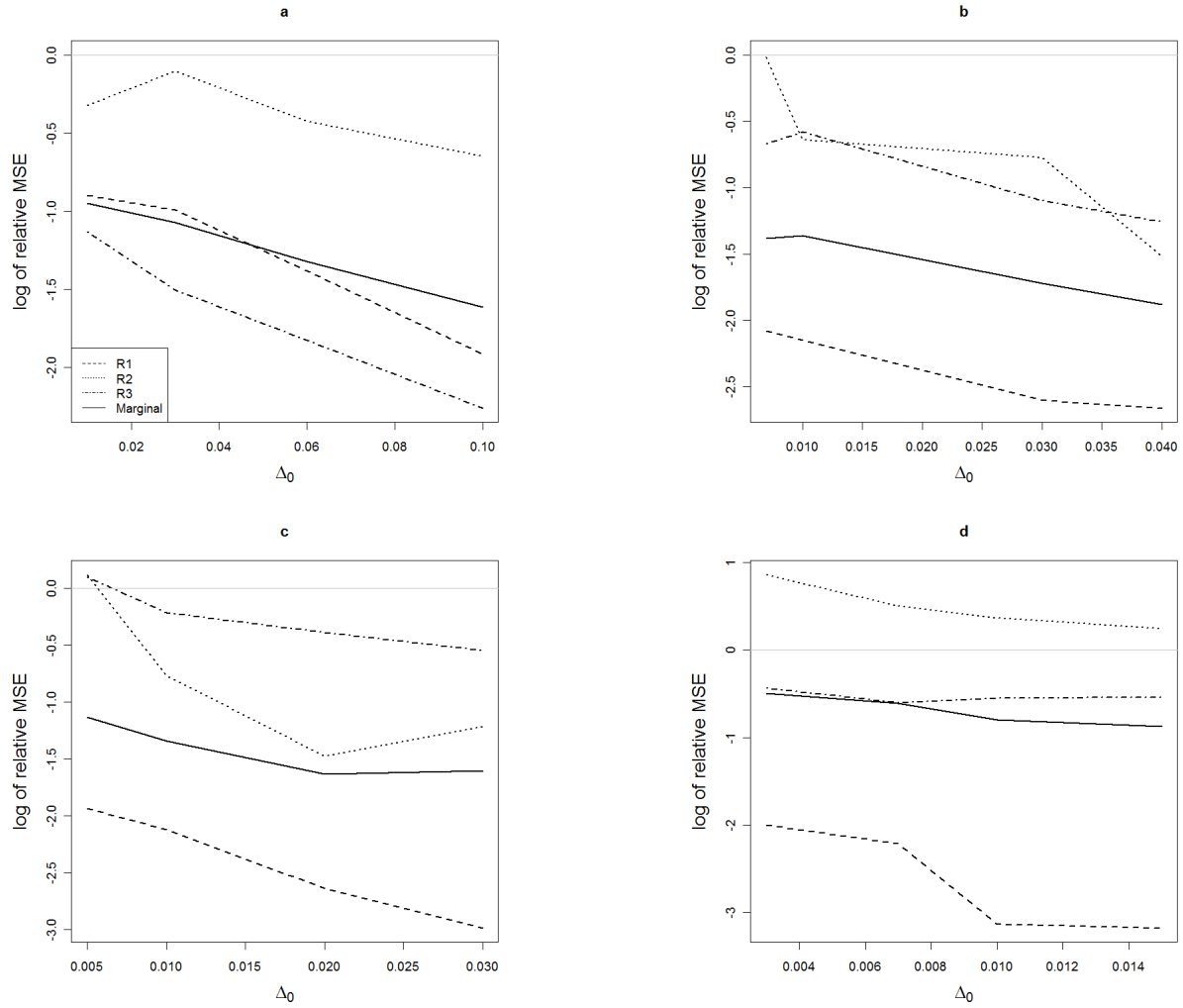

Figure 4.1: The $\log_2$ value of the relative MSE conditional on regions and marginal versus different values of Δ_0; for (a) $\bar{\pi}_0 = 0.60$, (b) $\bar{\pi}_0 = 0.80$, (c) $\bar{\pi}_0 = 0.90$ and (d) $\bar{\pi}_0 = 0.95$, where $\bar{\pi}_0 = \mathrm{E}(\pi_0(X_i))$ for $i = 1, ..., N$.

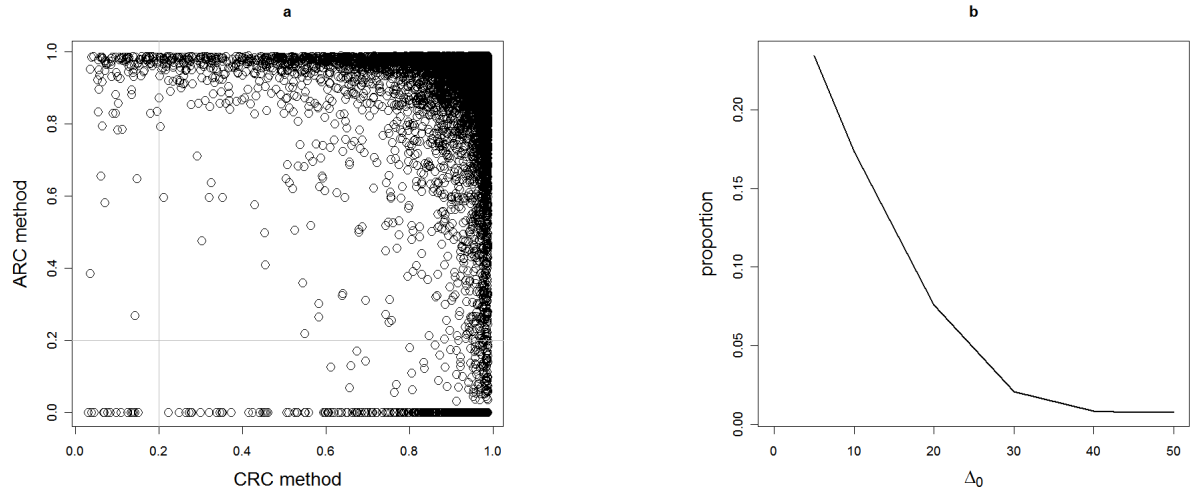

Figure 4.2: Brain data analysis. Panel (a) presents the LFDR estimate under the ARC method for $\Delta_0 = 20$ versus that for the CRC method and (b) illustrates the proportion of dislexic-non-dyslexic difference voxels under the ARC method when the LFDR estimate is less than 0.2 versus $\Delta_0 \in (0, 50)$.

5 Discussion and Conclusions

In this study, a novel approach in which a covariate (i.e., a scientific context corresponding to each hypothesis test) is incorporated to improve the LFDR estimate for identifying the alternative hypotheses. Through this approach, both the test statistic distribution under the alternative hypothesis and the prior probability that the null hypothesis is true are modulated by the covariate. In the case where the prior probability $\pi_0(X_i)$ is the step function in (21), the advantage of our method is that the LFDR estimator under the ARC method performs better than that under the CRC method when N becomes large. It would be interesting to investigate whether this result holds for a general prior probability $\pi_0(X_i)$. Our simulation results confirm that for the regions $\mathcal{R}_1(x_0, \Delta_0)$ and $\mathcal{R}_3(x_0, \Delta_0)$, the LFDR estimator associated with the ARC method has a smaller MSE than that of the CRC method

(see Fig. 4.1). In the region $\mathcal{R}_2(x_0, \Delta_0)$, the weak consistency of $\widehat{\mu}(\Delta_0, B)$ as an estimator of $\pi_0(x_i)$ could not be proven. The ARC method was applied to a set of brain data, as illustrated in Figure 4.2. By increasing the value of the tuning parameter Δ_0, the proportion of significant null hypotheses decreases, and approaches the proportion based on the CRC method.

Acknowledgments

The R packages *Biobase* by Gentleman et al. (2005) and *locfdr* by Efron (2007) facilitated the computational work. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, by a discovery grant from the Natural Sciences and Engineering Research Council of Canada OGP0009068, by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, by the Faculty of Science and Department of Mathematics and Statistics of the University of Ottawa, and by the Faculty of Medicine of the University of Ottawa.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300.

Benjamini, Y., Hochberg, Y., 1997. Multiple hypotheses testing with weights. Scandinavian Journal of Statistics 24 (3), 407–418.

Bickel, D. R., 2004. Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? Statistical Applications in Genetics and Molecular Biology 3, art. 8.

Bickel, D. R., 2013. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. International Statistical Review 81 (2), 188–206.

Bukszár, J., McClay, J. L., van den Oord, E. J. C. G., 2009. Estimating the posterior probability that genome-wide association findings are true or false. Bioinformatics 25, 1807–1813.

Deutsch, G. K., Dougherty, R. F., Bammer, R., Siok, W. T., Gabrieli, J. D., Wandell, B., 2005. Children's reading performance is correlated with white matter structure measured by diffusion tensor imaging. Cortex 41 (3), 354–363.

Efron, B., 2007. Size, power and false discovery rates. The Annals of Statistics, 1351–1377.

Efron, B., 2008. Simultaneous inference: When should hypothesis testing problems be combined? The Annals of Applied Statistics 197–223.

Efron, B., Tibshirani, R., 2002. Empirical Bayes methods and false discovery rates for microarrays. Genetic Epidemiology 23 (1), 70–86.

Efron, B., Tibshirani, R., Storey, J. D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96 (456), 1151–1160.

Genovese, C. R., Roeder, K., Wasserman, L., 2006. False discovery control with p-value weighting. Biometrika 93 (3), 509–524.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S. (Eds.), 2005. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York.

Hu, J. X., Zhao, H., Zhou, H. H., 2010. False discovery rate control with groups. Journal of the American Statistical Association 105 (491).

Schwartzman, A., Dougherty, R. F., Taylor, J. E., 2005. Cross-subject comparison of principal diffusion direction maps. Magnetic Resonance in Medicine 53 (6), 1423–1431.

Sundgren, P., Dong, Q., Gomez-Hassan, D., Mukherji, S., Maly, P., Welsh, R., 2004. Diffusion tensor imaging of the brain: review of clinical applications. Neuroradiology 46 (5), 339–350.

Wasserman, L., Roeder, K., 2006. Weighted hypothesis testing. arXiv preprint math/0604172.

Yang, Y., Aghababazadeh, F. A., Bickel, D. R., 2013. Parametric estimation of the local false discovery rate for identifying genetic associations. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 10 (1), 98–108.

Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M., Thompson, W. K., 2014. Covariate-modulated local false discovery rate for genome-wide association studies. Bioinformatics, btu145.

Appendix

Proof of Lemma 2. By Markov's inequality, it holds for any $\epsilon > 0$ that

$$\begin{aligned}\mathrm{P}\big(|\widehat{\mu}(\Delta_0, B) - \pi_{01}| > \epsilon | X_i = x_i\big) \leq & \frac{\mathrm{E}\big[|\widehat{\mu}(\Delta_0, B) - \pi_{01}| \big| X_i = x_i\big]}{\epsilon} \\ \leq & \frac{\mathrm{E}\big[|\widehat{\mu}(\Delta_0, B) - \mu_{\Delta_0}(x_i)| \big| X_i = x_i\big]}{\epsilon} \\ & + \frac{\mathrm{E}\big[|\mu_{\Delta_0}(x_i) - \pi_{01}| \big| X_i = x_i\big]}{\epsilon}.\end{aligned}$$

$\widehat{\mu}(\Delta_0, B)$ is an unbiased estimator of $\mu_{\Delta_0}(x_i)$, and has zero variance as B becomes large. That is,

$$\lim_{B\to\infty} \mathrm{E}\big[\big(\widehat{\mu}(\Delta_0, B) - \mu_{\Delta_0}(x_i)\big)^2 | X_i = x_i\big] = \lim_{B\to\infty} \frac{\sigma^2_{\Delta_0}(x_i)}{B} = 0. \tag{24}$$

Hence,

$$\lim_{B\to\infty} \mathrm{E}\big[|\widehat{\mu}(\Delta_0, B) - \mu_{\Delta_0}(x_i)| \big| X_i = x_i\big] \leq \lim_{B\to\infty} \mathrm{E}\big[\big(\widehat{\mu}(\Delta_0, B) - \mu_{\Delta_0}(x_i)\big)^2 | X_i = x_i\big]^{\frac{1}{2}} = 0.$$

On the other hand, when $x_i \in \mathcal{R}_1(x_0, \Delta_0)$ the expected dimension of the reference class $\boldsymbol{z}_i^{\Delta_0}$ as N becomes large is $\lim_{N\to\infty} d_i^{\Delta_0} = \infty$. By applying the consistency assumption of $\widehat{\Psi}_i$ on the reference class $\boldsymbol{z}_i^{\Delta_0}$, we have that

$$\lim_{N\to\infty} \mathrm{P}\big(|\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta_0}) - \Psi(z_i; x_i)| > \epsilon | X_i = x_i\big) = 0.$$

Because $|\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta_0}) - \Psi(z_i; x_i)| \leq 1$, the dominated convergence Theorem implies that

$$\lim_{N\to\infty} \mathrm{E}\big[\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta_0}) - \Psi(z_i; x_i)|X_i = x_i\big] = 0.$$

For $x_i \in \mathcal{R}_1(x_0, \Delta_0)$, $\mathrm{E}\big(\Psi(z_i; x_i)|X_i = x_i\big) = \pi_{01}$ and

$$\lim_{N\to\infty} \frac{\mathrm{E}\big[|\mu_{\Delta_0}(x_i) - \pi_{01}|\big|X_i = x_i\big]}{\epsilon} = 0.$$

Proof of Theorem 1. We know that

$$\mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^\star_{0i}})|\mathcal{R}_1(x_0, \Delta_0)\big) = \int_{\mathcal{R}_1(x_0,\Delta_0)} \mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big)\, d\mathrm{P}_{x_i},$$

$$\mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|\mathcal{R}_1(x_0, \Delta_0)\big) = \int_{\mathcal{R}_1(x_0,\Delta_0)} \mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i = x_i\big)\, d\mathrm{P}_{x_i}.$$

It suffices to show that

$$\lim_{N\to\infty}\lim_{B\to\infty} \Big[\mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) - \mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i = x_i\big)\Big] \le 0.$$

$$\begin{aligned}\mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big) - \mathrm{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i = x_i\big) =& \mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^\star_{0i}})|X_i = x_i\big)\\ &-\mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta^\star_{0i}})|X_i = x_i\big)\\ &+\mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta^\star_{0i}})|X_i = x_i\big)\\ &-\mathrm{err}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i = x_i\big).\end{aligned}$$

From Lemma 4, the weak consistency of $\widehat{\Delta}^\star_{0i}$ implies that

$$\lim_{N\to\infty}\lim_{B\to\infty}\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i=x_i\big)-\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta^{\star}_{0i}})|X_i=x_i\big)=0.$$

On the other hand, because $\Delta^{\star}_{0i}$ is optimal tuning parameter, it follows that

$$\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta^{\star}_{0i}})|X_i=x_i\big)-\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i=x_i\big)\leq\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i=x_i\big)-\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i=x_i\big)$$

for any $\Delta\in[\Delta_0,\infty)$, which indicates that

$$\begin{aligned}\lim_{N\to\infty}\lim_{B\to\infty}\big[\operatorname{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\widehat{\Delta}^{\star}_{0i}})|X_i=x_i\big)-\operatorname{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i=x_i\big)\big]&\leq\lim_{N\to\infty}\lim_{B\to\infty}\big[\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i=x_i\big)\\&\quad-\operatorname{err}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i=x_i\big)\big]\\&=\lim_{N\to\infty}\lim_{B\to\infty}\big[\operatorname{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i=x_i\big)\\&\quad-\operatorname{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i=x_i\big)\big]\\&=0.\end{aligned}$$

The facts that both $\lim_{N\to\infty}\operatorname{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z})|X_i=x_i\big)=0$ and $\lim_{N\to\infty}\operatorname{MSE}\big(\widehat{\Psi}_i(\boldsymbol{z}_i^{\Delta})|X_i=x_i\big)=0$ follow from the consistency and the dominated convergence Theorem.